On Faster Convergence of Scaled Sign Gradient Descent

Xiuxian Li, Senior Member, IEEE, Kuo-Yi Lin, Li Li, Member, IEEE, Yiguang Hong, Fellow, IEEE, and Jie Chen, Fellow, IEEE

Abstract—Communication has been seen as a significant bottleneck in industrial applications over large-scale networks. To alleviate the communication burden, sign-based optimization algorithms have gained popularity recently in both industrial and academic communities, which is shown to be closely related to adaptive gradient methods, such as Adam. Along this line, this article investigates faster convergence for a variant of sign-based gradient descent, called scaled SIGNSGD, in three cases: First, the objective function is strongly convex; second, the objective function is nonconvex but satisfies the Polyak–Lojasiewicz inequality; third, the gradient is stochastic, called scaled SIGNSGD in this case. For the first two cases, it can be shown that the scaled SIGNSGD converges at a linear rate. For case third, the algorithm is shown to converge linearly to a neighborhood of the optimal value when a constant learning rate is employed, and the algorithm converges at a rate $O(1/k + 1/k^2 + 1/k^3)$ when using a diminishing learning rate, where $k$ is the iteration number. The results are also extended to the distributed setting by majority vote in a parameter-server framework. Finally, numerical experiments are performed to corroborate the theoretical findings.

Index Terms—Gradient descent (GD), linear convergence, optimization, sign compression.

I. INTRODUCTION

A N OPTIMIZATION problem aims to maximize or minimize an objective function possibly subject to some constraints [1], which has numerous applications in industry, such as electric vehicles [2], [3], smart grid [4], [5], Internet of Things [6], network representation [7], recommender systems [8], and so on. To solve this problem, a quintessential algorithm is the gradient descent (GD) method [9], [10], [11], which requires to access true gradients. However, it is usually expensive or difficult to compute the true gradients in reality, and thereby a typical stochastic gradient descent (SGD) algorithm has become prevalent in deep neural networks [12], [13], which depends upon a lower computing cost for stochastic gradients.

As for large-scale neural networks, the training efficiency can be substantially improved in general by introducing multiple workers in a parameter-server framework, where a group of workers can train their own mini-batch datasets in parallel. Nonetheless, the communication between workers and the parameter server has been a nonnegligible handicap for its wide practical application. As such, as one of gradient compression techniques, sign-based methods have been popular in recent decades, not only because they can reduce the communication cost to 1 bit for each gradient coordinate, but also because they have good performance and close relationship with adaptive gradient methods [14], [15], [16]. As a matter of fact, it has been demonstrated in [15] and [17] that SIGNSGD with momentum often has pretty similar performance to Adam on deep learning missions in practice. Notice that a wide range of gradient compression approaches exist for reducing the communication cost in the literature, e.g., [18] and [19], whose elaboration is beyond the scope of this article. Particularly, sign-based methods considered in this article can be regarded as a special gradient compression scheme, which need to transmit only one bit per gradient component [20]. In addition, the sign-based idea has also been exploited in metaheuristic algorithms (e.g., beetle antennae search algorithm) for optimization, multiobjective optimization, and portfolio optimization problems [21], [22], [23], [24], [25].

Along this line, the sign gradient descent (SIGNGD) algorithm and its stochastic counterpart (SIGNSGD) have been extensively studied in recent years [15], [16], [26], [27], [28]. For instance, it was demonstrated in [15] that SIGNSGD enjoys an SGD-level convergence rate for nonconvex but smooth objective functions under a separable smoothness assumption, which, in combination with majority vote in distributed setup, was further shown to be efficient in terms of communication and fault tolerance in [27]. Recently, the authors in [16] found that the $\ell_{\infty}$-smoothness is a weaker and natural assumption than...
the separable smoothness and established two conditions under which the sign-based methods are preferable over GD.

Motivated by the abovementioned fact, this article studies a sign-based method, called scaled SIGNGD, for unconstrained optimization problems. To the best of authors knowledge, this article is the first to address faster convergence of SIGNGD. The contributions of this article can be summarized as follows.

1) First, it is found that SIGNGD is not generally convergent even for strongly convex and smooth objectives when using constant learning rates, although it is indeed convergent for vanilla GD. Therefore, scaled versions in Algorithms 1 and 2 are investigated. It is proved that Algorithm 1 converges linearly to the minimal value for two cases: strongly convex objectives and nonconvex objectives yet satisfying the Polyak–Łojasiewicz (PL) inequality. Meanwhile, Algorithm 2 converges linearly to a neighborhood of the minimal value when using a constant learning rate \( \alpha \) with an error being proportional to \( \alpha^2 \) and the variance of stochastic gradients. When applying a kind of diminishing learning rate, a rate \( O(1/k + 1/k^2 + 1/k^3) \) can be ensured, which enjoys extra faster rates \( O(1/k^2) \) and \( O(1/k^3) \) than the widely known rate \( O(1/k) \).

2) Second, the obtained results are extended to the distributed setup, where a group of workers compute their own (stochastic) gradients using individual dataset and then transmit the sign gradient and the gradient \( \ell_1 \)-norm to the parameter server that calculates the sign gradient by majority vote along with taking the average of the gradient \( \ell_1 \)-norms and transmits back to all the workers.

Notations: Denote by \([n] := \{1, 2, \ldots, n\}\) for an integer \( n > 0 \). Let \( \| \cdot \|, \| \cdot \|_1, \| \cdot \|_\infty \), and \( x^\top \) be the \( l_2 \)-norm, \( \ell_1 \)-norm, \( \ell_\infty \)-norm, and the transpose of \( x \in \mathbb{R}^n \), respectively. 1 and 0 stand for column vectors of compatible dimension with all entries being 1 and 0, respectively. \( \nabla f \) represents the gradient of a function \( f \), \( \mathbb{E}(\cdot) \) and \( \mathbb{P}(\cdot) \) denote the mathematical expectation and probability, respectively, \( \text{sign} \) denotes the sign function, which is operated componentwise for a vector. Other symbols are summarized in the following.

| Symbol   | Explanation                        |
|----------|------------------------------------|
| GD       | gradient descent                   |
| SGD      | stochastic gradient descent        |
| SIGNGD   | sign gradient descent              |
| SIGNEDG  | sign stochastic gradient descent   |
| PL       | Polyak–Łojasiewicz                 |

II. PROBLEM FORMULATION

This article studies an unconstrained optimization problem

\[
\min_{x \in \mathbb{R}^d} f(x)
\]

(1)

where the objective function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is proper, differentiable, and may be nonconvex. The goal of this article is to solve (1) by resorting to sign-based methods and investigate the convergence speed.

