The Min-Max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication

Blake Woodworth BLAKE@TTIC.EDU
Toyota Technological Institute at Chicago

Brian Bullins BBULLINS@TTIC.EDU
Toyota Technological Institute at Chicago

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We resolve the min-max complexity of distributed stochastic convex optimization (up to a log factor) in the intermittent communication setting, where M machines work in parallel over the course of R rounds of communication to optimize the objective, and during each round of communication, each machine may sequentially compute K stochastic gradient estimates. We present a novel lower bound with a matching upper bound that establishes an optimal algorithm.

Keywords: Distributed Stochastic Convex Optimization, Oracle Complexity of Optimization

1. Introduction

The min-max oracle complexity of stochastic convex optimization in a sequential (non-parallel) setting is very well-understood, and we have provably optimal algorithms that achieve the min-max complexity (Lan, 2012; Ghadimi and Lan, 2013). However, we do not yet have an understanding of the min-max complexity of stochastic optimization in a *distributed* setting, where oracle queries and computation are performed by different workers, with limited communication between them. Perhaps the simplest, most basic, and most important distributed setting is that of *intermittent communication*.

In the (homogeneous) intermittent communication setting, M parallel workers are used to optimize a single objective over the course of R rounds. During each round, each machine sequentially and locally computes K independent unbiased stochastic gradients of the global objective, and then all the machines communicate with each other. This captures the natural setting where multiple parallel "workers" or "machines" are available, and computation on each worker is much faster than communication between workers. It includes applications ranging from optimization using multiple cores or GPUs, to using a cluster of servers, to Federated Learning[1] where workers are edge devices.

The intermittent communication setting has been widely studied for over a decade, with many optimization algorithms proposed and analyzed (Zinkevich et al., 2010; Cotter et al., 2011; Dekel

[1] In a realistic Federated Learning setting, stochastic gradient estimates on the same machine might be correlated, or we might prefer thinking of a heterogeneous setting where each device has a different local objective. Nevertheless, much of the methodological and theoretical development in Federated Learning has been focused on the homogeneous intermittent communication setting we study here (see Kairouz et al., 2019, and citations therein).

et al., 2012; Zhang et al., 2013a,c; Shamir and Srebro, 2014), and obtaining new methods and improved analysis is still a very active area of research (Wang et al., 2017; Stich, 2018; Wang and Joshi, 2018; Khaled et al., 2019; Haddadpour et al., 2019; Woodworth et al., 2020b). However, despite these efforts, we do not yet know which methods are optimal, what the min-max complexity is, and what methodological or analytical improvements might allow us to make further progress.

Considerable effort has been made to formalize the setting and establish lower bounds for distributed optimization (Zhang et al., 2013b; Arjevani and Shamir, 2015; Braverman et al., 2016) and here, we follow the graph-oracle formalization of Woodworth et al. (2018). However, a key issue in the existing literature is that known lower bounds for the intermittent communication setting depend only on the product KR (i.e. the total number of gradients computed on each machine over the course of optimization), and not on the number of rounds, R, and the number of gradients per round, K, separately.

Thus, existing results cannot rule out the possibility that the optimal rate for fixed $T = KR$ can be achieved using only a single round of communication ($R = 1$), since they do not distinguish between methods that communicate very frequently ($R = T$, $K = 1$) and methods that communicate just once ($R = 1$, $K = T$). The possibility that the optimal rate is achievable with $R = 1$ was suggested by Zhang et al. (2013c), and indeed Woodworth et al. (2020b) proved that an algorithm that communicates just once is optimal in the special case of quadratic objectives. While it seems unlikely that a single round of communication suffices in the general case, none of our existing lower bounds are able to answer this extremely basic question.

In this paper, we resolve (up to a logarithmic factor) the minimax complexity of smooth, convex stochastic optimization in the (homogeneous) intermittent communication setting. Our main result in Section 3 is a lower bound on the optimal rate of convergence and a matching upper bound. Interestingly, we show that the combination of two extremely simple and naïve methods based on an accelerated stochastic gradient descent (SGD) variant called AC-SA (Lan, 2012) is optimal up to a logarithmic factor. Specifically, we show that the better of the following methods is optimal: "Minibatch Accelerated SGD" which executes R steps of AC-SA using minibatch gradients of size MK, and "Single-Machine Accelerated SGD" which executes KR steps of AC-SA on just one of the machines, completely ignoring the other $M - 1$.

