OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Distributed Online and Bandit Convex Optimization

Kumar Kshitij Patel KKPATEL@TTIC.EDU
Aadirupa Saha AADIRUPA@TTIC.EDU
Nati Srebro NATI@TTIC.EDU
Lingxiao Wang LINGXW@TTIC.EDU
Toyota Technology Institute at Chicago

Abstract

We study the problems of distributed online and bandit convex optimization against an adaptive adversary. Our goal is to minimize the average regret on M machines working in parallel over T rounds that can communicate R times intermittently. Assuming the underlying cost functions are convex, our results show collaboration is not beneficial if the machines have access to the first-order gradient information at the queried points. We show that in this setting, simple non-collaborative algorithms are min-max optimal, as opposed to the case for stochastic functions, where each machine samples the cost functions from a fixed distribution. Next, we consider the more challenging setting of federated optimization with bandit (zeroth-order) feedback, where the machines can only access values of the cost functions at the queried points. The key finding here is to identify the high-dimensional regime where collaboration is beneficial and may even lead to a linear speedup in the number of machines. Our results are the first attempts towards bridging the gap between distributed online optimization against stochastic and adaptive adversaries.

1. Introduction

We consider the following distributed regret minimization problem on M machines with horizon T:

$$\min_{\{x_t^m \in \mathcal{X}\}_{t\in[T]}^{m\in[M]}} \frac{1}{MT} \sum_{m\in[M], t\in[T]} f_t^m(x_t^m) - \min_{x^\star \in \mathcal{X}} \frac{1}{MT} \sum_{m\in[M], t\in[T]} f_t^m(x^\star), \tag{1}$$

where f_t^m is a non-negative, convex cost function observed by machine m at time t, and x_t^m is the model it plays. This formulation captures distributed learning problems where the data is generated in real-time but isn't stored, e.g., mobile keyboard prediction [11, 12] and self-driving vehicles [7, 20]. We want to solve this problem in the *intermittent communication* (IC) setting [29, 31] where the machines work in parallel and are allowed to communicate R times with K time steps in between communication rounds. The IC setting captures the expensive nature of communication in collaborative learning, such as in cross-device federated learning [15, 17].

The IC setting has been widely studied over the past decade [1, 2, 4, 5, 23, 25, 27, 32–34]. Most existing works consider the ***"stochastic"*** setting where $\{f_t^m\}$'s are sampled from a distribution specified in advance. However, real-world applications may have distribution shifts, unmodeled perturbations, or even an adversarial sequence of cost functions, all of which violate the fixed distribution assumption. To alleviate this issue, in this paper, we extend our understanding of distributed online optimization to ***"adaptive"*** adversaries that could potentially generate a worst-case sequence of cost functions. Although some recent works have underlined the importance of the adaptive setting [3, 9, 10, 16, 18], our understanding of the optimal regret guarantees for problem (1) is still lacking.

We first show that, under usual assumptions, there is no benefit of collaboration if all the machines have access to the gradients, a.k.a. first-order feedback for their cost functions. Specifically, in this setting, running online gradient descent on each device without any communication is min-max optimal for problem (1). Thus, we move to the harder setting of bandit convex optimization with two-point feedback. We study a natural variant of FEDAVG equipped with a stochastic gradient estimator due to Shamir [22]. We show that collaboration reduces the variance of the stochastic gradient estimator and is thus beneficial for problems of high enough dimension. We prove a linear speedup in the number of machines for high-dimensional problems, which mimics the stochastic setting [28, 31].

2. Setting

This section introduces notations, definitions, and assumptions used in our analysis.

Notation. We denote the horizon by $T = KR$. $\succeq, \preceq, \cong$ denote inequalities up to numerical constants. We denote the average function by $f_t(\cdot) := \frac{1}{M}\sum_{m\in[M]} f_t^m(\cdot)$ for all $t \in [T]$. We use $\mathbb{1}_A$ to denote the indicator function for the event A. Our model space is denoted by $\mathcal{X} \subseteq \mathbb{R}^d$. We denote the expected averaged regret by $Reg(M, K, R)$ in all the settings.

Function classes. We consider two common [13, 21] function classes in this paper: (i) $\boldsymbol{\mathcal{F}^{G,B}}$, the class of convex, differentiable, non-negative and G-Lipschitz functions, i.e., $\forall x, y \in \mathcal{X}$, $|f(x) - f(y)| \leq G\|x-y\|_2$, with bounded optima, i.e., $\|x^\star\|_2 \leq B$, $\forall\ x^\star \in \arg\min_{x\in\mathcal{X}} f(x)$; (ii) $\boldsymbol{\mathcal{F}^{H,B}}$, the class of convex, differentiable, non-negative and H-smooth functions, i.e., $\forall x, y \in \mathcal{X}$, $\|\nabla f(x) - \nabla f(y)\|_2 \leq H\|x-y\|_2$, with bounded optima. $\boldsymbol{\mathcal{F}^{G,B}}$ includes linear cost functions, while $\boldsymbol{\mathcal{F}^{H,B}}$ consists of quadratic functions denoted by $\boldsymbol{\mathcal{F}^{G,B}_{lin}}$. We also define $\boldsymbol{\mathcal{F}^{G,H,B}} := \boldsymbol{\mathcal{F}^{G,B}} \cap \boldsymbol{\mathcal{F}^{H,B}}$.

Adversary model. Note that in the most general setting, each machine will face arbitrary functions from a class $\mathcal{F}$ at each time step. Our algorithmic results are for this general model, which is usually referred to as an **"adaptive"** adversary. We also consider a weaker **"stochastic"** adversary model to aid comparison. More specifically, the adversary cannot adapt to the sequence of the models used by each machine but must fix a distribution in advance for each machine, i.e., $\forall m \in [M]$, $\mathcal{D}_m \in \Delta(\mathcal{F})$ such that at each time $t \in [T]$, $f_t^m \sim \mathcal{D}_m$. An example of this easier model is distributed stochastic optimization where $f_t^m(\cdot) := f(\cdot; z_t^m \sim \mathcal{D}_m) \in \mathcal{F}$ for $f(\cdot;\cdot) \in \mathcal{F}$.

