DISTRIBUTED RELATIVELY SMOOTH OPTIMIZATION

Sofia Jegnell and Stefan Vlaski

Department of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

Smoothness conditions, either on the cost itself or its gradients, are ubiquitous in the development and study of gradient-based algorithms for optimization and learning. In the context of distributed optimization and multi-agent systems, smoothness conditions and gradient bounds are additionally central to controlling the effect of local heterogeneity. We deviate from this paradigm and study distributed learning problems in relatively smooth environments, where cost functions may grow faster than a quadratic, and gradients need not be bounded. We generalize gradient noise conditions to cover this setting, and present convergence guarantees in relatively smooth and relatively convex environments. Numerical results corroborate the findings.

***Index Terms*—** Distributed learning, relative smoothness, federated learning, mirror descent, stochastic optimization.

1. INTRODUCTION AND RELATED WORKS

We consider a collection of K agents, where each agent k is equipped with a distinct, local cost function $J_k(w)$. We define the aggregate optimization problem:

$$J(w) \triangleq \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{1}$$

Our aim is to pursue an optimal solution to this global optimization problem in the sense that:

$$w^o \triangleq \arg\min_w J(w) = \arg\min_w \frac{1}{K}\sum_{k=1}^{K} J_k(w) \tag{2}$$

In the absence of communication constraints, one may pursue w^o by means of gradient descent:

$$w_i = w_{i-1} - \mu\nabla J(w_{i-1}) \tag{3}$$

Classical convergence guarantees for gradient descent are derived under smoothness conditions on the gradient of the form:

$$\|\nabla J(x) - \nabla J(y)\| \leq \delta\|x-y\| \tag{4}$$

Emails: { anna.jegnell18, s.vlaski}@imperial.ac.uk

For smooth loss functions satisfying (4), one can establish sublinear and linear convergence for convex and strongly-convex loss functions respectively [1]. Condition (4) can be equivalently formulated as:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \frac{\delta}{2}\|x-y\|^2 \tag{5}$$

When the cost $J(w)$ is not differentiable, one may instead resort to subgradient recursions, where sublinear convergence is classically established under Lipschitz conditions on the cost $J(\cdot)$ itself, rather than its (sub)gradient [2]. Such conditions are equivalent to assuming uniformly bounded (sub)gradients. In this work, we are interested in optimizing aggregate cost functions, which are differentiable, but do not satisfy the smoothness conditions (4)–(5). Instead, we will consider the recently introduced, and more general, relative smoothness condition [3, 4]:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \delta D_h(y,x) \tag{6}$$

Here, $D_h(y,x)$ denotes the Bregman divergence:

$$D_h(y,x) = h(y) - h(x) - \nabla h(x)^{\mathsf{T}}(y-x) \tag{7}$$

where $h(\cdot)$ is a 1-strongly convex proximity function ensuring [4]:

$$D_h(y,x) \geq \frac{1}{2}\|x-y\|^2 \tag{8}$$

Condition (6) is a direct generalization of (5). Indeed, setting $h(x) \triangleq \frac{1}{2}\|x\|^2$ yields $D_h(y,x) = \frac{1}{2}\|x-y\|^2$ and recovers (5) from (6) exactly. We note that while the analogy between $D_h(y,x)$ and the Euclidean distance $\frac{1}{2}\|x-y\|^2$ can provide useful intuition, $D_h(y,x)$ is not in fact a distance and as such does not satisfy several critical properties. For example, it is not symmetric in general, and does not satisfy the triangle inequality. We refer the reader to [3, 4] for a detailed discussion on relative smoothness, as well as a collection of examples. It is advocated in [3, 4], to pursue minimizers of relatively smooth cost functions by means of the generalized gradient scheme:

$$w_i \triangleq \arg\min_w \nabla J(w_{i-1})^{\mathsf{T}}(w-w_{i-1}) + \frac{1}{\mu}D_h(w,w_{i-1}) \tag{9}$$