These methods might seem to be horribly inefficient: Minibatch Accelerated SGD only performs one update per round of communication, and Single-Machine Accelerated SGD only uses one of the available workers! This perceived inefficiency has prompted many attempts at developing improved methods which take multiple steps on each machine locally in parallel including, in particular, numerous analyses of Local SGD (Zinkevich et al., 2010; Dekel et al., 2012; Stich, 2018; Haddadpour et al., 2019; Khaled et al., 2019; Woodworth et al., 2020b). Nevertheless, we establish that one or the other is optimal in every regime, so more sophisticated methods cannot yield improved guarantees for arbitrary smooth objectives. Our results therefore highlight an apparent dichotomy between exploiting the available parallelism but not the local computation (Minibatch Accelerated SGD) and exploiting the local computation but not the parallelism (Single-Machine Accelerated SGD).

Our lower bound applies quite broadly, including to the settings considered by much of the existing work on stochastic first-order optimization in the intermittent communication setting. But, like many lower bounds, we should not interpret this to mean we cannot make progress. Rather, it indicates that we need to expand our model or modify our assumptions in order to develop better methods. In Section 5 we explore several additional assumptions that allow for circumventing our lower bound. These include when the third derivative of the objective is bounded (as in recent work by Yuan and Ma (2020)), when the objective has a certain statistical learning-like structure, or when the algorithm has access to a more powerful oracle.

2. Setting and Notation

We aim to understand the fundamental limits of stochastic first-order algorithms in the intermittent communication setting. Accordingly, we consider a standard smooth, convex problem

$$\min_x F(x) \tag{1}$$

where F is convex, $\|x^*\| \leq B$, and F is H-smooth, so for all x, y

$$F(x) + \langle \nabla F(x),\, y - x\rangle \leq F(y) \leq F(x) + \langle \nabla F(x),\, y - x\rangle + \frac{H}{2}\|y - x\|^2 \tag{2}$$

We consider algorithms that gain information about the objective via a stochastic gradient oracle g with bounded variance[2], which satisfies for all x

$$\mathbb{E}_z g(x;z) = \nabla F(x) \quad \text{and} \quad \mathbb{E}_z \|g(x;z) - \nabla F(x)\|^2 \leq \sigma^2 \tag{3}$$

This is a well-studied class of optimization objectives: smooth, bounded, convex objectives with a bounded-variance stochastic gradient oracle.

To understand optimal methods for this class of problems requires specifying a class of optimization algorithms. We consider intermittent communication algorithms, which attempt to optimize F using M parallel workers, each of which is allowed K queries to g in each of R rounds of communication. Such intermittent communication algorithms can be formalized using the graph oracle framework of Woodworth et al. (2018) which focuses on the dependence structure between different stochastic gradient computations.

Finally, we are considering a "homogeneous" setting, where each of the machines have access to stochastic gradients from the same distribution, in contrast to the more challenging "heterogeneous" setting, where they come from *different* distributions, which could arise in a machine learning context when each machine uses data from a different source. The heterogeneous setting is interesting, important, and widely studied, but we focus here on the more basic question of min-max rates for homogeneous distributed optimization. We point out that our lower bounds also apply to heterogeneous objectives since homogeneous optimization is a special case of heterogeneous optimization, and there are also some lower bounds specific to the heterogeneous setting (e.g. Arjevani and Shamir, 2015) but they do not apply to our setting.

3. The Lower Bound

We now present our main result, which is a lower bound on what suboptimality can be guaranteed by any distributed zero-respecting intermittent communication algorithm in the worst case:

[2] This assumption can be strong, and does not hold for natural problems like least squares regression (Nguyen et al., 2019), nevertheless, this strengthens rather than weakens our lower bound.

Theorem 1 *For any $H, B, \sigma, K, R > 0$ and $M \geq 2$, and any intermittent communication algorithm, there exists a convex, H-smooth objective which has a minimizer with norm at most B in any dimension*

$$d \geq 2KR + \left(10^9\left(1 + KR + \left(\frac{HB}{\sigma}\right)^{3/2} M(KR)^{5/4}\right) + \frac{6144H^2B^2MKR}{\sigma^2}\right)\log(64MK^2R^2)$$

and a stochastic gradient oracle, g, with $\mathbb{E}_z\|g(x;z) - \nabla F(x)\|^2 \leq \sigma^2$ such that the algorithm's output will have error at least

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2(1+\log^2 M)}, \frac{\sigma B}{\sqrt{KR}}\right\}\right)$$

for a numerical constant c.