Oracle model. We consider two kinds of access to the cost functions in this paper. Each machine $m \in [M]$ for all time steps $t \in [T]$ has access to one of the following: (i) **gradient** of f_t^m at a single point, a.k.a., first-order feedback; or (ii) **function values** of f_t^m at two different points, a.k.a., two-point bandit feedback.

We consider two more assumptions controlling how similar the cost functions look across machines and the average regret at the comparator [24]:

Assumption 1 [1] $\forall\, t \in [T], x \in \mathcal{X}$, $\frac{1}{M}\sum_{m\in[M]} \|\nabla f_t^m(x) - \nabla f_t(x)\|_2^2 \leq \zeta^2 \leq 4G^2$.

1. Woodworth et al. [30] consider a more relaxed assumption in the stochastic setting: $\forall\ x \in \mathcal{X}$, $\frac{1}{M}\sum_{m\in[M]} \|\mathbb{E}_{z\sim\mathcal{D}_m}[\nabla f(x;z)] - \nabla f(x)\|_2^2 \leq \zeta^2 \leq 4G^2$ for $f(\cdot) := \frac{1}{M}\sum_{m\in[M]} \mathbb{E}_{z\sim\mathcal{D}_m}[\nabla f(x;z)]$.

Assumption 2 $\forall\, x^\star \in \arg\min_{x\in\mathcal{X}} \sum_{t\in[T]} f_t(x),\ \frac{1}{T}\sum_{t\in[T]} f_t(x^\star) \le F_\star$. *For non-negative functions in $\mathcal{F}^{G,H,B}$, this implies $\frac{1}{T}\sum_{t\in[T]} \|\nabla f_t(x^\star)\|_2^2 \ \le\ HF_\star$ (c.f., Lemma 4.1 [24]).*

Min-max regret. We can finally define our problem class. We use $\mathcal{P}_{M,K,R}(\mathcal{F}) := \mathcal{F}^{\otimes MKR}$ to denote all selections of MKR functions from a class $\mathcal{F}$. We use the argument ζ, $F_\star$ to further restrict this to selections that satisfy Assumptions 1 and 2 respectively. Furthermore, with a slight abuse of notation, we use the superscript 1 to denote first-order feedback and $(0,2)$ to denote two-point zeroth-order feedback to the cost functions. In this paper, we consider four problem classes: $\mathcal{P}^1_{M,K,R}(\mathcal{F}^{G,D},\zeta)$, $\mathcal{P}^{0,2}_{M,K,R}(\mathcal{F}^{G,B},\zeta)$, $\mathcal{P}^1_{M,K,R}(\mathcal{F}^{H,B},\zeta,F_\star)$, $\mathcal{P}^{0,2}_{M,K,R}(\mathcal{F}^{G,H,B},\zeta,F_\star)$. And we are interested in characterizing the min-max regret for these classes. In particular, for a problem class $\mathcal{P}$, we want to characterize up to numerical constants the following quantity:

$$\mathcal{R}(\mathcal{P}) := \min_{\mathcal{A}} \max_{P\in\mathcal{P}} \mathbb{E}_{\mathcal{A}}\left(\frac{1}{MT}\sum_{t\in[T],m\in[M]} f_t^m(x_t^m) - \min_{x\in\mathcal{X}} \frac{1}{MT}\sum_{t\in[T],m\in[M]} f_t^m(x)\right), \tag{2}$$

where $\mathcal{A}$ is a randomized algorithm producing models x_t^m's. For stochastic adversaries, the expectation is also taken over the randomness of sampling from the distributions $\mathcal{D}_m \in \Delta(\mathcal{F})$.

3. Collaboration doesn't help with First-order Feedback

We first consider the class $\mathcal{P}^1_{M,K,R}(\mathcal{F}^{G,B},\zeta)$. Note the following bound is always true for any stream of functions and sequence of models:

$$\frac{1}{M}\sum_{m\in[M]}\left(\sum_{t\in[KR]} f_t^m(x_t^m) - \min_{x^m\in\mathcal{X}}\sum_{t\in[KR]} f_t^m(x^m)\right) \ge \frac{1}{M}\sum_{t\in[KR],m\in[M]} f_t^m(x_t^m) - \min_{x\in\mathcal{X}}\sum_{t\in[KR]} f_t(x).$$

This means we can upper bound regret in equation 1 by running online gradient descent (OGD) independently on each machine and not collaborating at all. In other words:

$$\mathcal{R}\left(\mathcal{P}^1_{M,K,R}(\mathcal{F}^{G,D},\zeta)\right) \preceq \mathcal{R}\left(\mathcal{P}^1_{1,K,R}(\mathcal{F}^{G,D})\right) \cong \frac{GB}{\sqrt{T}}. \tag{3}$$

The min-max rate for a single machine follows classical results using vanilla OGD (c.f., Theorem 3.1 in [13]). But **can collaborative algorithms beat this natural baseline? No!**

Consider the problem where the functions don't vary across the machines but may change with time. This problem satisfies Assumption 1 with $\zeta = 0$. In this problem, the machines jointly see only T different functions but can make M first-order queries to the functions at each time step. However, these additional queries are not useful as there is a known sample-complexity lower bound of $GB/\sqrt{T}$ for $\mathcal{P}^1_{1,K,R}(\mathcal{F}^{G,B})$ (c.f., Theorem 3.2 [13]) which holds for any number of first-order queries at each time step. This implies that,

$$\frac{GB}{\sqrt{T}} \cong \mathcal{R}\left(\mathcal{P}^1_{1,K,R}(\mathcal{F}^{G,D})\right) \preceq \mathcal{R}\left(\mathcal{P}^1_{M,K,R}(\mathcal{F}^{G,B},\zeta)\right). \tag{4}$$

Combining equations (3) and (4), we conclude that $\mathcal{R}\left(\mathcal{P}^1_{M,K,R}(\mathcal{F}^{G,B},\zeta)\right) \cong GB/\sqrt{T}$. Or in other words, there is no benefit of collaboration when the machines have first-order feedback.