For $\mu \leq \frac{1}{\delta}$, the right-hand side of (9) is, in light of (6), an upper bound of $J(w)$ around $J(w_{i-1})$, and hence we can view (9) as a majorization-minimization scheme. Again, by setting $h(x) = \frac{1}{2}\|x\|^2$, we recover the ordinary gradient algorithm (3). We note that the generalized gradient scheme (9) is significantly older than the notion of relative smoothness introduced in [3, 4]. Indeed, it has been introduced much earlier under the name mirror-descent [5] and has since been studied extensively including composite and accelerated [6] as well as stochastic subgradient-based variants [7]. While this line of work illustrates the strength and flexibility of the generalized gradient scheme (9), the focus there has not been on relatively smooth functions satisfying (6), and the results generally rely on Lipschitz conditions on the gradient [6] or uniformly bounded stochastic (sub)gradients [7]. More recent works have provided convergence guarantees for centralized stochastic mirror descent in more general settings, focusing on relative continuity [8,9] and relative smoothness with uniformly bounded gradient noise variance [10]. In contrast, we will develop distributed schemes for the optimization of relatively smooth functions by generalizing the results of [3,4] to a federated setting.

1.1. Distributed Optimization

So far, we have disregarded communication constraints, and have focused on the centralized mode of operation, where a full gradient $\nabla J(\cdot)$ is evaluated at every iteration, which requires central aggregation of gradients $\nabla J_k(w)$ from all agents. Distributed algorithms avoid full central aggregation and can be broadly classified into two architectures. Federated approaches employ central aggregation of *partial* information from the network, for example by probing a random subset of agents at any given iteration, and averaging their gradients or update to obtain a stochastic estimate of the true gradient recursion (3) [11, 12]. Decentralized approaches avoid central aggregation altogether, and instead rely on local (stochastic) gradient updates followed by peer-to-peer interactions [13–20]. Distributed algorithms based on mirror descent have been introduced as well, and can be decomposed based on whether they involve the exchange of local primal estimates $w_{k,i}$ [21,22] or the dual counterpart $\nabla h(w_{k,i})$ [23–25]. All works on distributed mirror descent employ Lipschitz conditions, either on the gradients or the cost itself (implying a uniform gradient bound). As such, they do not apply to the relatively smooth setting (6) considered here. The aim of this work is to provide a framework and convergence guarantee for distributed optimization of relatively smooth functions. To this end, we will introduce new, relaxed gradient noise conditions, which are necessary to study the convergence of distributed relatively smooth optimization algorithms, and provide convergence analysis under this condition in the federated setting.

2. PROBLEM AND ALGORITHM FORMULATION

We return to the aggregate optimization problem (1) and consider a federated setting. At each iteration i, the server selects a subset $\mathcal{L}_i$ of L agents (sampled without replacement), each with equal probability, and provides them with the current model $\boldsymbol{w}_{i-1}$. Note that we now employ boldface notation to emphasize the fact that the models $\boldsymbol{w}_{i-1}$, as a result of random agent participation, will be random themselves. Each participating agent $k \in \mathcal{L}_i$ then performs a generalized gradient update (9) *along its local gradient* obtaining to find $\boldsymbol{w}_{k,i}$:

$$\boldsymbol{w}_{k,i} = \arg\min_{w} \nabla J_k(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w - \boldsymbol{w}_{i-1}) + \frac{D_h\left(w, \boldsymbol{w}_{i-1}\right)}{\mu} \tag{10}$$

The optimality conditions (10) yield the well-known equivalent representation in the dual mirror domain [5]:

$$\nabla h(\boldsymbol{w}_{k,i}) = \nabla h(\boldsymbol{w}_{i-1}) - \mu \nabla J_k(\boldsymbol{w}_{i-1}) \tag{11}$$

As in [23, 24], the agents send the dual estimates back to the server, where they are aggregated as:

$$\begin{aligned}\nabla h(\boldsymbol{w}_i) &= \frac{1}{L}\sum_{k\in\mathcal{L}_i} \nabla h(\boldsymbol{w}_{k,i}) \\ &= \frac{1}{L}\sum_{k\in\mathcal{L}_i} \nabla h(\boldsymbol{w}_{i-1}) - \mu\frac{1}{L}\sum_{k\in\mathcal{L}_i} \nabla J_k(\boldsymbol{w}_{i-1}) \\ &= \nabla h(\boldsymbol{w}_{i-1}) - \mu\frac{1}{L}\sum_{k\in\mathcal{L}_i} \nabla J_k(\boldsymbol{w}_{i-1}) \end{aligned} \tag{12}$$