As an extensively studied sign-based algorithm, SIGNGD [15], [16], [27] may be a promising option for solving (1), which is of the form

\[
x_{k+1} = x_k - \alpha_k \text{sign}(\nabla f(x_k))
\]

(2)

where \( \alpha_k > 0 \) is the learning rate. Unfortunately, SIGNGD is not applicable to the studied settings here, as illustrated in the next section.

III. COUNTEREXAMPLES FOR SIGNGD

For SIGNGD, an interesting result can also be found in the continuous-time setup, which demonstrates obvious advantages of SIGNGD compared with GD. Particularly, SIGNGD converges linearly, while GD is only sublinearly convergent.

Example 1: Consider \( f(x) = x_1^2 + x_2^2 \) for \( x \in \mathbb{R}^2 \), which is strongly convex and smooth with \( \nabla f(x) = (2x_1, 2x_2)^\top \). By choosing the initial point as \( x_0 = (\alpha/2, \alpha/2)^\top \), it is easy to verify for (2) that for \( l = 0, 1, 2, \ldots \),

\[
x_{2l} = \left( -\frac{\alpha}{2}, -\frac{\alpha}{2} \right)^\top, \quad x_{2l+1} = \left( \frac{\alpha}{2}, \frac{\alpha}{2} \right)^\top
\]

(3)

which is obviously not convergent.

To fix the issue observed in Example 1, one may attempt to consider the sign counterpart of adaptive gradient methods. However, it generally does not work as well. For instance, the AdaGrad-Norm [29]

\[
b_{k+1}^2 = b_k^2 + \| \nabla f(x_k) \|^2, \quad b_{k+1} = \frac{\eta}{b_{k+1}} \nabla f(x_k), \quad \eta > 0
\]

(4)

is shown to converge linearly without knowing any function parameters beforehand [30], while the linear convergence cannot be ensured in general for its sign counterpart, as illustrated in the following for its two sign variants.

Example 2: Consider the first sign variant as

\[
b_{k+1}^2 = b_k^2 + \| \nabla f(x_k) \|^2
\]

\[
x_{k+1} = x_k - \frac{\eta}{b_{k+1}} \text{sign}(\nabla f(x_k)), \quad \eta > 0
\]

(5)

and \( f(x) = x^2/2 \) (strongly convex and smooth) with \( x \in \mathbb{R} \). For simplicity, set \( b_0 = 0 \) and \( x_0 \neq 0 \). Then, simple manipulations give rise to \( b_{k+1}^2 = \sum_{i=0}^{k} x_i^2 \).

In what follows, we show that the convergence rate of (5) is not linear. To do so, it is easy to see that \( x_{k+1} = x_k - \frac{\eta}{b_{k+1}} \text{sign}(x_k) \), which leads to that

\[
x_{k+1}^2 = x_k^2 - \frac{2\eta}{b_{k+1}} |x_k| + \frac{\eta^2}{b_{k+1}^2}
\]
of classic GD can be invoked for (7), which is known to be sublinear [31]. As shown in Section III, the sign counterparts of GD and AdaGrad-Norm are indeed linearly convergent for strongly convex and smooth objectives, their sign counterparts fail to converge linearly in general.

**Algorithm 1:** Scaled SIGNGD.

**Input:** learning rate $\alpha$, current point $x_k$

$$x_{k+1} = x_k - \alpha \|g_k\|_1 \text{sign}(g_k), \quad g_k := \nabla f(x_k)$$

**Algorithm 2:** Scaled SINGSGD.

**Input:** learning rate $\alpha_k$, current point $x_k$

$$\tilde{g}_k = \text{StochasticGradient}(x_k)$$

$$x_{k+1} = x_k - \alpha_k \|\tilde{g}_k\|_1 \text{sign}(\tilde{g}_k)$$

**A. Deterministic Setting**

Consider the deterministic setting with full gradients, i.e., (13), for which we have the following results. Note that all proofs are given in the Appendix.

**Theorem 1:** The following statements are true for (13).

1) Under Assumptions 1 and 3, if $0 < \alpha < \frac{2}{L}$, then

$$f(x_k) - f^* \leq \zeta^k (f(x_0) - f^*)$$

where $\zeta := 1 - 2\alpha \mu (1 - \frac{\ell_0}{\mu}) \in [0, 1)$.

2) Under Assumptions 2 and 3 with $\alpha$ satisfying $0 < \alpha < \frac{2}{L}$, (15) still holds.
3) If Assumption 3 holds only, then
\[
\min_{\ell \in \{0, 1, \ldots, k\}} \|g_\ell\|^2 \leq \frac{f(x_0) - f^*}{\gamma (k + 1)} \tag{16}
\]
where \(\gamma := \alpha \left(1 - \frac{L_2}{\sigma^2}\right)\).

Remark 3: In view of cases 1) and 2) in Theorem 1, algorithm (13) is proved to be linearly convergent, which is in contrast to SIGNGD and sign AdaGrad-Norm as discussed in Section III. Moreover, for the nonconvex but smooth with respect to the Euclidean norm, by leveraging the similar argument to Theorem 1, it is easy to obtain for SIGNGD with a constant learning rate that \(\min_{\ell \in \{0, 1, \ldots, k\}} \|g_\ell\|^2 \leq \frac{d L_1(x_0) - f^*}{d L_1(k+1)}\) by choosing the learning rate as \(\alpha = \sqrt{\frac{2 (f(x_0) - f^*)}{d L_1(k+1)}}\). In comparison, (16) becomes \(\frac{2 L_1(x_0) - f^*}{L_1(k+1)}\) when \(\alpha\) is chosen as \(\frac{1}{k}\). In this regard, our result is independent of the dimension constant \(d\) and tighter when \(d > 4\), and the learning rate here is easier to implement. In addition, if the smoothness is with respect to the maximum norm, then the result here has the same convergence bound as SIGNGD but with a less conservative learning rate selection.