We recall several interesting functions, such as quadratics, that don't lie in $\mathcal{F}^{G,B}$ but lie in $\mathcal{F}^{H,B}$. To understand the latter class we look at problems in $\mathcal{P}^1_{M,K,R}(\mathcal{F}^{H,B}, \zeta, F_\star)$. In the single machine setting, we know that OGD incurs a regret of $HB^2/T + \sqrt{HF_\star}B/\sqrt{T}$ (c.f., Theorem 3 [24]). This serves as the non-collaborative baseline. Unfortunately, there is again a matching sample complexity lower bound for $\mathcal{P}^1_{1,K,R}(\mathcal{F}^{H,B}, F_\star)$ (c.f., Theorem 4 [28]). Using a similar argument as before, we can obtain that,

$$\mathcal{R}(\mathcal{P}^1_{M,K,R}(\mathcal{F}^{H,B}, \zeta, F_\star)) \cong \frac{HB^2}{T} + \frac{\sqrt{HF_\star}B}{\sqrt{T}}, \tag{5}$$

which suggests that regret doesn't improve with collaboration, either.

Thus, when the machines have first-order feedback for their own objectives, they do not benefit from collaboration. The commonality between these problems is that even when the functions are the same across the machines, the hardest instances within the problem class do not benefit from the additional gradient accesses. This is not surprising because linear functions are the hardest lipschitz and smooth functions in the adversarial online setting, and they are fully specified by their gradient. This suggests that we should consider settings where machines have weaker oracles than first-order and may benefit through collaboration. One such setting is with stochastic first-order oracles because, with additional stochastic gradients, the machines can reduce the variance of their gradient estimator. This is one mechanism through which collaboration helps in the stochastic setting [28, 31], and we see next that it naturally arises in bandit convex optimization.

4. Online Local SGD Algorithm with Two-point Bandit Feedback

Algorithm 1: FEDOSGD (η, δ) with two-point bandit feedback

```
Initialize x_0^m = 0 on all machines m ∈ [M]
for t ∈ {0, ..., KR − 1} do
    for m ∈ [M] in parallel do
        Sample u_t^m ~ Unif(S_{d−1}), i.e., a random unit vector
        Query function f_t^m at points (x_t^{m,1}, x_t^{m,2}) := (x_t^m + δu_t^m, x_t^m − δu_t^m)
        Incur loss (f_t^m(x_t^m + δu_t^m) + f_t^m(x_t^m − δu_t^m))
        Compute stochastic gradient at point x_t^m as g_t^m = d(f(x_t^m + δu_t^m) − f(x_t^m − δu_t^m)) u_t^m / (2δ)
        if (t + 1) mod K = 0 then
            Communicate to server: (x_t^m − η · g_t^m)
            On server x_{t+1} ← (1/M) Σ_{m∈[M]} (x_t^m − η · g_t^m)
            Communicate to machine: x_{t+1}^m ← x_{t+1}
        else
            x_{t+1}^m ← x_t^m − η · g_t^m
```

In this section, we study distributed bandit convex optimization with two-point feedback [6, 22], i.e., at each time step, the machines can query the value (and not the full gradient) of their cost functions at two different points. We analyze the online variant of the FEDAVG or LOCAL-SGD algorithm, which is common in the stochastic setting. We call the algorithm FEDOSGD and describe it in Algorithm 1. In line 7, we use the stochastic gradient estimator, originally proposed by Shamir [22] and based on a similar estimator by Duchi et al. [6]. For a smoothed version of the function $\hat{f}_t^m(x) := \mathbb{E}_u[f_t^m(x+\delta u)]$, this estimator satisfies (c.f., Lemmas 3 and 5 [22]) for all $t \in [T]$, $m \in [M]$ and $x \in \mathcal{X}$,

$$\mathbb{E}_u[g_t^m(x)] = \nabla \hat{f}_t^m(x) \quad \text{and} \quad \mathbb{E}_u\left[\left\|g_t^m(x) - \nabla \hat{f}(x)\right\|_2^2\right] \preceq dG^2.$$

Equipped with this gradient estimator, we can prove the following guarantee for $\mathcal{P}_{M,K,R}^{1,\sigma}(\mathcal{F}^{G,B}, \zeta)$.

Theorem 1 *Consider the problem class $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{G,B}, \zeta)$. With $\eta = \frac{B}{G\sqrt{T}} \cdot \min\left\{1, \frac{\sqrt{M}}{\sqrt{d}}, \frac{1}{\mathbb{1}_{K>1}\sqrt{K}d^{1/4}}\right\}$, the queried points $\{x_t^{m,j}\}_{t,m,j=1}^{T,M,2}$ of Algorithm 1 satisfy:*

$$\frac{1}{2MKR} \sum_{t\in[KR], m\in[M], j\in[2]} \mathbb{E}\left[f_t^m(x_t^{m,j}) - f_t^m(x^\star)\right] \preceq \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1} \cdot \frac{GBd^{1/4}}{\sqrt{R}},$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d} \sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the choice of function queries.

When $K = 1$, the above bound reduces to the first two terms, which are known to be tight for two-point bandit feedback [6, 13] (see Appendix A), making FEDOSGD optimal. When $K > 1$, we would like to compare our results to the non-collaborative baseline as we did in section 3. Using the gradient estimator proposed by Shamir [22], the non-collaborative baseline gets regret $\mathcal{O}\left(GB\sqrt{d}/\sqrt{KR}\right)$. Thus, as long as $d \succeq K^2$, FEDOSGD is better than the non-collaborative baseline. Furthermore, if $d \succeq K^2M^2$, then the second term in the upper bound dominates, and FEDOSGD gets a *"linear speed-up"* in the number of machines. Unfortunately, the bound doesn't improve with smaller ζ.

Note that the lipschitzness assumption is crucial to the two-point gradient estimator in algorithm 1. While there are gradient estimators that don't require lipschitzness or bounded gradients [8], they do require stronger assumptions such as bounded function values. To avoid making these assumptions, we skip looking at the problems in $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{H,B}, \zeta, F_\star)$ and look at the problems in $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$.