Inverting the same optimality argument that led to (11), we can conclude equivalently that:

$$\boldsymbol{w}_i = \arg\min_{w} \widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w - \boldsymbol{w}_{i-1}) + \frac{1}{\mu} D_h\left(w, \boldsymbol{w}_{i-1}\right) \tag{13}$$

where we defined the stochastic gradient approximation:

$$\widehat{\nabla J}(\boldsymbol{w}_{i-1}) \triangleq \frac{1}{L}\sum_{k\in\mathcal{L}_i} \nabla J_k(\boldsymbol{w}_{i-1}) \tag{14}$$

This insight provides justification for exchanging and averaging the mirror maps $\nabla h(\boldsymbol{w}_{k,i})$, rather than the primal estimates $\boldsymbol{w}_{k,i}$. In this way, the distributed algorithm can be viewed as a centralized mirror descent algorithm, which utilizes a stochastic estimate of the gradient defined in (14). This fact will be central to the convergence analysis that follows. It does not, however, allow us to apply results from the literature on centralized stochastic mirror descent [7–10]. This is because the conditions on the gradient approximation employed in [7–10], are in general not satisfied by the gradient approximation $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$, resulting from federated sampling of relatively smooth functions. We will illustrate this in the sequel, and proceed to present an alternative gradient noise conditions for distributed optimization along with performance guarantees.

3. ANALYSIS

3.1. Modeling Conditions

We introduce the gradient noise term:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = \widehat{\nabla J}(\boldsymbol{w}_{i-1}) - \nabla J(\boldsymbol{w}_{i-1}) \tag{15}$$

We can then immediately verify that:

$$\mathbb{E}\left\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\middle|\,\boldsymbol{w}_{i-1}\right\} = 0 \tag{16}$$

The critical quantity, which determines performance in most stochastic optimization algorithms, is then the variance $\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\middle|\,\boldsymbol{w}_{i-1}\right\}$. We propose the following condition:

Assumption 1 (Relative Gradient Noise Variance Bound). *The gradient noise process $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ satisfies the variance bound:*

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\middle|\,\boldsymbol{w}_{i-1}\right\} \leq 2\beta^2 D_h(w^o, \boldsymbol{w}_{i-1}) + \sigma^2 \tag{17}$$

for some $\beta, \sigma \geq 0$. □

The condition merits some discussion and comparison with related works. For $\beta^2 = 0$, we recover:

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\middle|\,\boldsymbol{w}_{i-1}\right\} \leq \sigma^2 \tag{18}$$

which corresponds to the most classical condition on gradient noise variance, employed for example in [7,10]. In the special case where $\beta, \sigma > 0$, but $h(\cdot) = \frac{1}{2}\|\cdot\|^2$, we recover:

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\middle|\,\boldsymbol{w}_{i-1}\right\} \leq \beta^2\|w^o - \boldsymbol{w}_{i-1}\,\|^2 + \sigma^2 \tag{19}$$

which corresponds to the relative gradient noise bounds studied in [14, 15]. Allowing for a relative component $\beta^2\|w^o - \boldsymbol{w}_{i-1}\,\|^2$ or $2\beta^2 D_h(w^o, \boldsymbol{w}_{i-1}) + \sigma^2$ turns out to be critical to allow for gradient noise components to grow in variance away from the minimum w^o. This is necessary, for example, in the case of least mean squares [15].

Comparison with the condition employed in [8, 9], requires slight reformulation. The authors there impose:

$$\mathbb{E}\left\{\|\widehat{\nabla J}(\boldsymbol{x})\|^2\middle|\,\boldsymbol{x}\right\} \leq G^2\,\frac{D_h(y,\boldsymbol{x})}{\frac{1}{2}\|y-\boldsymbol{x}\,\|^2} \tag{20}$$

for some $G > 0$ and all $\boldsymbol{x}, y$. By conditional independence, we have:

$$\mathbb{E}\left\{\|\widehat{\nabla J}(\boldsymbol{x})\|^2\middle|\,\boldsymbol{x}\right\} = \|\nabla J(\boldsymbol{x})\|^2 + \mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{x})\|^2\middle|\,\boldsymbol{x}\right\} \tag{21}$$