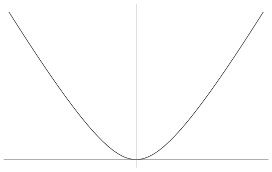
Proof Sketch The first two terms of this lower bound follow directly from previous work (Woodworth et al., 2018); the $\frac{HB^2}{K^2R^2}$ term corresponds to optimizing a function with a deterministic gradient oracle, and the $\frac{\sigma B}{\sqrt{MKR}}$ term is a very well-known statistical limit (see, e.g., Nemirovsky and Yudin, 1983). The distinguishing feature of our lower bound is the second $\min$ term, which depends differently on K than on R. For quadratics, the min-max complexity actually does depend only on the product KR, and is given by just the first two terms (Woodworth et al., 2020b). Consequently, proving our lower bound necessitates going beyond quadratics (in contrast, all the lower bounds for sequential smooth convex optimization that we are aware of can be obtained using quadratics). We therefore prove the Theorem using the following non-quadratic hard instance

$$F(x) = \psi'(-\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{N-1}\psi(x_{i+1} - x_i) \tag{4}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\psi(x) := \frac{\sqrt{H}x}{2\beta}\arctan\left(\frac{\sqrt{H}\beta x}{2}\right) - \frac{1}{2\beta^2}\log\left(1 + \frac{H\beta^2x^2}{4}\right) \tag{5}$$

and where β, ζ, and N are hyperparameters that are chosen depending on H, B, σ, M, K, R so that F satisfies the necessary conditions. This construction closely resembles the classic lower bound for deterministic first-order optimization of Nesterov (2004), which corresponds to $\psi(x) = x^2$. To describe our stochastic gradient oracle, we will use $\mathrm{prog}_\alpha(x) := \max\{j : |x_j| > \alpha\}$, which denotes the highest index of a coordinate of x that is significantly non-zero. We also define F^- to be equal to the objective with the $\mathrm{prog}_\alpha(x)^{\text{th}}$ term removed:



The function $\psi(x)$

$$F^-(x) = \psi'(-\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{\mathrm{prog}_\alpha(x)-1}\psi(x_{i+1} - x_i) + \sum_{i=\mathrm{prog}_\alpha(x)+1}^{N-1}\psi(x_{i+1} - x_i) \tag{6}$$

The stochastic gradient oracle for F that we use then resembles

$$g(x) = \begin{cases}\nabla F^-(x) & \text{with probability } 1-p \\ \nabla F(x) + \frac{1-p}{p}(\nabla F(x) - \nabla F^-(x)) & \text{with probability } p\end{cases} \tag{7}$$

This stochastic gradient oracle is similar to the one used by Arjevani et al. (2019) to prove lower bounds for non-convex optimization, and its key property is that $\mathbb{P}[\mathrm{prog}_\alpha(g(x)) \leq \mathrm{prog}_\alpha(x)] = 1-p$. Therefore, each oracle access only reveals information about the next coordinate of the gradient the algorithm with probability p, and therefore the algorithm is essentially only able to make progress with probability p. The rest of the proof revolves around bounding the total progress of the algorithm and showing that if $\mathrm{prog}_\alpha(x) \leq \frac{N}{2}$, then x has high suboptimality.

Since each machine makes KR sequential queries and only makes progress with probability p, the total progress scales like $KR \cdot p$. By taking p smaller, we decrease the amount of progress made by the algorithm, and therefore increase the lower bound. Indeed, when $p \approx 1/K$, the algorithm only increases its progress by about $\log M$ per round, which gives rise to the key $(HB^2)/(R^2 \log^2 M)$ term in the lower bound. However, we are constrained in how small we can take p since our stochastic gradient oracle has variance

$$\sup_x \mathbb{E}\|g(x) - \nabla F(x)\|^2 \approx \frac{1}{p} \sup_x \psi'(x)^2 \tag{8}$$

This is where our choice of ψ comes in. Specifically, we chose the function ψ to be convex and smooth so that F is, but we also made it Lipschitz:

$$\psi'(x) = \frac{\sqrt{H}}{2\beta} \arctan\left(\frac{\sqrt{H}\beta x}{2}\right) \in \left[-\frac{\pi\sqrt{H}}{4\beta}, \frac{\pi\sqrt{H}}{4\beta}\right] \tag{9}$$

Notably, this Lipschitz bound on ψ, which implies a bound on $\|\nabla F(x)\|_\infty$, is the key non-quadratic property that allows for our lower bound. Since ψ' is bounded, we are able to able to choose $p \approx H\sigma^{-2}\beta^{-2}$ without violating the variance constraint on the stochastic gradient oracle. Carefully balancing β completes the argument.

Another important aspect of our lower bound is that it applies to arbitrary randomized algorithms, rather than more restricted families of algorithms like "zero-respecting" methods (see Appendix D). We therefore prove our theorem using techniques similar to Woodworth and Srebro (2016), Carmon et al. (2017), Arjevani et al. (2019), and others, who introduce a random rotation matrix, U; construct a hard instance like $F(U^\top x)$; and argue that any algorithm behaves almost as if it were zero-respecting. For further discussion of this proof technique, we refer readers to (Woodworth and Srebro, 2016; Carmon et al., 2017). All of the details of the proof can be found in Appendices A-C.