Remark 4: A similar result can be also obtained from the most related work [32] by resorting to the \(\delta\)-approximate compressor. To be specific, \(C(v) := \text{sign}(v)\) can be viewed as \(\frac{1}{\delta}\)-approximate compressor, and then applying [32, Th. 13] leads to the learning rate \(\alpha \in (0, \frac{1}{\delta})\) and convergence rate \(1 - \frac{\alpha^2}{\sigma^2}\). In contrast, Theorem 1 of this article (need to replace \(\alpha\) by \(\frac{1}{\delta}\) here) is for \(\alpha \in (0, \frac{2\sigma}{\delta})\) with the convergence rate \(1 - \frac{2L_1\alpha}{\sigma^2} (1 - \frac{\alpha}{\delta})\). It is easy to verify that our learning rate is more relaxed and the convergence rate is faster due to \(\frac{\alpha^2}{\sigma^2} \leq \frac{2\sigma^2}{\delta^2} (1 - \frac{\alpha}{\delta})\). Note that for different objectives with different landscapes (thus different gradients), it is possible to have distinct convergence performances. However, all of them at least have the rates obtained here, guaranteed by the theoretical results established in this article, as long as they satisfy the required assumptions, e.g., strong convexity.

Remark 5: The results here have three application merits as follows.

1) As discussed in Remarks 3 and 4, the convergence rates established here are faster than existing ones, indicating that the algorithms here can generally compute solutions more quickly in realistic applications.

2) Algorithms here are based on the sign of gradients, which are more robust to gradient’s noises or uncertainties, since the sign of a positive or negative scalar will still be the true one even though the scalar deviates from its true value as long as it is still positive or negative.

3) The third one will be discussed in Remark 8.

B. Stochastic Setting

This section considers the stochastic gradient case, where the true gradient \(g_k = \nabla f(x_k)\) is expensive to compute and instead a stochastic gradient \(\tilde{g}_k\) is relatively cheap to evaluate as an estimate of \(g_k\). To move forward, some standard assumptions are imposed on stochastic gradients [15], [17].

Assumption 4: The stochastic gradients \(\tilde{g}_k\) are unbiased and have bounded variances with respect to \(\ell_1\)-norm, i.e., there exists a constant \(\sigma > 0\) such that
\[
\mathbb{E}(|\tilde{g}_k|) = g_k, \quad \mathbb{E}(\|\tilde{g}_k - g_k\|^2) \leq \sigma^2. \tag{17}
\]

In this case, the algorithm becomes (14). For brevity, define \(p_{k,i} := \mathbb{P}(\text{sign}(\tilde{g}_{k,i}) = \text{sign}(g_{k,i}))\) for \(k \geq 0\) and \(i \in [d]\), where \(\tilde{g}_{k,i}\) and \(g_{k,i}\) represent the \(i\)th components of \(\tilde{g}_k\) and \(g_k\), respectively.

Remark 6: For stochastic gradient \(g_k\), when leveraging a mini-batch of size \(n_k\) at \(x_k\), the oracle gives us \(n_k\) gradient estimates and in this case, the stochastic gradient \(\tilde{g}_k\) can be chosen as the average of \(n_k\) estimates. In this respect, the variance bound can be reduced to \(\frac{\sigma^2}{n_k}\). In addition, it was shown in [28] that the success probability \(p_{k,i}\) should be greater than \(1/2\), and otherwise the sign algorithm generally fails to work, and a multitude of cases can ensure \(p_{k,i} > 1/2\), for instance, each component \(\tilde{g}_{k,i}\) possesses a unimodal and symmetric distribution [15], [28].

We are now in a position to present the main result on (14).

Theorem 2: For (14), under Assumptions 1, 3, 4, or 2–4, the following statements are true.

1) If \(\alpha_k = \alpha \in (0, \frac{2\mu\sigma}{L_1})\), then
\[
\mathbb{E}(f(x_k)) - f^* \leq \zeta^k_0 (\mathbb{E}(f(x_0)) - f^*) + \frac{L_2\alpha}{2\mu(2\mu_{\text{min}} - 1 - L_0)} \tag{18}
\]
where \(\mu_{\text{min}} := \min_{i \in [d], k \geq 0} p_{k,i}\) and \(\zeta_1 := 1 - 2\mu\alpha(2\mu_{\text{min}} - 1 - L_0) \in \left[\frac{1}{2}, 1\right]\).

2) If \(\alpha_k = \frac{3}{\mu(2\mu_{\text{min}} - 1)(k+1)}\), then
\[
\mathbb{E}(f(x_k)) - f^* \leq \frac{9L_2^2}{\mu^2(2\mu_{\text{min}} - 1)^2} \left(\frac{32}{k} + \frac{1}{k^2}\right) + \frac{f(x_0) - f^*}{(k + 1)^2}. \tag{19}
\]

Remark 7: The first result in Theorem 2 shows that algorithm (14) converges linearly at a rate \(\zeta_1\). This is comparable to vanilla SGD in [33], where the convergence rate is \(1 - \alpha\mu\), which is slower than \(\zeta_1\) (i.e., \(\zeta_1 \leq 1 - \alpha\mu\)) when \(\alpha \in (0, \frac{1}{\sqrt{2}})\). Moreover, the result in (19) is the exact convergence with rate \(O(\frac{1}{k})\) for both strongly convex case and nonconvex case with PL inequality, which is the same as both vanilla SGD [34] and compression methods [35]. In addition, the same rate \(O(\frac{1}{k})\) was established in [36]. However, the condition in [36] for convergence does not always hold, e.g., \(t_k = 1\) of [36, Th. II.2], and our result (19) includes additional faster rates \(O(\frac{1}{k^2})\) and \(O(\frac{1}{k^3})\) except for \(O(\frac{1}{k})\) obtained in [36].

V. DISTRIBUTED SETTING

Now, we extend the results in Section IV to the distributed setting within a parameter-server framework. For simplicity, we only focus on scaled SIGNSGD in this section, but the results can be similarly obtained for scaled SIGNGD.
Algorithm 3: Distributed scaled SIGNGD by majority vote.

**Input:** learning rate $\alpha$, current point $x_k$, # workers $M$

each with an i.i.d. gradient estimate $\hat{g}_k^m$, $m \in [M]$

**On server**

Pull $\text{sign}(\hat{g}_k^m)$ and $\|\hat{g}_k^m\|_1$ from each worker

Push $\text{sign}(\hat{g}_k^m)$ and $M_k$ to each worker

$\hat{g}_k := \frac{1}{M} \sum_{m=1}^{M} \text{sign}(\hat{g}_k^m)$

$M_k := \frac{1}{M} \sum_{m=1}^{M} \|\hat{g}_k^m\|_1$

**On each worker**

$x_{k+1} = x_k - \alpha M_k \text{sign}(\hat{g}_k)$

To proceed, the distributed scaled SIGNGD by majority vote is given in Algorithm 3, for which the following convergence result is obtained.