Hence, relation (20) imposes a bound on both the gradient norm $\|\nabla J(\boldsymbol{x})\|^2$ and the gradient noise $\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{x})\|^2\middle|\,\boldsymbol{x}\right\}$. For general choices of $h(\cdot)$, the right-hand side of (20) may grow with $\boldsymbol{x}$. For $h(\cdot) = \frac{1}{2}\|\cdot\|^2$ the bound simplifies to:

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_i(\boldsymbol{x})\|^2\middle|\,\boldsymbol{x}\right\} \leq G^2 \tag{22}$$

and hence, in contrast to (17), allows for no relative growth of the gradient noise variance when $h(\cdot) = \frac{1}{2}\|\cdot\|^2$.

In addition to the gradient noise condition in Assumption 1, we formally state the relative smoothness and convexity conditions as assumptions [3,4]:

Assumption 2. *The global objective function is δ-smooth relative to $h(\cdot)$, i.e. for all x, y:*

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \delta D_h(y,x) \tag{23}$$

Additionally, we require the objective function to be ν-strongly convex relative to $h(\cdot)$, i.e. for all x, y:

$$J(y) \geq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \nu D_h(y,x) \tag{24}$$

3.2. Convergence Analysis

Our argument essentially follows that of [4], after accounting for the terms arising from the gradient noise and the fact that mirror descent in the presence of gradient perturbations is no longer a true descent method. The proof is provided in the appendix.

Theorem 1 (Convergence Guarantee). *Under Assumptions 1 and 2, we have:*

$$\begin{aligned}&\min_{n=0,\ldots,i} \mathbb{E}J(\boldsymbol{w}_n) - J(w^o)\\ &\leq \frac{1}{\mu\sum_{n=0}^{i-1}\gamma^{-n}} D_h(w^o, w_o) + \mu\sigma^2\end{aligned} \tag{25}$$

where $\gamma \triangleq 1 - \mu\nu + 2\mu^2\beta^2$.

Proof. Appendix A. □

We note that $\sum_{n=0}^{i-1}\gamma^{-n}$ forms a divergent geometric sum, and hence 1 implies linear convergence. To verify this, observe that:

$$\frac{1}{\sum_{n=0}^{i-1}\gamma^{-n}} = \frac{\gamma^{-1}-1}{\gamma^{-i}-1} = \frac{\gamma^{i-1}-\gamma^i}{1-\gamma^i} \approx \gamma^{i-1}\,(1-\gamma) \tag{26}$$

where the approximation is accurate for large i.

4. NUMERICAL EXAMPLE

We consider a regularized least-squares problem, where each agent is equipped with:

$$J_k(w) = \frac{1}{2}\|b_k - A_k w\|_2^2 + \frac{\rho_1}{2}\|w\|_2^2 + \frac{\rho_2}{4}\|w\|_4^4 \tag{27}$$

The aggregate objective is then given by:

$$J(w) = \frac{1}{2K}\sum_{k=1}^{K}\|b_k - A_k w\|_2^2 + \frac{\rho_1}{2}\|w\|_2^2 + \frac{\rho_2}{4}\|w\|_4^4 \tag{28}$$

One direct approach to establishing an appropriate proximity function for $J(w)$ in this particular case is to examine its Hessian matrix $\nabla^2 J(w)$ and verify the equivalent conditions $\nu\nabla^2 h(w) \preceq \nabla^2 J(w) \preceq \delta\nabla^2 h(w)$ [3, 4]. We have:

$$\nabla^2 J(w) = \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k + \rho_1 I + \frac{\rho_2}{4}\nabla^2 \|w\|_4^4 \tag{29}$$

If we set $\delta = \max\left\{\lambda_{\max}\left(\frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right) + \rho_1, \rho_2\right\}$ and $\nu = \min\left\{\lambda_{\min}\left(\frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right) + \rho_1, \rho_2\right\}$, it then follows that $J(w)$ is δ-smooth and ν-strongly convex relative to the proximity function:

$$h(w) = \frac{1}{2}\|w\|_2^2 + \frac{1}{4}\|w\|_4^4 \tag{30}$$

We conclude that $J(w)$ satisfies Assumption 2. To verify the gradient noise condition in Assumption 1 we observe:

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 | \boldsymbol{w}\right\} \\
&= \mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k\in\mathcal{L}} \nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\right\|^2 | \boldsymbol{w}\right\}
\end{aligned} \tag{31}$$

We introduce the participation indicator $\mathbb{1}_{k,i}$, indicating whether agent k has been chosen to participate at iteration i. If agent k participates, we have $\mathbb{1}_{k,i} = 1$, and $\mathbb{1}_{k,i} = 0$ otherwise. Since L agents are chosen at each iteration, we have $\sum_{k=1}^{K} \mathbb{1}_{k,i} = L$ with probability one, and since agents are chosen with equal probability, we have $\Pr\{\mathbb{1}_{k,i} = 1\} = \frac{L}{K}$. It then follows that:

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 | \boldsymbol{w}\right\} \\
&= \mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k\in\mathcal{L}} \nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\right\|^2 | \boldsymbol{w}\right\} \\
&= \mathbb{E}\left\{\left\|\frac{1}{L}\sum_{k=1}^{K} \mathbb{1}_{k,i}\left(\nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\right)\right\|^2 | \boldsymbol{w}\right\}
\end{aligned} \tag{32}$$

The difficulty in evaluating (32) is that, since agents are sampled *without replacement*, the participation indicators $\mathbb{1}_{k,i}$ are not pairwise independent. Nevertheless, following argument analogous to [26, Lemma 1], we can establish:

$$\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 | \boldsymbol{w}\right\} = \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\|\nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\|^2 \tag{33}$$

Relation (33) captures the interplay between the participation rate $\frac{K}{L}$ and the level of heterogeneity $\frac{1}{K}\sum_{k=1}^{K}\|\nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\|^2$. Finally, we bound the local heterogeneity:

$$\begin{aligned}
&\|\nabla J_k(\boldsymbol{w}) - \nabla J(\boldsymbol{w})\|^2 \\
&= \left\|A_k^{\mathsf{T}} A_k\,\boldsymbol{w} - \left(\frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right)\boldsymbol{w}\right\|^2 \\
&= \left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right)(\boldsymbol{w} - w^o + w^o)\right\|^2 \\
&\le 2\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right)\right\|^2 \|w^o - \boldsymbol{w}\|^2 \\
&\quad + 2\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right) w^o\right\|^2 \\
&\overset{(a)}{\le} 4\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right)\right\|^2 D_h(w^o, \boldsymbol{w}) \\
&\quad + 2\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right) w^o\right\|^2 \\
&\overset{(b)}{=} \beta_k^2 D_h(w^o, \boldsymbol{w}) + \sigma_k^2
\end{aligned} \tag{34}$$

where in (a) we made use of 1-strong convexity of $D_h(\cdot, \boldsymbol{w})$, and in (b) we introduced:

$$\beta_k^2 \triangleq 4\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right)\right\|^2 \tag{35}$$

$$\sigma_k^2 \triangleq 2\left\|\left(A_k^{\mathsf{T}} A_k - \frac{1}{K}\sum_{k=1}^{K} A_k^{\mathsf{T}} A_k\right) w^o\right\|^2 \tag{36}$$

Returning to (33), we conclude that the construction satisfies the proposed gradient noise condition in Assumption 1 with:

$$\beta^2 = \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\beta_k^2 \tag{37}$$

$$\sigma^2 = \frac{1}{KL}\frac{K-L}{K-1}\sum_{k=1}^{K}\sigma_k^2 \tag{38}$$

Note in particular that the relative component $\beta_k^2 D_h(w^o, \boldsymbol{w})$ in (34) is necessary to account for the unbounded growth in heterogeneity as $\boldsymbol{w}$ moves away from w^o.