Theorem 2 *Consider the problem class $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$. With appropriate η (c.f., lemma 6), the queried points $\{x_t^{m,j}\}_{t,m,j=1}^{T,M,2}$ of Algorithm 1 satisfy:*

$$\begin{aligned}\frac{1}{2MKR} \sum_{t\in[KR], m\in[M], j\in[2]} \mathbb{E}\left[f_t^m(x_t^{m,j}) - f_t^m(x^\star)\right] &\preceq \frac{HB^2}{KR} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \frac{GB}{\sqrt{KR}} + \frac{\sqrt{HF_\star}B}{\sqrt{KR}} \\ &+ \mathbb{1}_{K>1} \cdot \min\left\{\frac{H^1/3B^{4/3}G^{2/3}d^{1/3}}{K^{1/3}R^{2/3}} + \frac{H^{1/3}B^{4/3}\zeta^{2/3}}{R^{2/3}}\right. \\ &\left. + \frac{\sqrt{\zeta G}Bd^{1/4}}{K^{1/4}\sqrt{R}} + \frac{\zeta B}{\sqrt{R}}, \frac{GBd^{1/4}}{K^{1/4}\sqrt{R}} + \frac{\sqrt{G\zeta}B}{\sqrt{R}}\right\},\end{aligned}$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d} \sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the choice of function queries. The regret is also upper bounded as in theorem 1 for the corresponding step size.

The above result is a bit technical, so to interpret it, we consider the simpler class $\mathcal{F}_{lin}^{G,B}$ of linear functions with bounded gradients. Linear functions are the "smoothest" Lipschitz functions as their smoothness constant $H = 0$. We can get the following guarantee for this class:

Corollary 3 *Consider the problem class $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}_{lin}^{G,0,B}, \zeta, F_\star)$. With appropriate η (c.f., lemma 6), the queried points $\{x_t^{m,j}\}_{t,m,j=1}^{T,M,2}$ of Algorithm 1 satisfy:*

$$\frac{1}{2MKR}\sum_{t\in[KR],m\in[M],j\in[2]}\mathbb{E}\left[f_t^m(x_t^{m,j}) - f_t^m(x^\star)\right] \preceq \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1}\cdot\left(\frac{\sqrt{\zeta G}Bd^{1/4}}{K^{1/4}\sqrt{R}} + \frac{\zeta B}{\sqrt{R}}\right),$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d}\sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the choice of function queries.

Unlike general Lipschitz functions, the last two terms are zero for linear functions when $\zeta = 0$ and the upper bound is smaller for smaller ζ. In fact, when $K = 1$ or $\zeta = 0$, the upper bound is tight [6]. More generally, when $K \leq \max(1, G^2\zeta^2 d, G^2 d/\zeta^2 M^2)$ then FEDOSGD is optimal. Recall that in this setting, the non-collaborative baseline obtains a regret [24] of $\mathcal{O}(GB\sqrt{d}/\sqrt{KR})$. Thus, the benefit of collaboration through FEDOSGD again appears in high-dimensional problems in $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$ similar to what we discussed for $\mathcal{P}_{M,K,R}^{0,2}(\mathcal{F}^{G,B}, \zeta, F_\star)$.

5. Conclusion

In this paper, we show that, in the adaptive bandit setting, the benefit of collaboration is very similar to the stochastic setting, where the collaboration is useful when: (i) There is stochasticity in the problem and (ii) The variance of the gradient estimators is *"high"* [31] and reduces with collaboration. There are several open questions and directions:

1. Does collaboration provably not help for the smaller class $\mathcal{P}_{M,K,R}^{1}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$? This might require new proof techniques.
2. Is the final term tight in Theorems 1 and 2? We don't know any lower bounds in the intermittent communication setting against an adaptive adversary. Perhaps there is no gap between the stochastic and adaptive adversaries, and we can use existing techniques and online-to-batch conversion to provide a tight lower bound.
3. When K is large, but R is a fixed constant, the average regret of the non-collaborative baseline goes to zero, but our upper bounds for FEDOSGD don't. It is unclear if our analysis is loose or if we need to modify the algorithm, for instance, add projection steps.
4. How to obtain second-order methods in the distributed online setting, especially in the intermittent communication setting? This only makes sense when the worst-case functions are not linear, which we might expect in the distributed setting [26].
5. For stochastic linear bandits, collaborative methods have been shown to attain optimal regret with very few rounds of communication [14]. What structures in the problem can we further exploit to reduce communication?

Acknowledgements

We thank the anonymous reviewers who helped us improve the writing of the paper. This research was partly supported by NSF-BSF award 1718970, NSF TRIPOD IDEAL award, and the NSF-Simons funded Collaboration on the Theoretical Foundations of Deep Learning.

References

[1] Brian Bullins, Kshitij Patel, Ohad Shamir, Nathan Srebro, and Blake E Woodworth. A stochastic newton algorithm for distributed convex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24, 2011.

[3] Shuang Dai and Fanlin Meng. Addressing modern and practical challenges in machine learning: A survey of online federated and transfer learning. *arXiv preprint arXiv:2202.03070*, 2022.

[4] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

[5] Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for local-sgd with large step size. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[7] Ahmet M Elbir, Burak Soner, and Sinem Coleri. Federated learning in vehicular networks. *arXiv preprint arXiv:2006.01412*, 2020.

[8] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

[9] Francois Gauthier, Vinay Chakravarthi Gogineni, Stefan Werner, Yih-Fang Huang, and Anthony Kuh. Resource-aware asynchronous online federated learning for nonlinear regression. In *ICC 2022-IEEE International Conference on Communications*, pages 2828–2833. IEEE, 2022.

[10] Vinay Chakravarthi Gogineni, Stefan Werner, Yih-Fang Huang, and Anthony Kuh. Communication-efficient online federated learning framework for nonlinear regression. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5228–5232. IEEE, 2022.

[11] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Fran*c*coise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[12] Florian Hartmann. Predicting text selections with federated learning, Nov 2021. URL `https://ai.googleblog.com/2021/11/predicting-[]text-[]selections-[]with.html`.

[13] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[14] Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34:27057–27068, 2021.

[15] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. corr. *arXiv preprint arXiv:1912.04977*, 2019.

[16] Anthony Kuh. Real time kernel learning for sensor networks using principles of federated learning. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 2089–2093. IEEE, 2021.

[17] H Brendan McMahan, Eider Moore, Daniel Ramage, S Hampson, and B Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016.