**Theorem 3:** For Algorithm 3, under Assumptions 1, 3, 4 or 2-4, if $0 < \alpha < \frac{2L_{\text{PL}}(\kappa, \kappa)}{L}$, then

$$\mathbb{E}(f(x_k)) - f^* \leq \zeta_k^2 (\mathbb{E}(f(x_0)) - f^*) + \frac{L\sigma^2\alpha^2}{1 - \zeta_2}$$

where $p_{\text{min}} = \min_{t \in [d], k \geq 0} p_{k,t}$, $\zeta_2 := 1 - 2\mu\alpha(2L_{\text{PL}}(\kappa, \kappa) - 1 - L\alpha) \in \left(\frac{1}{2}, 1\right)$, $\kappa := \left\lceil \frac{M-1}{3} \right\rceil$ with $\lceil \cdot \rceil$ being the floor function, and $I_p(a, b)$ is the regularized incomplete beta function, defined by

$$I_p(a, b) := \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

**Remark 8:** It is noteworthy that the exact convergence can be similarly established as (19) in Theorem 2, which is omitted in Theorem 3. As one of the application merits, sign-based algorithm 3 can largely alleviate the communication burden since only one bit is required for the transmission of each gradient component. However, when using true gradients, it usually needs infinite or extremely large bits to transmit a real number for each gradient component, which is generally computationally prohibitive in practical applications.

VI. EXPERIMENTS

Numerical experiments are provided to corroborate the efficacy of the obtained theoretical results here.

**Example 4:** Consider a simple example with $f(x) = \|x\|^2 + 3\sin^2(x_1) + 3\sin^2(x_2)$ for $x = (x_1, x_2)^T \in \mathbb{R}^2$. It is easy to verify that $f(x)$ is nonconvex, but satisfying the PL condition. To verify the performance of scaled SIGNGD, it is compared with several existing algorithms, i.e., vanilla GD, SIGNGD, SIGNGD (i.e., SIGNUM) [15], and EF-SIGNGD (i.e., (11) [20]). Whereas, SIGNUM is of the form [15], [20]

$$s_{k+1} = \beta s_k + (1 - \beta)g_k$$

$$x_{k+1} = x_k - \alpha s_{k+1}$$

where $\alpha > 0$ is the learning rate and $\beta \in (0, 1)$ is a momentum constant. By setting $\alpha = 0.05$ for all algorithms, $\beta = 0.1$ for SIGNUM, and choosing the same initial state randomly for all algorithms, simulation results are shown in Fig. 1. It can be observed that scaled SIGNGD has the fastest linear convergence, while SIGNGD and SIGNUM cannot converge, behaving oscillations. In summary, this example shows the efficiency of the scaled SIGNGD, and supports the observation in Example 1.

**Example 5:** Consider another example for comparing performances of SIGNGD with different objective functions. To do so, consider the following objectives with $x = (x_1, x_2)^T \in \mathbb{R}^2$:

$$f_1(x) = \|x\|^2 + 3\sin^2(x_1) + 2\sin^2(x_2)$$

$$f_2(x) = 3f_1(x) = 3\|x\|^2 + 9\sin^2(x_1) + 6\sin^2(x_2)$$

$$f_3(x) = \|x\|^2$$

$$f_4(x) = 3f_3(x) = 3\|x\|^2$$

$$f_5(x) = \log(1 + e^{-x_1 - 2x_2}) + \|x\|^2$$

$$f_6(x) = 3f_5(x) = 3\log(1 + e^{-x_1 - 2x_2}) + 3\|x\|^2.$$
although different objectives have different performances (but all are linearly convergent).

**Example 6:** Consider the logistic regression problem, where the objective is 
\[ f(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)) + \frac{\lambda}{2n} \|x\|^2 \] with a standard L2-regularizer [35], and \( a_i \in \mathbb{R}^d \) and \( b_i \in \{-1, +1\} \) are the data samples.

To test the performance of scaled SIGNSGD, the epsilon dataset with \( n = 400,000 \) and \( d = 2,000 \) is exploited [37], and the baseline is calculated using the standard optimizer LogisticSGD of scikit-learn [38]. To marginalize out the effect of initial choices, the numerical result is averaged over repeated runs with \( x_0 \approx \mathcal{N}(0, I) \). We compare scaled SIGNSGD with vanilla SGD, SIGNSGD, SIGNSGDM, and EF-SIGNSGD [20], as shown in Fig. 4 on a platform with the Intel Core i7-4300U CPU. Fig. 4 shows that SIGNSGD has a similar performance to SGD and performs better than SIGNSGD and SIGNSGDM. It can be also observed that EF-SIGNSGD is comparable with SGD, which is consistent with the discussion in Remark 2. Moreover, the case in Fig. 4(a) with a constant learning rate converges faster than that in Fig. 4(b) with a diminishing learning rate. Meanwhile, Fig. 5 shows that more workers can improve the performance. Therefore, the numerical results support our theoretical findings.

**VII. CONCLUSION**

This article has investigated faster convergence of scaled SIGNSGD/SGD, which can relieve the communication cost compared with vanilla SGD. To further motivate the study of sign methods, continuous-time algorithms have been addressed, indicating that sign SGD can significantly improve the convergence speed of SGD. Subsequently, it has been proven that scaled SIGNSGD is linearly convergent for both strongly convex and nonconvex (satisfying the PL inequality) objectives. Also, the convergence for SIGNSGD has been analyzed in two cases with constant and decaying learning rates. The results are also extended to the distributed setting in the parameter-server framework. The efficacy of scaled sign methods has been validated by numerical experiments.

**ACKNOWLEDGMENT**

The authors are grateful to the Editor, Associate Editor, and anonymous reviewers for their insightful suggestions. The authors would also like to thank Mr. Yang Yu for his assistance in numerical experiments.

**APPENDIX**

**A. Further Motivations for SIGNSGD**

Let us provide more evidences for studying sign-based GD from the continuous-time perspective. In doing so, consider the continuous-time dynamics corresponding to the discrete-time GD and SIGNSGD, i.e.,

\[ \dot{x} = -\beta \nabla f(x) \] (23)

\[ \dot{x} = -\beta \text{sign}(\nabla f(x)) \] (24)

where \( \beta > 0 \) is a constant learning rate.