4.1. Simulation Results

We simulate the numerical example for a collection of $K = 100$ agents with participation rate of $\frac{L}{K} = 10\%$. We set $\rho_1 = \rho_2 = 0.1$. The regressor matrix A_k, for each agent k, is generated by filling its rows with $N = 5$ normally distributed samples $a_k \sim \mathcal{N}(0, I_{10}) \in \mathbb{R}^{10}$, resulting in $A \in \mathbb{R}^{5\times 10}$. We similarly generate $w^{\text{true}} \in \mathcal{N}(0, I_{10})$ and measure:

$$d_k = A_k w^{\text{true}} + v_k \tag{39}$$

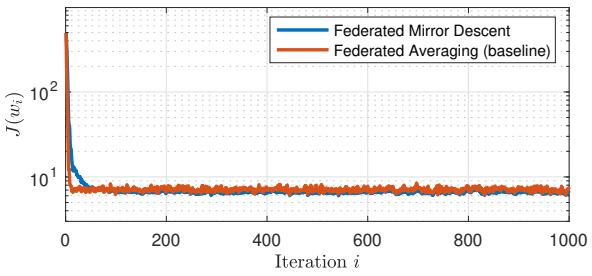

Fig. 1: Learning curve when initializing close to the origin ($b_v = 1$). Iterates remain close to the origin, where gradients are approximately Lipchitz continuous, resulting in comparable performance of both approaches.

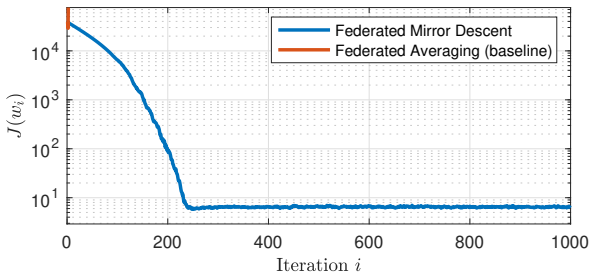

Fig. 2: Learning curve when initializing far from the origin ($b_v = 10$, all other parameters fixed). The scheme based on federated averaging and ordinary gradient descent diverges due to lack of local smoothness.

where $v_k \in \mathcal{N}(0, 0.1I_5)$. The step-size for all implementations is set to $\mu = 0.1$. All quantities are randomly sampled once prior to simulating the algorithm, and then kept fixed throughout. The only remaining randomness arises from the sampling of agents at each iteration. We average over 10 runs. We compare the federated mirror descent scheme (13) to the federated averaging algorithm [12] based on ordinary gradient updates as a baseline. We present two runs of the algorithm, differing only in the initialization:

$$w_0 = b_v \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^{\mathsf{T}} \in \mathbb{R}^M \tag{40}$$

In Fig. 1 we show a sample learning curve when both algorithms are initialized close to the origin ($b_v = 1$). We observe comparable performance between the schemes. This is consistent with empirical observations on the convergence of gradient-based algorithms in the absence of global smoothness conditions. In the vicinity of the origin, the gradients of the objective function are locally approximately Lipschitz continuous. As long as the gradients of the objective are locally smooth over a subset of the solutions space, and iterates of the algorithm do not leave this subset, one can expect gradient-based algorithms to perform well. We contrast this in Fig. 2, where both algorithms are initialized slightly further from the origin ($b_v = 10$). This change in initialization is sufficient to drive the gradient-based federated averaging scheme unstable, while the federated mirror descent algorithm (13) remains stable and exhibits comparable performance.

5. CONCLUSION

We have presented a linear convergence guarantee of a federated mirror descent algorithm in relatively smooth and relatively convex environments. To this end, we introduced a generalized condition on the gradient noise process, and quantified the effect of this noise on downstream algorithm performance. A numerical example illustrated the practical advantage of the proposed approach.

A. PROOF OF THEOREM 1

From (23), we have:

$$\begin{aligned}
&J(\boldsymbol{w}_i) \\
&\leq J(\boldsymbol{w}_{i-1}) + \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) + \delta D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) \\
&\leq J(\boldsymbol{w}_{i-1}) + \widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) + \delta D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) \\
&\quad + \left(\nabla J(\boldsymbol{w}_{i-1}) - \widehat{\nabla J}(\boldsymbol{w}_{i-1})\right)^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) \\
&\leq J(\boldsymbol{w}_{i-1}) + \widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) + \delta D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) \\
&\quad + (\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1})
\end{aligned} \tag{41}$$