[18] Aritra Mitra, Hamed Hassani, and George J Pappas. Online federated learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4083–4090. IEEE, 2021.

[19] Arkadi Nemirovski. Efficient methods in convex programming. *Lecture notes*, 1994.

[20] Anh Nguyen, Tuong Do, Minh Tran, Binh X Nguyen, Chien Duong, Tu Phan, Erman Tjiputra, and Quang D Tran. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1824–1830. IEEE, 2022.

[21] Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 636–642. JMLR Workshop and Conference Proceedings, 2011.

[22] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

[23] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[24] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

[25] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[26] Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

[27] Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[28] Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in Neural Information Processing Systems*, 34:7333–7345, 2021.

[29] Blake E Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in neural information processing systems*, 31, 2018.

[30] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020.

[31] Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.

[32] Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

[33] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

[34] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Appendix A. Proof of Theorem 1

In this section and the next one, we consider access to a first-order stochastic oracle as an intermediate step before considering the zeroth-order oracle. Specifically, each machine has access to a **stochastic gradient** g_t^m of f_t^m at point x_t^m, such that it is unbiased and has bounded variance, i.e., for all $x \in \mathcal{X}$,

$$\mathbb{E}[g_t^m(x_t^m)|x_t^m] = \nabla f_t^m(x_t^m) \text{ and } \mathbb{E}\left[\|g_t^m(x_t^m) - \nabla f_t^m(x_t^m)\|_2^2 \,|x_t^m\right] \leq \sigma^2.$$

In algorithm 1, we constructed a particular stochastic gradient estimator at x_t^m with $\sigma^2 = G^2 d$. We can define the corresponding problem class $\mathcal{P}_{M,K,R}^{1,\sigma}(\mathcal{F}^{G,B}, \zeta)$ when the agents can access a stochastic first-order oracle. We prove the following lemma about this problem class:

Lemma 4 *Consider the problem class $\mathcal{P}_{M,K,R}^{1,\sigma}(\mathcal{F}^{G,B}, \zeta)$. If we choose $\eta = \frac{B}{G\sqrt{T}} \cdot \min\left\{1, \frac{G\sqrt{M}}{\sigma}, \frac{\sqrt{G}}{\mathbb{1}_{K>1}\sqrt{\sigma K}}, \frac{1}{\mathbb{1}_{K>1}\sqrt{K}}\right\}$, then the models $\{x_t^m\}_{t,m=1}^{T,M}$ of Algorithm 1 satisfy the following guarantee:*

$$\frac{1}{MKR} \sum_{t\in[KR],m\in[M]} \mathbb{E}\left[f_t^m(x_t^m) - f_t^m(x^\star)\right] \preceq \frac{GB}{\sqrt{KR}} + \frac{\sigma B}{\sqrt{MKR}} + \mathbb{1}_{K>1} \cdot \left(\frac{\sqrt{\sigma G}B}{\sqrt{R}} + \frac{GB}{\sqrt{R}}\right),$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d} \sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the stochastic gradients.

Proof Consider any time step $t \in [KR]$ and define ghost iterate $\bar{x}_t = \frac{1}{M}\sum_{m\in[M]} x_t^m$ (which not might actually get computed). If $K = 1$, the machines calculate the stochastic gradient at the same point, $\bar{x}_t$. Then using the update rule of Algorithm 1, we can get the following:

$$\begin{aligned}
\mathbb{E}_t\left[\|\bar{x}_{t+1} - x^\star\|_2^2\right] &= \mathbb{E}_t\left[\left\|\bar{x}_t - \frac{\eta_t}{M}\sum_{m\in[M]} \nabla f_t^m(x_t^m) - x^\star + \frac{\eta_t}{M}\sum_{m=1}^{M}\left(\nabla f_t^m(x_t^m) - g_t^m(x_t^m)\right)\right\|_2^2\right] \\
&= \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\langle \bar{x}_t - x^\star, \nabla f_t^m(x_t^m)\rangle + \frac{\eta_t^2\sigma^2}{M} \\
&= \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\langle x_t^m - x^\star, \nabla f_t^m(x_t^m)\rangle \\
&\quad + \mathbb{1}_{K>1} \cdot \frac{2\eta_t}{M}\sum_{m\in[M]}\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m)\rangle + \frac{\eta_t^2\sigma^2}{M} \\
&\leq \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\left(f_t^m(x_t^m) - f_t^m(x^\star)\right) \\
&\quad + \mathbb{1}_{K>1} \cdot \frac{2\eta_t}{M}\sum_{m\in[M]}\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m)\rangle + \frac{\eta_t^2\sigma^2}{M},
\end{aligned}$$

where $\mathbb{E}_t$ is the expectation conditioned on the filtration at time t under which x_t^m's are measurable, and the last inequality is due to the convexity of each function. Re-arranging this leads to

$$\begin{aligned}
\frac{1}{M}\sum_{m\in[M]}\left(f_t^m(x_t^m)-f_t^m(x^\star)\right) &\leq \frac{1}{2\eta_t}\left(\|\bar{x}_t-x^\star\|_2^2-\mathbb{E}_t\left[\|\bar{x}_{t+1}-x^\star\|_2^2\right]\right)+\frac{\eta_t}{2M^2}\left\|\sum_{m\in[M]}\nabla f_t^m(x_t^m)\right\|_2^2 \\
&\quad + \mathbb{1}_{K>1}\cdot\frac{1}{M}\sum_{m\in[M]}\mathbb{E}_t\left\langle x_t^m-\bar{x}_t,\nabla f_t^m(x_t^m)\right\rangle+\frac{\eta_t\sigma^2}{2M} \\
&\leq \frac{1}{2\eta_t}\left(\|\bar{x}_t-x^\star\|_2^2-\mathbb{E}_t\left[\|\bar{x}_{t+1}-x^\star\|_2^2\right]\right)+\frac{\eta_t}{2}\left(G^2+\frac{\sigma^2}{M}\right) \\
&\quad + \mathbb{1}_{K>1}\cdot\frac{G}{M}\sum_{m\in[M]}\mathbb{E}\left[\|x_t^m-\bar{x}_t\|_2\right]. \qquad (6)
\end{aligned}$$