Theorem 1 also implies a lower bound for strongly convex objectives:

Corollary 2 *There is a numerical constant, c, such that no intermittent communication algorithm can guarantee for any H-smooth, λ-strongly convex objective F and stochastic gradient oracle with variance less than σ^2 that its output will have suboptimality*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{F(0)-F^*}{K^2R^2} \exp\left(-\sqrt{\frac{\lambda}{H}}KR\right) + \frac{\sigma^2}{\lambda MKR} \right.$$
$$\left. + \min\left\{\frac{F(0)-F^*}{R^2\log^2 M} \exp\left(-\sqrt{\frac{\lambda}{H}}R\log M\right), \frac{\sigma^2}{\lambda KR}\right\}\right)$$

This lower bound is more limited than Theorem 1, since we prove it using a reduction from convex to strongly convex optimization, rather than directly. We also do not expect the exponential terms to be tight. Nevertheless, the Corollary gives some indication of the optimal rate in the strongly convex setting and, as with Theorem 1, it distinguishes between R and K unlike previous results. A simple proof can be found in Appendix C.

4. A Matching Upper Bound and an Optimal Algorithm

The lower bound in Theorem 1 is matched (up to $\log$ factors) by the combination of two simple distributed zero-respecting algorithms, which are distributed variants of an accelerated SGD algorithm called AC-SA due to Lan (2012). In the sequential setting, AC-SA algorithm maintains two iterates y_t and x_t which it updates according to

$$\begin{aligned} y_{t+1} &= y_t - \gamma_t g_t\big(\beta_t^{-1} y_t + (1-\beta_t^{-1})x_t\big) \\ x_{t+1} &= \beta_t^{-1} y_{t+1} + (1-\beta_t^{-1})x_t \end{aligned} \tag{10}$$

where γ_t and β_t are carefully chosen stepsize parameters. In the smooth, convex setting, this algorithm converges at a rate (see Corollary 1, Lan, 2012)

$$\mathbb{E}[F(x_T) - F^*] \leq c \cdot \left(\frac{HB^2}{T^2} + \frac{\sigma B}{\sqrt{T}}\right) \tag{11}$$

To describe the optimal algorithm for the intermittent communication setting, we will first define two distributed variants of AC-SA.

The first algorithm, which we will refer to as **Minibatch Accelerated SGD**, implements R iterations of AC-SA using minibatch gradients of size MK (c.f. Cotter et al., 2011). Specifically, the method maintains two iterates y_r and x_r which are shared across all the machines. During each round of communication, each machine computes K independent stochastic estimates of $\nabla F\big(\beta_r^{-1} y_r + (1-\beta_r^{-1})x_r\big)$; the machines then communicate their minibatches, averaging them together into a larger minibatch of size MK, and then they update y_r and x_r according to (10). Because the minibatching reduces the variance of the stochastic gradients by a factor of MK, (11) implies this method converges at a rate

$$\mathbb{E}[F(x_R) - F^*] \leq c \cdot \left(\frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}}\right) \tag{12}$$

The second algorithm, which we will call **Single-Machine Accelerated SGD**, "parallelizes" AC-SA in a different way. In contrast to Minibatch Accelerated SGD, Single-Machine Accelerated SGD simply ignores $M-1$ of the available machines and runs $T = KR$ steps of AC-SA on the remaining one, therefore converging like

$$\mathbb{E}[F(x_{KR}) - F^*] \leq c \cdot \left(\frac{HB^2}{K^2R^2} + \frac{\sigma B}{\sqrt{KR}}\right) \tag{13}$$

From here, we point out that lower bound in Theorem 1 is equal (up to $\log$ factors) to the minimum of (12) and (13). Furthermore, one can determine which of these algorithms achieves the minimum based on the problem parameters:

Theorem 3 *For any $H, B, \sigma, K, R, M > 0$, the algorithm which returns the output of Minibatch Accelerated SGD when $K \leq \frac{\sigma^2 R^3}{H^2B^2}$ and returns the output of Single-Machine Accelerated SGD when $K > \frac{\sigma^2 R^3}{H^2B^2}$ is optimal up to a factor of $O(\log^2 M)$.*

This optimal algorithm is computationally efficient and requires no significant overhead. Each machine needs to store only a constant number of vectors, it performs only a constant number of vector additions for each stochastic gradient oracle access, and it communicates just one vector per round. Therefore, the total storage complexity is $O(d)$ per machine, the sequential runtime complexity is $O(KR \cdot d)$, and the total communication complexity is $O(MR \cdot d)$. In fact, the communication complexity is 0 when Single-Machine Accelerated SGD is used. Therefore, we do not expect a substantially better algorithm from the standpoint of computational efficiency either.