To proceed, let us construct a Lyapunov candidate as

\[ V(t) := f(x) - f^* \quad \forall t \geq 0 \] (25)

where \( f^* \) denotes the minimum value attained by \( f \).

For algorithms (23) and (24), the following results can be obtained. **Proposition 1:** For algorithm (23)
1) if $f$ is convex, then $V (t) \leq \frac{D^2_1 V(0)}{D^2_t + D^2 (0)^{\beta t}}$, where $D_1 := \max_{x, f(x) \leq f(x_0)} \min_{x' \in X'} \| x - x' \|$ with $X'$ being the set of minimizers;
2) if $f$ is nonconvex, then $\min_{x \in [0, t]} \| \nabla f(x(s)) \| \leq \sqrt{\frac{\beta}{D^2_t}}$. 

**Proposition 2:** For algorithm (24)

1) if $f$ is convex, then $V (t) \leq V(0) e^{-\frac{\beta t}{D_1}}$, where $D_2 := \max_{x, f(x) \leq f(x_0)} \min_{x' \in X'} \| x - x' \|_{\infty}$; 
2) if $f$ is nonconvex, then $\min_{x \in [0, t]} \| \nabla f(x(s)) \|_1 \leq \frac{D_2 \| \nabla f(x(t)) \|_1}{\beta t}$.

By the abovementioned results, one can easily observe that (24) with sign gradients converges faster than GD (23) in the continuous-time domain, indicating that the performance of GD can be largely improved by sign gradient compression. For instance, in the scenario with convex objectives, GD (23) is sublinearly convergent whereas SIGNGD (24) is linearly convergent. Thus, the abovementioned results provide a new perspective for showing advantages of SIGNGD compared with GD.

**B. Proof of Proposition 1**

Consider the case with convex objectives. In light of (23), it can be calculated that

$$\dot{V}(t) = \nabla f(x(t))^T \dot{x} = -\beta \| \nabla f(x(t)) \|^2 \leq 0 \tag{26}$$

which implies $V(t) \leq V(0)$ for all $t \geq 0$ and thus $f(x(t)) \leq f(x(0))$.

Meanwhile, invoking the convexity of $f$ yields

$$V(t) = f(x(t)) - f(x^*)$$

$$\leq \nabla f(x(t))^T (x(t) - x^*)$$

$$\leq \| \nabla f(x(t)) \| \cdot \| x(t) - x^* \|$$

$$\leq D_1 \| \nabla f(x(t)) \| \tag{27}$$

where the second inequality has used the Cauchy–Schwarz inequality and the definition of $D_1$ as defined in Proposition 1 is employed for deriving the last inequality. Combining (27) with (26) gives rise to $\dot{V}(t) \leq -\frac{\beta}{D_1} V(t)^2$, further implying the claimed result.

For the case with nonconvex objectives, by integrating (26) from 0 to $t$, one can obtain that

$$\beta \int_0^t \| \nabla f(x(s)) \|^2 ds = V(0) - V(t)$$

$$= f(x(0)) - f(x(t))$$

$$\leq f(x(0)) - f^* \tag{28}$$

where the inequality has employed the fact that $f(z) \geq f^*$ for all $z \in \mathbb{R}^d$. Then, taking the minimum of $\| \nabla f(x(s)) \|$ over $[0, t]$ ends the proof.

**C. Proof of Proposition 2**

Consider first the convex case. Similar to (26), it can be obtained that

$$\dot{V}(t) = -\beta \| \nabla f(x(t)) \|_1 \tag{29}.$$ 
Akin to (27), one has that

$$V(t) \leq \nabla f(x(t))^T (x(t) - x^*)$$

$$\leq \| \nabla f(x(t)) \|_1 \cdot \| x(t) - x^* \|_\infty$$

$$\leq D_2 \| \nabla f(x(t)) \|_1 \tag{30}$$

where the second inequality has used Holder’s inequality and the last inequality has applied the definition of $D_2$ defined in Proposition 2. Combining (29) with (30) yields $\dot{V}(t) \leq -\frac{\beta}{D_1} V(t)$, from which it is easy to verify the claimed result.

Consider now the nonconvex case. The desired result can be obtained by (29) and the similar argument to that in the convex case. This completes the proof.

**D. Proof of Theorem 1**

To facilitate the subsequent analysis, define

$$V_k := f(x_k) - f^* \quad \forall k \geq 0. \tag{31}$$

In view of (13) and Assumption 3, it can be concluded that

$$V_{k+1} - V_k = f(x_{k+1}) - f(x_k)$$

$$\leq \nabla f(x_{k+1})^T (x_{k+1} - x_k) + \frac{L}{2} \| x_{k+1} - x_k \|_2^2$$

$$= -\alpha \| g_k \|_1^2 + \frac{L \alpha^2}{2} \| g_k \|_1 \cdot \| \text{sign}(g_k) \|_\infty^2$$

$$\leq -\alpha \| g_k \|_1^2 + \frac{L \alpha^2}{2} \| g_k \|_1^2$$

$$= -\gamma \| g_k \|_1^2 \tag{32}$$

where $\gamma = \alpha (1 - L \alpha/2)$ and the second inequality has used the fact $\| \text{sign}(g_k) \|_\infty \leq 1$.

In what follows, let us prove this theorem one by one. First, for case 1, invoking Assumption 1 yields

$$V_k \leq g_k^T (x_k - x^*) - \frac{\mu}{2} \| x_k - x^* \|_\infty^2$$

$$\leq \frac{1}{2} \left( \frac{\| g_k \|_1^2}{\mu} + \mu \| x_k - x^* \|_2^2 \right) - \frac{\mu}{2} \| x_k - x^* \|_\infty^2$$

$$= \frac{\| g_k \|_1^2}{2 \mu} \tag{33}$$

where the second inequality has employed the Holder inequality. Then, one has that $\| g_k \|_1^2 \geq 2 \mu V_k$. Therefore, in combination with (32), one can obtain that $V_{k+1} - V_k \leq -2 \mu \gamma V_k$, further leading to $V_{k+1} \leq \zeta V_k$, where $\zeta = 1 - 2 \mu \gamma$. Consequently, by iteration, this completes the proof of case 1.