We bound terms individually. Beginning with the term involving the gradient noise, we have:

$$\begin{aligned}
&(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) \\
&\overset{(a)}{\leq} \mu\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 + \frac{1}{4\mu}\|\,\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\,\|^2 \\
&\overset{(b)}{\leq} \mu\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 + \frac{1}{2\mu}D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1})
\end{aligned} \tag{42}$$

where (a) follows from Young's inequality and (b) follows from 1-strong convexity of $D_h(\cdot, \boldsymbol{w}_{i-1})$. Upon taking conditional expectation, we can further bound:

$$\begin{aligned}
&\mathbb{E}\left\{\left(\nabla J(\boldsymbol{w}_{i-1}) - \widehat{\nabla J}(\boldsymbol{w}_{i-1})\right)^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1})|\,\boldsymbol{w}_{i-1}\right\} \\
&\leq \mu\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\,\boldsymbol{w}_{i-1}\right\} \\
&\quad + \frac{1}{2\mu}\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1})|\,\boldsymbol{w}_{i-1}\right\} \\
&\leq 2\mu\beta^2 D_h(w^o, \boldsymbol{w}_{i-1}) + \mu\sigma^2 \\
&\quad + \frac{1}{2\mu}\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1})|\,\boldsymbol{w}_{i-1}\right\}
\end{aligned} \tag{43}$$

For the term involving the stochastic gradient, we employ Tseng's three-point-property [6, Property 1]. It states that,

if:

$$\boldsymbol{w}_i = \arg\min_{w} \phi(w) + D_h(w, \boldsymbol{w}_{i-1}) \tag{44}$$

then for all w:

$$\begin{aligned} &\phi(w) + D_h(w, \boldsymbol{w}_{i-1}) \\ &\geq \phi(\boldsymbol{w}_i) + D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) + D_h(w, \boldsymbol{w}_i) \end{aligned} \tag{45}$$

If we set $\phi(w) = \mu\widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w - \boldsymbol{w}_{i-1})$, then (44) holds in light of (13). After rearranging, we can conclude:

$$\begin{aligned} &\widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) \\ &\leq \widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) + \frac{1}{\mu} D_h(w^o, \boldsymbol{w}_{i-1}) \\ &\quad - \frac{1}{\mu} D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) - \frac{1}{\mu} D_h(w^o, \boldsymbol{w}_i) \end{aligned} \tag{46}$$

Upon taking expectations:

$$\begin{aligned} &\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\leq \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) + \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \frac{1}{\mu}\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} - \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} \end{aligned} \tag{47}$$

where we used the fact that:

$$\begin{aligned} &\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &= \mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\}^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) \\ &= \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) \end{aligned} \tag{48}$$

Putting everything together, we can bound (41) after taking conditional expectations and grouping terms as:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} \\ &\leq J(\boldsymbol{w}_{i-1}) + \mathbb{E}\left\{\delta D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_i\right\} \\ &\quad + \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) + \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \frac{1}{\mu}\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} - \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad + 2\mu\beta^2 D_h(w^o, \boldsymbol{w}_{i-1}) + \frac{1}{2\mu}\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad + \mu\sigma^2 \\ &= J(\boldsymbol{w}_{i-1}) + \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) \\ &\quad + \left(\frac{1}{\mu} + 2\mu\beta^2\right)\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \left(\frac{1}{2\mu} - \delta\right)\mathbb{E}\left\{D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} + \mu\sigma^2 \end{aligned} \tag{49}$$

Since $D_h(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}) \geq 0$, whenever $\mu \leq \frac{1}{2\delta}$, we further have:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} \\ &\leq J(\boldsymbol{w}_{i-1}) + \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) \\ &\quad + \left(\frac{1}{\mu} + 2\mu\beta^2\right)\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} + \mu\sigma^2 \end{aligned} \tag{50}$$

Finally, relative strong convexity (24) guarantees:

$$\begin{aligned} &J(\boldsymbol{w}_{i-1}) + \nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}(w^o - \boldsymbol{w}_{i-1}) \\ &\leq J(w^o) - \nu D_h(w^o, \boldsymbol{w}_{i-1}) \end{aligned} \tag{51}$$

ensuring:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} \\ &\leq J(w^o) \\ &\quad + \left(\frac{1}{\mu} - \nu + 2\mu\beta^2\right)\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_{i-1}) |\, \boldsymbol{w}_{i-1}\right\} \\ &\quad - \frac{1}{\mu}\mathbb{E}\left\{D_h(w^o, \boldsymbol{w}_i) |\, \boldsymbol{w}_{i-1}\right\} + \mu\sigma^2 \end{aligned} \tag{52}$$