In light of Theorem 3 and the second $\min$ term in Theorem 1, we see that algorithms in this setting are offered the following dilemma: they may either attain the optimal statistical rate $\sigma B/\sqrt{MKR}$ but suffer an optimization rate $HB^2/(R^2 \log^2 M)$ that does not benefit from K at all, or they may attain the optimal optimization rate of $HB^2/(K^2R^2)$ but suffer a statistical rate $\sigma B/\sqrt{KR}$ as if only single machine were available. In this sense, there is a very real dichotomy between exploiting parallelism and leveraging local computation.

The main shortcoming of the optimal algorithm is the need to know the problem parameters H, B, and σ to implement it. However, knowledge of these parameters is anyway needed in order to choose the stepsizes for AC-SA, and we are not aware of accelerated variants of SGD that can be implemented without knowing them, even in the sequential setting. This algorithm is also somewhat unnatural because of the hard switch between Minibatch and Single-Machine Accelerated SGD. It would be nice, if only aesthetically, to have an algorithm that more naturally transitions from the Minibatch to the Single-Machine rate. Accelerated Local SGD (Yuan and Ma, 2020) or something similar is a contender for such an algorithm, although it is unclear whether or not this method can match the optimal rate in all regimes. Local SGD methods can also be augmented by using two stepsizes—a smaller, conservative stepsize for the local updates between communications, and a larger, aggressive stepsize when the local updates are aggregated—this two-stepsize approach allows for interpolation between Minibatch-like and Single-Machine-like behavior, and could be used to design a more “natural” optimal algorithm (see Section 6, Woodworth et al., 2020a).

Finally, the upper and lower bounds match up to a factor of $\log^2 M$. While this is generally a minor gap, it does raise the question of what the optimal error would be in a massively parallel regime where exponentially many machines are available. In this case, it is conceivable that a brute-force approach might be available that could converge at the rate $1/(R \log M)^2$ in a certain regime, as is suggested by the lower bound, improving over the $1/R^2$ rate achieved by Minibatch Accelerated SGD. Nevertheless, it is not obvious how this could be achieved without any dependence on the dimension and related work by Duchi et al. (2018) suggests that such a rate would not be possible without depending on the dimension. We therefore conjecture that the $\log^2 M$ factor can be removed from the lower bound.

5. Better than Optimal: Breaking the Lower Bound

Perhaps the most important use of a lower bound is in understanding how to break it. Instead of viewing the lower bound as telling us to give up any hope of improving over the naïve optimal method in Section 4, we should view it as informing us about possible means of making progress.

One way to break our lower bound is by introducing additional assumptions that are not satisfied by the hard instance. These assumptions could then be used to establish when and how some alternate method improves over the “optimal” method in Section 4. Several methods, which operate within the intermittent communication framework of Section 2, have been shown to be better than the "optimal algorithm" in practice *for specific instances*. However, attempts to demonstrate the benefit of these methods theoretically have so far failed, and we now understand why. In order to understand such benefits, we *must* introduce additional assumptions, and ask not "is this alternate method better" but rather "under what assumption is this alternate method better?" Below we suggest possible additional assumptions, including ones that have appeared in recent analysis and also other plausible assumptions one could rely on.

Another way to break the lower bound is by considering algorithms that go beyond the stochastic oracle framework of Section 2, utilizing more powerful oracles that nevertheless could be equally easy to implement. Understanding the lower bound can inform us of what type of such extensions might be useful, thus guiding development of novel types of optimization algorithms.

5.1. Relying on a Bounded Third Derivative

As we have mentioned, Theorem 1 does not hold in the special case of quadratic objectives of the form $Q(x) = \frac{1}{2}x^\top Ax + b^\top x$ for p.s.d. A, e.g. least squares problems, in which case the min-max rate is much better, and Accelerated Local SGD achieves:

$$\mathbb{E}Q(\hat{x}) - Q^* \leq c \cdot \left(\frac{HB^2}{K^2R^2} + \frac{\sigma B}{\sqrt{MKR}}\right) \tag{14}$$

Since improvement over the lower bound is possible when the objective is *exactly* quadratic, it stands to reason that similar improvement should be possible when the objective is sufficiently *close* to quadratic. Indeed, Yuan and Ma (2020) analyze another accelerated variant of Local SGD in the smooth, convex setting with the additional assumption that the Hessian $\nabla^2 F(x)$ is α-Lipschitz. Their algorithm converges at a rate

$$\mathbb{E}F(\hat{x}) - F^* \leq \tilde{O}\left(\frac{HB^2}{KR^2} + \frac{\sigma B}{\sqrt{MKR}} + \left(\frac{H\sigma^2B^4}{MKR^3}\right)^{1/3} + \left(\frac{\alpha\sigma^2B^5}{R^4K}\right)^{1/3}\right) \tag{15}$$