The last inequality comes from each function's G-Lipschitzness. For the last term in (6), we can upper bound it similar to lemma 8 in Woodworth et al. [30] to get that

$$\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\|x_t^m-\bar{x}_t\|_2\right]\leq 2(\sigma+G)K\eta. \qquad (7)$$

Plugging (7) into (6) and choosing a constant step-size η, and taking full expectation we get

$$\begin{aligned}
\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[f_t^m(x_t^m)-f_t^m(x^\star)\right] &\leq \frac{1}{2\eta}\left(\left\|\mathbb{E}\left[\bar{x}_t-x^\star\right]^2\right\|_2-\mathbb{E}\left[\|\bar{x}_{t+1}-x^\star\|_2^2\right]\right) \\
&\quad +\frac{\eta}{2}\left(G^2+\frac{\sigma^2}{M}\right)+\mathbb{1}_{K>1}\cdot 2G(\sigma+G)K\eta.
\end{aligned}$$

Summing this over time $t\in[KR]$ we get,

$$\begin{aligned}
\frac{1}{M}\sum_{m\in[M],t\in[T]}\mathbb{E}\left[f_t^m(x_t^m)-f_t^m(x^\star)\right] &\preceq \frac{\|\bar{x}_0-x^\star\|_2^2}{\eta}+\eta\left(G^2+\frac{\sigma^2}{M}+\mathbb{1}_{K>1}\cdot\sigma GK+\mathbb{1}_{K>1}\cdot\zeta GK\right)T \\
&\preceq \frac{B^2}{\eta}+\eta\left(G^2+\frac{\sigma^2}{M}+\mathbb{1}_{K>1}\cdot\sigma GK+\mathbb{1}_{K>1}\cdot G^2K\right)T.
\end{aligned}$$

Finally choosing,

$$\eta=\frac{B}{G\sqrt{T}}\cdot\min\left\{1,\frac{G\sqrt{M}}{\sigma},\frac{\sqrt{G}}{\mathbb{1}_{K>1}\sqrt{\sigma K}},\frac{1}{\mathbb{1}_{K>1}\sqrt{K}}\right\},$$

we can obtain,

$$\frac{1}{M}\sum_{m\in[M],t\in[T]}\mathbb{E}\left[f_t^m(x_t^m)-f_t^m(x^\star)\right]\preceq GB\sqrt{T}+\mathbb{1}_{K>1}\cdot\sqrt{\sigma G}B\sqrt{KT}+\mathbb{1}_{K>1}\cdot GB\sqrt{KT}+\frac{\sigma B\sqrt{T}}{\sqrt{M}}. \qquad (8)$$

Dividing by KR finishes the proof. ■

Remark 5 *Note that when $K = 1$, the upper bound in Lemma 4 reduces to the first two terms, both of which are known to be optimal due to lower bounds in the stochastic setting, i.e., against a stochastic online adversary [13, 19]. We now use this lemma to guarantee bandit two-point feedback oracles for the same function class. We recall that one can obtain a stochastic gradient for a "smoothed-version" $\hat{f}$ of a Lipschitz function f at any point $x \in \mathcal{X}$, using two function value calls to f around the point x [6, 22].*

With this lemma, we can prove Theorem 1.

Theorem 1 *Consider the problem class $\mathcal{P}^{0,2}_{M,K,R}(\mathcal{F}^{G,B}, \zeta)$. With $\eta = \frac{B}{G\sqrt{T}} \cdot \min\left\{1, \frac{\sqrt{M}}{\sqrt{d}}, \frac{1}{\mathbb{1}_{K>1}\sqrt{K}d^{1/4}}\right\}$, the queried points $\{x_t^{m,j}\}_{t,m,j=1}^{T,M,2}$ of Algorithm 1 satisfy:*

$$\frac{1}{2MKR}\sum_{t\in[KR],m\in[M],j\in[2]} \mathbb{E}\left[f_t^m(x_t^{m,j}) - f_t^m(x^\star)\right] \preceq \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1}\cdot\frac{GBd^{1/4}}{\sqrt{R}},$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d}\sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the choice of function queries.

Proof First, we consider smoothed functions

$$\hat{f}_t^m(x) := \mathbb{E}_{u\sim Unif(S_{d-1})}[f_t^m(x+\delta u)],$$

for some $\delta > 0$ and S_{d-1} denoting the euclidean unit sphere. Based on the gradient estimator in (**??**) proposed by Shamir [22] (which can be implemented with two-point bandit feedback) and Lemma 4, we can get the following regret guarantee (noting that $\sigma \le c_1\sqrt{d}G$ for a numerical constant c_1, c.f., [22]):

$$\mathbb{E}\left[\frac{1}{MKR}\sum_{t\in[KR],m\in[M]} \hat{f}_t^m(\hat{x}_t^m)\right] - \frac{1}{MKR}\sum_{t\in[KR],m\in[M]} \hat{f}_t^m(x^\star) \preceq \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1}\cdot\frac{GBd^{1/4}}{\sqrt{R}},$$

where the expectation is w.r.t. the stochasticity in the stochastic gradient estimator. To transform this into a regret guarantee for f we need to account for two things:

1. The difference between the smoothed function $\hat{f}$ and the original function f. This is easy to handle because both these functions are pointwise close, i.e., $\sup_{x\in\mathcal{X}}|f(x) - \hat{f}(x)| \le G\delta$.

2. The difference between the points $\hat{x}_t^m$ at which the stochastic gradient is computed for $\hat{f}_t^m$ and the actual points $x_t^{m,1}$ and $x_t^{m,2}$ on which we incur regret while making zeroth-order queries to f_t^m. This is also easy to handle because due to the definition of the estimator in **??**, $x_t^{m,1}, x_t^{m,1} \in B_\delta(\hat{x}_t^m)$, where $B_\delta(x)$ is the L_2 ball of radius δ around x.