Rearranging and taking expectations to remove conditioning yields the critical inequality:

$$\begin{aligned} \mathbb{E}J(\boldsymbol{w}_i) - J(w^o) \leq &\left(\frac{1}{\mu} - \nu + 2\mu\beta^2\right)\mathbb{E}D_h(w^o, \boldsymbol{w}_{i-1}) \\ &- \frac{1}{\mu}\mathbb{E}D_h(w^o, \boldsymbol{w}_i) + \mu\sigma^2 \end{aligned} \tag{53}$$

We introduce:

$$\gamma = \frac{\frac{1}{\mu} - \nu + 2\mu\beta^2}{\frac{1}{\mu}} = 1 - \mu\nu + 2\mu^2\beta^2 \tag{54}$$

Then, with proper scaling, relation (53) can be used in a telescoping sum to obtain:

$$\begin{aligned} &\sum_{n=1}^{i} \gamma^{-n}\left(\mathbb{E}J(\boldsymbol{w}_n) - J(w^o)\right) \\ &\leq \frac{1}{\mu} D_h(w^o, w_0) + \mu\sigma^2 \cdot \left(\sum_{n=1}^{i} \gamma^{-n}\right) \end{aligned} \tag{55}$$

From $\min_{n=0,\ldots,i} \mathbb{E}J(\boldsymbol{w}_n) \leq \mathbb{E}J(\boldsymbol{w}_n)$ for all $n = 0, \ldots, i$, we conclude:

$$\begin{aligned} &\left(\sum_{n=1}^{i} \gamma^{-n}\right)\left(\min_{n=0,\ldots,i} \mathbb{E}J(\boldsymbol{w}_n) - J(w^o)\right) \\ &\leq \frac{1}{\mu} D_h(w^o, w_0) + \mu\sigma^2 \cdot \left(\sum_{n=1}^{i} \gamma^{-n}\right) \end{aligned} \tag{56}$$

Theorem 1 follows after rearranging.

B. REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1997.

[3] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[4] H. Lu, R. M. Freund, and Y. Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 333–354, 2018.

[5] A. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, 1983.

[6] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization. technical report," *technical note*, 2008.

[7] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[8] H. Lu, "Relative-continuity for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent," *available as arXiv:1710.04718*, 2017.

[9] S. Zhang and N. He, "On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization," *available as arXiv:1806.04781*, 2018.

[10] F. Hanzely and P. Richtárik, "Fastest rates for stochastic mirror descent methods," *Computational Optimization and Applications*, vol. 79, no. 3, pp. 717–766, July 2021.

[11] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[12] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[13] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[14] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks – Part II: Performance analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3518–3548, June 2015.

[15] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, July 2014.

[16] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[17] P. Di Lorenzo and G. Scutari, "NEXT: in-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, June 2016.

[18] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *Proc. IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 8353–8358.

[19] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part II: Polynomial escape from saddle-points," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1257–1270, 2021.

[20] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," in *2020 Information Theory and Applications Workshop (ITA)*, 2020, pp. 1–37.

[21] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 714–725, 2018.

[22] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2015, pp. 517–520.

[23] A. Agarwal, M. J. Wainwright, and J. C. Duchi, "Distributed dual averaging in networks," in *Advances in Neural Information Processing Systems*, 2010, vol. 23.

[24] H. Yuan, M. Zaheer, and S. Reddi, "Federated composite optimization," in *Proceedings of the 38th International Conference on Machine Learning*, Jul 2021, vol. 139, pp. 12253–12266.

[25] Y. Sun, M. Fazlyab, and S. Shahrampour, "On centralized and distributed mirror descent: Convergence analysis using quadratic constraints," *available as arXiv:2105.14385*, 2021.

[26] S. Vlaski, E. Rizk, and A. H. Sayed, "Second-order guarantees in federated learning," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 915–922.