Second, for case 2, in view of (31), invoking Assumption 2 with $x = x_k$ leads to $2 \mu V_k \leq \| g_k \|_1^2$, which, together with the similar argument to case 1, follows the conclusion in this case.
Third, for case 3, invoking (32) gives \( \|g_k\|^2 \leq V_k - V_{k+1} \), which, by summation over \( l = 0, 1, \ldots, k \), implies that
\[
\gamma \sum_{l=0}^{k} \|g_l\|^2 \leq V_0 - V_{k+1} = f(x_0) - f(x_{k+1}) \leq f(x_0) - f^*
\] (33)
where the last inequality has used the fact that \( f(x_{k+1}) \geq f^* \). Then, taking the minimum of \( \|g_l\|^2 \) over \( l = 0, 1, \ldots, k \) ends the proof. \( \square \)

E. Proof of Theorem 2
Recalling \( V_k \) in (31). Invoking Assumption 3 gives rise to
\[
V_{k+1} - V_k \leq \|g_k \|^2 (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2_\infty = -\alpha \|\tilde{g}_k\|_1^2 \|\text{sign}(\tilde{g}_k)\|^2 + \frac{L\alpha^2}{2} \|\tilde{g}_k\|^2_1 \|\text{sign}(\tilde{g}_k)\|^2_1 \leq -\alpha \|\tilde{g}_k\|_1^2 \|\text{sign}(\tilde{g}_k)\|^2 + \frac{L\alpha^2}{2} \|\tilde{g}_k\|^2_1 \] (34)
where the last inequality has used the fact \( \|\text{sign}(\tilde{g}_k)\|_1 \leq 1 \). By taking the conditional expectation on \( x_k \), one has
\[
E(V_{k+1}|x_k) - V_k \leq -\alpha \|\tilde{g}_k\|_1^2 E(\|\tilde{g}_k\|_1^2 \|\text{sign}(\tilde{g}_k)\|_1^2 \|x_k\) \]+ \frac{L\alpha^2}{2} E(\|\tilde{g}_k\|_1^2 \|x_k\). (35)

Consider now the coordinate \( \tilde{g}_{k,i} \) for \( i \in [d] \). One has that
\[
E(\|\tilde{g}_k\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) = E[E(\|\tilde{g}_k\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) \]
\[
= E[\tilde{g}_k\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) - \|\tilde{g}_k\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) \]
\[
E[g_{k,i} \|\tilde{g}_k\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) \]
\[
= (2p_{k,i} - 1) \|\tilde{g}_{k,i}\|_1 \|\text{sign}(\tilde{g}_{k,i})\|_1 \|x_k\) \] (36)
where the first equality is obtained by the tower property of conditional expectation and the second equality is derived using the fact that \( \|\tilde{g}_{k,i}\|_1 \) is \( \tilde{g}_{k,i} \)-measurable.

The abovementioned equality, together with (35), implies that
\[
E(V_{k+1}|x_k) - V_k \leq -\alpha \sum_{i=1}^{d} (2p_{k,i} - 1) \|g_{k,i}\|_1 E(\|\tilde{g}_k\|_1^2 \|x_k\) \]+ \frac{L\alpha^2}{2} E(\|\tilde{g}_k\|_1^2 \|x_k\). (37)

By Jesen’s inequality in conditional expectation for the function \( \|g_l\|^2 \), it follows that \( E(\|\tilde{g}_k\|_1^2 \|x_k\) \geq E(\|\tilde{g}_k\|_1^2 \|x_k\) \), where (17) has been utilized to obtain the equality. Because \( p_{k,i} \geq p_{min} \) for \( i \in [d] \), taking the expectation on (37) implies that
\[
E(V_{k+1}) - E(V_k) \leq -\alpha (2p_{min} - 1) E(\|g_k\|^2) \]
\[
+ \frac{L\alpha^2}{2} E(\|g_k\|^2 - \|\tilde{g}_k\|^2_1) \]
\[
\leq -\alpha (2p_{min} - 1 - L\alpha) E(\|g_k\|^2) + L\alpha^2 \alpha^2 \] (38)
where the second term in the first inequality has applied the fact \( \|x + y\|^2 \leq (\|x\|^1 + |y|^1)^2 \leq 2(\|x\|^2 + \|y\|^2) \) for all \( x, y \in \mathbb{R}^d \), and (17) is leveraged to obtain the second inequality.

Now, under Assumptions 1 or 2, using the similar argument to the proof of Theorem 1 can both lead to that \( E(\|g_k\|^2) \geq 2\mu E(V_k) \), which, together with (38) and \( \alpha \in (0, (2p_{min} - 1)/L \), yields that
\[
E(V_{k+1}) \leq E(V_k) \leq \zeta I E(V_k) + L\sigma^2 \alpha^2 \] (39)
where \( \zeta_1 = 1 - 2\mu (2p_{min} - 1 - L\alpha) \).

Iteratively applying the abovementioned inequality leads to (18).

It remains to show (19). Invoking the similar analysis for (39) yields that
\[
E(V_{k+1}) \leq \Pi_{l=0}^{k-1} \zeta \leq c_k E(V_0) + L\sigma^2 \alpha^2 \]
where \( c_k := 1 - \alpha \mu (2p_{min} - 1) = 1 - 3/(k+1) \) (note that \( \alpha_k = 3/(\mu(2p_{min} - 1)/(k+1)) \)), further implying that
\[
E(V_k) \leq \Pi_{l=0}^{k-1} c_l E(V_0) + L\sigma^2 \]
\[
\leq \Pi_{l=0}^{k-1} c_l E(V_0) + \frac{9}{\mu^2 (2p_{min} - 1)^2} \] (40)
where the second inequality has employed the expression of \( \alpha_k \) and \( c_k \) and the fact \( 1 - a \leq e^{-a} \) for all \( a \geq 0 \).