In light of the last two observations, the average regret between the smoothed and original functions only differs by a factor of $2G\delta$, i.e.,

$$\mathbb{E}\left[\frac{1}{2MKR}\sum_{t\in[KR],m\in[M],j\in[2]} f_t^m(x_t^{m,j})\right] - \frac{1}{MKR}\sum_{t\in[KR],m\in[M]} f_t^m(x^\star)$$

$$\preceq G\delta + \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1} \cdot \frac{GBd^{1/4}}{\sqrt{R}}$$
$$\preceq \frac{GB}{\sqrt{KR}} + \frac{GB\sqrt{d}}{\sqrt{MKR}} + \mathbb{1}_{K>1} \cdot \frac{GBd^{1/4}}{\sqrt{R}},$$

where the last inequality is due to the choice of δ such that $\delta \preceq \frac{Dd^{1/4}}{\sqrt{R}}\left(1 + \frac{d^{1/4}}{\sqrt{MK}}\right)$. ■

Appendix B. Proof of Theorem 2

Similar to before, we start by looking at $\mathcal{P}^{1,\sigma}_{M,K,R}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$. We first prove the following Lemma:

Lemma 6 *Consider the problem class $\mathcal{P}^{1,\sigma}_{M,K,R}(\mathcal{F}^{G,H,B}, \zeta, F_\star)$. The models $\{x_t^m\}_{t,m=1}^{T,M}$ of Algorithm 1 with appropriate η (specified in the proof) satisfy the following regret guarantee:*

$$\frac{1}{MKR}\sum_{t\in[KR],m\in[M]} \mathbb{E}\left[f_t^m(x_t^m) - f_t^m(x^\star)\right] \preceq \frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{GB}{\sqrt{KR}}, \frac{\sqrt{HF_\star}B}{\sqrt{KR}}\right\},$$
$$+ \mathbb{1}_{K>1}\cdot \min\left\{\frac{H^1/3B^{4/3}\sigma^{2/3}}{K^{1/3}R^{2/3}} + \frac{H^{1/3}B^{4/3}\zeta^{2/3}}{R^{2/3}} + \frac{\sqrt{\zeta\sigma}B}{K^{1/4}\sqrt{R}} + \frac{\zeta B}{\sqrt{R}},\right.$$
$$\left.\frac{\sqrt{G\sigma}B}{K^{1/4}\sqrt{R}} + \frac{\sqrt{G\zeta}B}{\sqrt{R}}\right\},$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d}\sum_{t\in[KR]} f_t(x)$, and the expectation is w.r.t. the stochastic gradients. The models also satisfy the guarantee of lemma 4 with the same step-size.

Proof Consider any time step $t \in [KR]$ and define ghost iterate $\bar{x}_t = \frac{1}{M}\sum_{m\in[M]} x_t^m$ (which not might actually get computed). Then using the update rule of Algorithm 1, we can get:

$$\mathbb{E}_t\left[\|\bar{x}_{t+1} - x^\star\|_2^2\right] = \mathbb{E}_t\left[\left\|\bar{x}_t - \frac{\eta_t}{M}\sum_{m\in[M]}\nabla f_t^m(x_t^m) - x^\star + \frac{\eta_t}{M}\sum_{m=1}^{M}\left(\nabla f_t^m(x_t^m) - g_t^m(x_t^m)\right)\right\|_2^2\right],$$
$$= \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]}\nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\langle \bar{x}_t - x^\star, \nabla f_t^m(x_t^m)\rangle + \frac{\eta_t^2\sigma^2}{M}$$
$$= \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]}\nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\langle x_t^m - x^\star, \nabla f_t^m(x_t^m)\rangle$$
$$+ \mathbb{1}_{K>1}\cdot\frac{2\eta_t}{M}\sum_{m\in[M]}\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m)\rangle + \frac{\eta_t^2\sigma^2}{M}$$
$$\leq \|\bar{x}_t - x^\star\|_2^2 + \frac{\eta_t^2}{M^2}\left\|\sum_{m\in[M]}\nabla f_t^m(x_t^m)\right\|_2^2 - \frac{2\eta_t}{M}\sum_{m\in[M]}\left(f_t^m(x_t^m) - f_t^m(x^\star)\right)$$

$$+ \mathbb{1}_{K>1} \cdot \frac{2\eta_t}{M} \sum_{m\in[M]} \langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m) \rangle + \frac{\eta_t^2 \sigma^2}{M},$$

where $\mathbb{E}_t$ is the expectation taken with respect to the filtration at time t, and the last line comes from the convexity of each function. Re-arranging this and taking expectation gives leads to

$$\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left(f_t^m(x_t^m) - f_t^m(x^\star)\right) \le \frac{1}{2\eta_t}\left(\mathbb{E}\left\|\bar{x}_t - x^\star\right\|_2^2 - \mathbb{E}\left[\left\|\bar{x}_{t+1} - x^\star\right\|_2^2\right]\right) + \frac{\eta_t}{2M^2}\mathbb{E}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2$$
$$+ \mathbb{1}_{K>1} \cdot \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m) \right\rangle + \frac{\eta_t \sigma^2}{2M} \tag{9}$$

Bounding the blue term. We consider two different ways to bound the term. First note that similar to lemma 4 we can just use the following bound,

$$\frac{\eta_t}{2M^2}\mathbb{E}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2 \le \frac{\eta_t G^2}{2} \tag{10}$$

However, since we also have smoothness, we can use the self-bounding property (c.f., Lemma 4.1 [24]) to get,

$$\frac{\eta_t}{2M^2}\mathbb{E}\left\|\sum_{m\in[M]} \nabla f_t^m(x_t^m)\right\|_2^2 \le \frac{\eta_t H}{2M} \sum_{m\in[M]} \left(f_t^m(x_t^m) - f_t^m(x^\star)\right) + \frac{\eta_t H}{2M} \sum_{m\in[M]} f_t^m(x^\star) \tag{11}$$