This *can* improve over the lower bound in Theorem 1 in certain parameter regimes, for instance, (15) is better if

$$\frac{H^2B^2}{\sigma^2} \leq \frac{R^3}{MK} \qquad \text{and} \qquad \alpha \leq \tilde{O}\left(\min\left\{\frac{\sigma R^{5/2}}{B^2K^{1/2}}, \frac{H^3BK}{\sigma^2R^2\log^6 M}\right\}\right) \tag{16}$$

However, Yuan and Ma's guarantee does not *always* improve over the lower bound, and it is not completely clear to what extent further improvement over their algorithm might be possible. In an effort to understand when it may or may not be possible to improve, we extend our lower bound to the case where $\nabla^2 F$ is α-Lipschitz:

Theorem 4 *For any $H, B, \sigma, Q, K, R > 0$ and any $M \geq 2$, there exists a convex, H-smooth objective F with $\|x^*\| \leq B$ and with $\nabla^2 F$ being Q-Lipschitz with respect to the L2 norm, and a stochastic gradient oracle g with $\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ for all x, such that with probability at least $\frac{1}{2}$ all of the oracle queries $\{x_{k,r}^m\}$ made by any distributed-zero-respecting intermittent* *communication algorithm (see Definition 18 in Appendix D) will have suboptimality*

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sqrt{Q\sigma}B^2}{K^{1/4}R^2\log^{7/4} M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right]$$

We prove this lower bound in Appendix D using the same construction (4) as we used for Theorem 1, but using the parameter β to control the third derivative of F. This lower bound does not match the guarantee of Yuan and Ma's algorithm, so it does not resolve the min-max complexity. However, there is reason to suspect that the lower bound is closer to the min-max rate, at least in certain regimes. For instance, when Q is taken to zero, i.e. the objective becomes quadratic, we know that Theorem 4 is tight while (15) can be larger by a factor of K. For that reason, we suspect that (15) is suboptimal, but further analysis will be needed. At any rate, our lower bound does establish that there is a limit to the utility of assuming a Lipschitz Hessian. Specifically, there can be no advantage over the optimal algorithm from Section 4 once $Q \geq O\left(\max\left\{\frac{H^2\sqrt{K}}{\sigma\log^{1/2} M}, \frac{\sigma R^3 \log^{7/2} M}{B^2\sqrt{K}}\right\}\right)$.

Theorem 4 and Yuan and Ma's algorithm also highlight a substantial qualitative difference between distributed and sequential optimization: in the sequential setting, there is never any advantage to assuming that the objective is close to quadratic. In fact, worst-case instances for sequential optimization are exactly quadratic (Nemirovsky and Yudin, 1983; Nesterov, 2004; Simchowitz, 2018).

Beyond requiring that the Hessian be Lipschitz, there are other ways of measuring an objective's closeness to a quadratic. Two notable examples are self-concordance (Nesterov, 1998) and quasi-self-concordance (Bach et al., 2010), which bound the third derivative of F in terms of the second derivative: we say that F is Q-self-concordant when for all x, v, $f(t) = F(x + tv)$ satisfies $|f'''(t)| \leq 2Qf''(t)^{3/2}$ and we say it is Q-quasi-self-concordant if $|f'''(t)| \leq Qf''(t)$. There has been recent interest in such objectives (Bach et al., 2010; Zhang and Xiao, 2015; Karimireddy et al., 2018; Carmon et al., 2020) which arise e.g. in logistic regression problems. In Appendix D, we extend the lower bound in Theorem 4 to these settings.

5.2. Statistical Learning Setting: Assumptions on Components

Stochastic optimization commonly arises in the context of statistical learning, where the goal is to minimize the expected loss with respect to a model's parameters. In this case, the objective can be written $F(x) = \mathbb{E}_{z\sim\mathcal{D}} f(x; z)$, where $z \sim \mathcal{D}$ represents data drawn i.i.d. from an unknown distribution, and the "components" $f(x; z)$ represent the loss of the model parametrized by x on the example z.

In the setting of Theorem 1, we only place restrictions on the F itself, and on the first and second moments of g. However, in the statistical learning setting, it is often natural to assume that the loss function $f(\cdot; z)$ itself satisfies particular properties *for each* z *individually*. For instance, in our setting we might assume f is convex and smooth and furthermore that the gradient oracle is given by $g(x) = \nabla f(x; z)$ for an i.i.d. $z \sim \mathcal{D}$. This is a non-trivial restriction on the stochastic gradient oracle, and it is conceivable that this property could be leveraged to design and analyze a method that converges faster than the lower bound in Theorem 1 would allow.