For the last two terms in (40), in light of the fact that \( \Pi_{l=0}^{n} (1 - a_l) \leq e^{-\sum_{l=0}^{n} a_l} \) for \( a_l \in [0, 1] \), one has that
\[
\Pi_{l=0}^{k-1} c_l \leq e^{-3 \sum_{l=0}^{k-1} \frac{1}{(k+1)^3}} \leq e^{-3 \int_{k=1}^{k+1} \frac{1}{x^3} dx} = e^{-3\ln(k+1)} = \frac{1}{(k+1)^3} \] (41)
and
\[
\sum_{m=1}^{k-1} e^{-\sum_{l=m}^{k-1} \frac{1}{(k+1)^3}} m^2 \leq \sum_{m=1}^{k-1} \frac{(m+1)^3}{(k+1)^3} \frac{1}{m^2} \leq \sum_{m=1}^{k-1} \frac{8m^3}{(k+1)^3 m^2} \]
\[
= \frac{8}{(k+1)^3} \sum_{m=1}^{k-1} \frac{k-1}{(k+1)^3} \frac{4k(k-1)}{(k+1)^3} \]
\[
\leq \frac{4k(k-1)}{k^2(k-1)} = \frac{4}{k} \] (42)
where the first inequality in (42) is similarly obtained to (41) and the second inequality in (42) is due to \( m+1 \leq 2m \) for all \( m \geq 1 \).
Then, inserting (41) and (42) into (40) leads to the conclusion (19). The proof is complete. □

F. Proof of Theorem 3

To ease the exposition, define \( \hat{g}_k := \{ \hat{g}_k^m, m \in [M] \} \). Similar to (34), invoking Assumption 3 and Algorithm 3 yields

\[
V_{k+1} - V_k \leq -\alpha M_k g_k^T \text{sign}(\hat{g}_k^m) + \frac{L\alpha^2}{2} M_k^2
\]

which, by taking the conditional expectation, implies that

\[
\mathbb{E}(V_{k+1}|x_k) - V_k \leq -\alpha g_k^T \mathbb{E}(M_k \text{sign}(\hat{g}_k^m)|x_k) + \frac{L\alpha^2}{2} \mathbb{E}(M_k^2|x_k).
\]

(43)

For \( g_k^T \mathbb{E}(M_k \text{sign}(\hat{g}_k^m)|x_k) \) in (43), similar to (36), one has

\[
\mathbb{E}(M_k \text{sign}(\hat{g}_k^m)|x_k) = g_k^T \mathbb{E}(M_k \text{sign}(\hat{g}_k^m)|\hat{g}_k^m|x_k)
\]

\[
= \mathbb{E} \left( M_k \sum_{i=1}^d g_{k,i} \mathbb{E}(\text{sign}(\hat{g}_{k,i}^m)|\hat{g}_k^m|x_k) \right)
\]

\[
= \mathbb{E} \left( \sum_{i=1}^d |g_{k,i}| \mathbb{E}(\text{sign}(\hat{g}_{k,i}^m)|\hat{g}_k^m|x_k) \right)
\]

\[
= \mathbb{E} \left( \sum_{i=1}^d |g_{k,i}| (2I_{p_{k,i}}(\kappa, \kappa) - 1) \right)
\]

\[
\geq (2I_{p_{\min}}(\kappa, \kappa) - 1) \| \hat{g}_k^m \|_1 \mathbb{E}(M_k|x_k)
\]

(44)

where the last equality has exploited [28, Lemma 7], and the inequality comes from the fact that \( p_{k,i} \geq p_{\min} \) for \( i \in [d], k \geq 0 \).

As for the last term in (44), by the definition of \( M_k \), it can be concluded that

\[
\mathbb{E}(M_k|x_k) = \mathbb{E} \left( \frac{1}{M} \sum_{m=1}^M \| \hat{g}_k^m \|_1 |x_k) \right)
\]

\[
\geq \mathbb{E} \left( \| \frac{1}{M} \sum_{m=1}^M \hat{g}_k^m \|_1 |x_k \right)
\]

\[
\geq \mathbb{E} \left( \frac{1}{M} \sum_{m=1}^M \| \hat{g}_k^m |x_k \right) \|_1
\]

\[
= \| g_k \|_1
\]

where the last inequality has applied the Jesen’s inequality in conditional expectation for the function \( \| . \|_1 \).

By substituting the abovementioned inequality and (44) into (43) and then taking the expectation, it can be obtained that

\[
\mathbb{E}(V_{k+1}) - \mathbb{E}(V_k) \leq -\alpha (2I_{p_{\min}}(\kappa, \kappa) - 1) \mathbb{E}(\| g_k \|_1^2) + \frac{L\alpha^2}{2} \mathbb{E}(M_k^2).
\]

(45)

Now, for the last term in (45), one has

\[
\mathbb{E}(M_k^2) \leq \frac{1}{M} \sum_{m=1}^M \mathbb{E}(\| \hat{g}_k^m \|_1^2)
\]

\[
\leq 2 \frac{M}{M} \sum_{m=1}^M \mathbb{E}(\| g_k^m - g_k \|_1^2) + 2 \mathbb{E}(\| g_k \|_1^2)
\]

\[
\leq 2\alpha^2 + 2 \mathbb{E}(\| g_k \|_1^2)
\]

(46)

where the first inequality has used the fact \( \sum_{m=1}^M \| g_k^m \|_1^2 \leq M \sum_{m=1}^M \| g_k^m \|_1^2 \), the second inequality has employed the fact that \( \| x + y \|_1 \leq 2 \| x \|_1^2 + \| y \|_1^2 \) for all \( x, y \in \mathbb{R}^d \), and the last inequality is obtained by using \( \mathbb{E}(\| g_k^m - g_k \|_1^2) \leq \alpha^2 \) in Assumption 4.

Combining (46) with (45) yields

\[
\mathbb{E}(V_{k+1}) - \mathbb{E}(V_k) \leq -\alpha (2I_{p_{\min}}(\kappa, \kappa) - 1 - L\alpha) \mathbb{E}(\| g_k \|_1^2) + L\alpha^2\alpha^2.
\]