Bounding the red term. We will bound the term in three different ways. Similar to lemma 4, we can bound the term after taking expectation and then bounding the consensus term similar to Lemma 8 in Woodworth et al. [30] as follows,

$$\begin{aligned}\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\left\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m) \right\rangle\right] &\le \frac{G}{M} \sum_{m\in[M]} \mathbb{E}\left[\left\|x_t^m - \bar{x}_t\right\|_2\right] \\ &\le 2G(\sigma + G) \sum_{t'=\tau(t)}^{\tau(t)+K-1} \eta_{t'},\end{aligned} \tag{12}$$

where $\tau(t)$ maps t to the last time step when communication happens. Alternatively, we can use smoothness as follows after assuming $\eta_t \le 1/2H$,

$$\begin{aligned}\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\left\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m) \right\rangle\right] &= \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\left\langle x_t^m - \bar{x}_t, \nabla f_t^m(x_t^m) - \nabla f_t(\bar{x}_t) \right\rangle\right], \\ &\le \sqrt{\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left\|x_t^m - \bar{x}_t\right\|_2^2}\sqrt{\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left\|\nabla f_t^m(x_t^m) - \nabla f_t(\bar{x}_t)\right\|_2^2}, \\ &\le \sqrt{\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left\|x_t^m - \bar{x}_t\right\|_2^2}\sqrt{\frac{2}{M} \sum_{m\in[M]} H^2\mathbb{E}\left\|x_t^m - \bar{x}_t\right\|_2^2 + 2\zeta^2},\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2H}{M}\sum_{m\in[M]}\mathbb{E}\left\|x_t^m-\bar{x}_t\right\|_2^2+2\zeta\sqrt{\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left\|x_t^m-\bar{x}_t\right\|_2^2},\\
&\leq 2\eta_t^2H(\sigma^2K+\zeta^2K^2)+2\eta_t\zeta(\sigma\sqrt{K}+\zeta K), \qquad (13)
\end{aligned}
$$

where we used lemma 8 from Woodworth et al. [30] in the last inequality. We can also use the lipschitzness and smoothness assumption together with a constant step size $\eta < 1/2H$ to obtain,

$$
\begin{aligned}
\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\langle x_t^m-\bar{x}_t,\nabla f_t^m(x_t^m)\rangle\right] &\leq \frac{G}{M}\sum_{m\in[M]}\mathbb{E}\left[\|x_t^m-\bar{x}_t\|_2\right]\\
&\leq \eta G(\sigma\sqrt{K}+\zeta K). \qquad (14)
\end{aligned}
$$

Combining everything. After using a constant step-size η, summing (9) over time, we can use the upper bound of the red and blue terms in different ways. If we plug in (10) and (12) we recover the guarantee of lemma 4. This is not surprising because $\mathcal{F}^{G,H,B}\subseteq\mathcal{F}^{G,B}$. Combining the upper bounds in all other combinations assuming $\eta<\frac{1}{2H}$, we can show the following upper bound

$$
\begin{aligned}
\frac{Reg(M,K,R)}{KR} \preceq\ & \frac{HB^2}{KR}+\frac{\sigma B}{\sqrt{MKR}}+\min\left\{\frac{GB}{\sqrt{KR}},\frac{\sqrt{HF_\star}B}{\sqrt{KR}}\right\},\\
&+\mathbb{1}_{K>1}\min\left\{\frac{H^{1/3}B^{4/3}\sigma^{2/3}}{K^{1/3}R^{2/3}}+\frac{H^{1/3}B^{4/3}\zeta^{2/3}}{R^{2/3}}+\frac{\sqrt{\zeta\sigma}B}{K^{1/4}\sqrt{R}}+\frac{\zeta B}{\sqrt{R}},\frac{\sqrt{G\sigma}B}{K^{1/4}\sqrt{R}}+\frac{\sqrt{G\zeta}B}{\sqrt{R}}\right\},
\end{aligned}
$$

where we used step size,

$$
\begin{aligned}
\eta = \min\Bigg\{&\frac{1}{2H},\frac{B\sqrt{M}}{\sigma\sqrt{KR}},\max\left\{\frac{B}{G\sqrt{KR}},\frac{B}{\sqrt{HF_\star KR}}\right\},\\
&\frac{1}{\mathbb{1}_{K>1}}\cdot\max\Bigg\{\min\left\{\frac{B^{2/3}}{H^{1/3}\sigma^{2/3}K^{2/3}R^{1/3}},\frac{B^{2/3}}{H^{1/3}\zeta^{2/3}KR^{1/3}},\frac{B}{K^{3/4}\sqrt{\zeta\sigma R}},\frac{B}{\zeta K\sqrt{R}}\right\},\\
&\min\left\{\frac{B}{K^{3/4}\sqrt{G\sigma R}},\frac{B}{K\sqrt{\zeta GR}}\right\}\Bigg\}\Bigg\}
\end{aligned}
$$

This finishes the proof. ■

It is now straightforward to prove Theorem 2 similar to the proof for Theorem 1 by replacing σ^2 with G^2d:

Theorem 2 *Consider the problem class $\mathcal{P}^{0,2}_{M,K,R}(\mathcal{F}^{G,H,B},\zeta,F_\star)$. With appropriate η (c.f., lemma 6), the queried points $\{x_t^{m,j}\}_{t,m,j=1}^{T,M,2}$ of Algorithm 1 satisfy:*

$$
\begin{aligned}
\frac{1}{2MKR}\sum_{t\in[KR],m\in[M],j\in[2]}\mathbb{E}\left[f_t^m(x_t^{m,j})-f_t^m(x^\star)\right] \preceq\ & \frac{HB^2}{KR}+\frac{GB\sqrt{d}}{\sqrt{MKR}}+\frac{GB}{\sqrt{KR}}+\frac{\sqrt{HF_\star}B}{\sqrt{KR}}\\
&+\mathbb{1}_{K>1}\cdot\min\left\{\frac{H^1/3B^{4/3}G^{2/3}d^{1/3}}{K^{1/3}R^{2/3}}+\frac{H^{1/3}B^{4/3}\zeta^{2/3}}{R^{2/3}}\right.
\end{aligned}
$$

$$+ \frac{\sqrt{\zeta G} B d^{1/4}}{K^{1/4}\sqrt{R}} + \frac{\zeta B}{\sqrt{R}}, \frac{G B d^{1/4}}{K^{1/4}\sqrt{R}} + \frac{\sqrt{G\zeta} B}{\sqrt{R}} \Bigg\},$$

where $x^\star \in \arg\min_{x\in\mathbb{R}^d} \sum_{t\in[KR]} f_t(x)$*, and the expectation is w.r.t. the choice of function queries. The regret is also upper bounded as in theorem 1 for the corresponding step size.*