In particular, the specific stochastic gradient oracle (7) used to prove Theorem 1 *cannot* be written as the gradient of a random smooth function. In this sense, the lower bound construction is somewhat "unnatural," however, we are not aware of any analysis that meaningfully[3] exploits the fact that $g = \nabla f(\cdot; z)$. An interesting question is whether such an assumption can be used to prove a better convergence guarantee, or whether Theorem 1 can be proven using a stochastic gradient oracle that obeys this constraint.

5.3. Statistical Learning Setting: Repeated Access to Components

In the statistical learning setting, it is also natural to consider algorithms that can evaluate the gradient at multiple points for the same datum z. Specifically, allowing the algorithm access to a pool of samples $z_1, \dots, z_N$ drawn i.i.d. from $\mathcal{D}$ and to compute $\nabla f(x; z)$ for any chosen x and z_n opens up additional possibilities. Indeed, Arjevani et al. (2019) showed that multiple—even just two—accesses to each component enables substantially faster convergence ($T^{-1/3}$ vs. $T^{-1/4}$) in sequential stochastic non-convex optimization. Similar results have been shown for zeroth-order and bandit convex optimization (Agarwal et al., 2010; Duchi et al., 2015; Shamir, 2017; Nesterov and Spokoiny, 2017), where accessing each component twice allows for a quadratic improvement in the dimension-dependence.

In sequential smooth convex optimization, if F has "finite-sum" structure (i.e. $\mathcal{D}$ is the uniform distribution on $\{1, \dots, N\}$), then allowing the algorithm to pick a component and access it multiple times opens the door to variance-reduction techniques like SVRG (Johnson and Zhang, 2013). These methods have updates of the form:

$$x_{t+1} = x_t - \eta_t(\nabla f(x_t; z_t) - \nabla f(\tilde{x}; z_t) + \nabla F(\tilde{x})) \tag{17}$$

Computing this update therefore requires evaluating the gradient of $f(x; z_t)$ at two different points, which necessitates multiple accesses to a chosen component. This stronger oracle access allows faster rates compared with a single-access oracle (see discussion in, e.g., Arjevani et al., 2020).

Most relevantly, in the intermittent communication setting, distributed variants of SVRG are able to improve over the lower bound in Theorem 1 (Wang et al., 2017; Lee et al., 2017; Shamir, 2016; Woodworth et al., 2018). For example, in the intermittent communication setting when f is H-smooth and L-Lipschitz, and where the algorithm can access each component multiple times, Woodworth et al. show that using distributed SVRG to optimize an empirical objective composed of suitably many samples is able to achieve convergence at the rate

$$\mathbb{E}F(\hat{x}) - F^* \le c \cdot \left(\left(\frac{HB^2}{RK} + \frac{LB}{\sqrt{MKR}} \right) \log \frac{MKR}{LB} \right) \tag{18}$$

While this guarantee (necessarily!) holds in a different setting than Theorem 1, the Lipschitz bound L is generally analogous to the standard deviation of the stochastic gradient variance, σ (indeed, L is an upper bound on σ). With this in mind, this distributed SVRG algorithm can beat the lower bound in Theorem 1 when σ, L, and K are sufficiently large.

[3] Numerous papers assume that $F(x) = \mathbb{E}_{z\sim\mathcal{D}} f(x; z)$ and $g = \nabla f(\cdot; z)$ for some smooth, convex f (e.g. Bottou et al., 2018; Nguyen et al., 2019; Koloskova et al., 2020; Woodworth et al., 2020a). Nevertheless, the *purpose* of this assumption is to bound $\mathbb{E}\|g(x)\|^2$ or $\mathbb{E}\|g(x) - \nabla F(x)\|^2$ in terms of $\sigma_*^2 = \mathbb{E}\|g(x^*)\|^2$. In other words, one could prove the same guarantees in the setting of Theorem 1 with the additional constraint of the form $\mathbb{E}\|g(x)\|^2 \le \sigma_*^2 + \Gamma\|x - x^*\|^2$ for some parameter Γ. Since the variance of the gradient oracle in our lower bound construction is bounded everywhere by a constant σ^2, it therefore applies to these analyses.

5.4. Higher Order and Other Stronger Oracles

Yet another avenue for improved algorithms in the intermittent communication setting is to use stronger stochastic oracles. For instance, a stochastic second-order oracle that estimates $\nabla^2 F(x)$ (Hendrikx et al., 2020) or a stochastic Hessian-vector product oracle that estimates $\nabla^2 F(x)v$ given a vector v, which can typically be computed as efficiently as stochastic gradients. In the statistical learning setting, some recent work also considers a stochastic prox oracle which returns $\arg\min_y f(y;z) + \frac{1}{2}\|x-y\|^2$ (Wang et al., 2017; Chadha et al., 2021).