The rest of the proof is similar to that after (38). This ends the proof. □

REFERENCES

[1] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge U.K.: Cambridge Univ. Press, 2004.
[2] J. Shen, S. Duszcz, and A. Khaligh, “Optimization of sizing and battery cycle life in battery/ultracapacitor hybrid energy storage systems for electric vehicle applications,” IEEE Trans. Ind. Inform., vol. 10, no. 4, pp. 2112–2121, Nov. 2014.
[3] X. Li, X. Yi, and L. Xie, “Distributed online optimization for multi-agent networks with coupled inequality constraints,” IEEE Trans. Autom. Control, vol. 66, no. 8, pp. 3575–3591, Aug. 2021.
[4] W. Su, H. Eichi, W. Zeng, and M.-Y. Chow, “A survey on the electrification of transportation in a smart grid environment,” IEEE Trans. Ind. Inform., vol. 8, no. 1, pp. 1–10, Feb. 2012.
[5] Y. Wang, S. Liu, B. Sun, and X. Li, “A distributed proximal primal-dual algorithm for energy management with transmission losses in smart grid,” IEEE Trans. Ind. Inform., vol. 18, no. 11, pp. 7608–7618, Nov. 2022.
[6] S. Messaoud, A. Bradaï, and E. Moulay, “Online GMM clustering and mini-batch gradient descent based optimization for industrial IoT 4.0,” IEEE Trans. Ind. Inform., vol. 16, no. 2, pp. 1427–1435, Feb. 2020.
[7] X. Luo, H. Wu, Z. Wang, J. Wang, and D. Meng, “A novel approach to large-scale dynamically weighted directed network representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9756–9773, Dec. 2022.
[8] X. Luo, Y. Zhou, Z. Liu, and M. Zhou, “Fast and accurate non-negative latent factor analysis on high-dimensional and sparse matrices in recommender systems,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 4, pp. 3897–3911, Apr. 2023, doi: 10.1109/TKDE.2021.3125252, 2021.
[9] S. Rudor, “An overview of gradient descent optimization algorithms,” 2016, arXiv:1609.04747.
[10] M. Meng and X. Li, “Aug-PDG: Linear convergence of convex optimization with inequality constraints,” Control Theory & Appl., vol. 39, no. 10, pp. 1969–1977, 2022.
[11] X. Li, L. Xie, and Y. Hong, “Distributed aggregative optimization over multi-agent networks,” IEEE Trans. Automat. Control, vol. 67, no. 6, pp. 3165–3171, June. 2022.
[12] L. Bottou, “Stochastic gradient descent tricks,” in Neural Networks: Tricks of the Trade. Berlin, Germany: Springer, 2012, pp. 421–436.
[13] S. Bonnabel, “Stochastic gradient descent on Riemannian manifolds,” IEEE Trans. Automat. Control, vol. 58, no. 9, pp. 2217–2229, Sep. 2013.
[14] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” in Proc. Int. Conf. Neural Networks, 1993, pp. 586–591.
[15] J. Bernstein, Y.-X. Wang, K. Azizicusneshi, and A. Anandkumar, “signSGD: Compressed optimisation for non-convex problems,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 560–569.
[16] L. Balleses, F. Pedregosa, and N. L. Roux, “The geometry of sign gradient descent,” 2020, arXiv:2002.08056.
[17] L. Balles and P. Henning, “Dissecting Adam: The sign, magnitude and variance of stochastic gradients,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 404–413.
[18] J. Hamer, M. Mohri, and A. T. Suresh, “FedBoost: A communication-efficient algorithm for federated learning,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 3973–3983.
Xiuxian Li (Senior Member, IEEE) received the B.S. degree in mathematics and applied mathematics and the M.S. degree in pure mathematics from Shandong University, Jinan, China, in 2009 and 2012, respectively, and the Ph.D. degree in mechanical engineering from the University of Hong Kong, Hong Kong, in 2016.

From 2016 to 2020, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. In 2020, he joined Tongji University, Shanghai, China, where he is currently a Professor. His research interests include distributed control and optimization, game theory, machine learning, and applications to autonomous vehicles.

Kuo-Yi Lin received the B.S. degree in statistics from Cheng Kung University, Taiwan, China, in 2007 and the M.S. degree in industrial engineering and management from Tsing Hua University, Taiwan, China, in 2009 and 2014, respectively.

He is a Professor with the College of Business, Guilin University of Electronic Technology, Guilin, China. His research interests include intelligent manufacturing and federated learning.

Dr. Lin is the Director of China Excellent Business Decision Making Society, Executive Committee Members of Intelligent Simulation Optimization and Scheduling Committee of China Simulation Society, and a Member of Natural Computing and Digital Intelligent City Committee of China Artificial Intelligence Society.

Li Li (Member, IEEE) received the B.S. and M.S. degrees in electrical automation from Shenyang Agriculture University, Shenyang, China, in 1996 and 1999, respectively, and the Ph.D. degree in mechatronics engineering from the Shenyang Institute of Automation, Chinese Academy of Science, Shenyang, in 2003.

She then joined Tongji University, Shanghai, China, where she is currently a Professor of control theory and engineering. She has authored or coauthored more than 50 publications, including five books, over 30 journal papers, and two book chapters. Her current research interests include unmanned systems, data-driven modeling and optimization, and energy systems.

Yiguang Hong (Fellow, IEEE) received the B.S. and M.S. degrees in mechanics from Peking University, Beijing, China, in 1987 and 1990, respectively, and the Ph.D. degree in operations research and cybernetics from the Chinese Academy of Sciences (CAS), Beijing, China, in 1993.

He is currently a Professor with Tongji University, Shanghai, China. Before October 2020, he was a Professor with the Academy of Mathematics and Systems Science, CAS. His current research interests include nonlinear control, multigant systems, distributed optimization and game, machine learning, and social networks.

Dr. Hong is a Fellow of Chinese Association for Artificial Intelligence and Chinese Association of Automation (CAA). In addition, he is the Chair of Technical Committee of Control Theory of CAA and was a Board of Governor of IEEE Control Systems Society. He is the Editor-in-Chief of Control Theory and Technology. He is or was the Associate Editor for many journals including the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, and IEEE CONTROL SYSTEMS MAGAZINE. He was the recipient of the Guan Zhaozhi Award at the Chinese Control Conference, Young Author Prize of the IFAC World Congress, Young Scientist Award of CAS, the Youth Award for Science and Technology of China, and the National Natural Science Prize of China.

Jie Chen (Fellow, IEEE) received the B.Sc., M.Sc., and the Ph.D. degrees in control theory and control engineering from the Beijing Institute of Technology, Beijing, China, in 1986, 1996, and 2001, respectively.

From 2018 to 2023, he was the President of Tongji University, Shanghai, China. He is currently a Professor with the Department of Control Science and Engineering, Tongji University and Beijing Institute of Technology, where he is the Director of the National Key Laboratory of Autonomous Intelligent Unmanned Systems. His research interests include complex systems, multigant systems, multiobjective optimization and decision, and constrained nonlinear control.

Dr. Chen is currently the Editor-in-Chief of Unmanned Systems and the Journal of Systems Science and Complexity. He was on the editorial boards of several journals, including the IEEE TRANSACTIONS ON CYBERNETICS, International Journal of Robust and Nonlinear Control, and Science China Information Sciences. He is a Fellow of IFAC and a Member of the Chinese Academy of Engineering.