As an example, a stochastic Hessian-vector product oracle, in conjunction with a stochastic gradient oracle can be used to efficiently implement a distributed Newton algorithm. Specifically, the Newton update $x_{t+1} = x_t - \eta_t \nabla^2 F(x_t)^{-1}\nabla F(x_t)$ can be rewritten as

$$x_{t+1} = x_t + \eta_t \arg\min_y \left\{ \frac{1}{2} y^\top \nabla^2 F(x_t) y + \nabla F(x_t)^\top y \right\} \tag{19}$$

That is, each update can be viewed as the solution to a quadratic optimization problem, and its stochastic gradients can be computed using stochastic Hessian-vector and gradient access to F. The DiSCO algorithm (Zhang and Xiao, 2015) uses distributed preconditioned conjugate gradient descent to find an approximate Newton step. Alternatively, as previously discussed, this quadratic can be minimized to high accuracy using a single round of communication using Accelerated Local SGD. Under suitable assumptions (e.g., that F is convex, smooth and self-concordant), this algorithm may converge substantially faster than the lower bounds in Theorems 1 and 4 would allow for first-order methods.

Differences from Sequential Setting: Interestingly, in the sequential setting there is no benefit to using stochastic Hessian-vector products over and above what can be achieved using just a stochastic gradient oracle. This is because the worst-case instances are simply quadratic, in which case Hessian-vector products and gradients are essentially equivalent. This adds to a list of structures that facilitate distributed optimization while being essentially useless in the sequential setting. Likewise, objectives being quadratic or near-quadratic facilitates distributed optimization but does not help sequential algorithms since, again, the hard instances for sequential optimization are already quadratic. Furthermore, accessing a statistical learning gradient oracle $\nabla f(\cdot;z)$ multiple times can allow for faster distributed algorithms—e.g. distributed SVRG or using the stochastic gradients to implement stochastic Hessian-vector products via finite-differencing—but it does not generally help in the sequential case without further assumptions (like the problem having finite-sum structure).

5.5. Beyond Single-Sample Oracles

Another class of distributed optimization algorithms, which includes ADMM (Boyd et al., 2011) and DANE (Shamir et al., 2014), involve solving an optimization problem on each machine $m = 1..M$ at each round r of the form

$$\min_x \frac{1}{K}\sum_{k=1}^{K} f(x; z_{k,r}^m) + \lambda_{r,m}\|x - y_{r,m}\|^2, \tag{20}$$

where $f(\cdot;z)$ are components of the objective $F(x) = \mathbb{E}f(x,z)$, and the vectors $y_{r,m}$ and scalars $\lambda_{r,m}$ are chosen by the algorithm. Although these methods also involve processing K samples, or components, at each round on each machine, and then communicating between the machines, they are quite distinct from the stochastic optimization algorithms we consider, and fall well outside the "stochastic optimization with intermittent communication" model we study. The main distinction is that in this paper we are focused on stochastic optimization methods, where each oracle access or "atomic operation" involves a single "data point" $z_{k,r}^m$ (a single component of a stochastic objective), or in our first-order model, a single stochastic gradient estimate, and can generally be performed in time $O(d)$, where d is the dimensionality of x. In particular, each round consists of K separate accesses, and in all the methods we consider, can be implemented in time $O(dK)$. In contrast, (20) is a complex optimization problem involving many data points, and cannot be solved with $O(K)$ atomic operations[4]. This distinction results in the first term of the lower bound in Theorem 1, namely the "optimization term" $HB^2/(K^2R^2)$, not applying for methods using (20). In particular, even ignoring $M-1$ machines and running the Mini-Batch Prox method (Wang et al., 2017) on a single machine results ensures a suboptimality of

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\frac{\sigma B}{\sqrt{KR}}\right), \tag{21}$$

entirely avoiding the first term of Theorem 1, and beating the lower bound when σ^2 is small.

Another difference is that DANE, as well as other methods which target Empirical Risk Minimization such as DiSCO (Zhang and Xiao, 2015) and AIDE (Reddi et al., 2016), work on the same batch of K examples per machine in all rounds, i.e. they use $z_{k,r}^m = z_k^m$ with only KM (rather than KRM) random samples $\{z_k^m\}_{k\in[K],m\in[M]}$. In our setup and terminology, they thus require repeated access to components, as discussed above in Section 5.3. Furthermore, since they only use KM samples overall, they cannot guarantee suboptimality better than $\sigma B/\sqrt{KM}$, a factor of $\sqrt{R}$ worse than the second term in Theorem 1.

The Mini-Batch Prox guarantee (21) is disappointing, and suboptimal, once σ^2 and M are large, and DANE is not optimal, at least when R is large. Understanding the min-max complexity of the class of methods which solve (20) at each round on each machine thus remains an important and interesting open problem. We note that lower bounds and the optimality of some of these methods were studied in Arjevani and Shamir (2015), but in a somewhat different, non-statistical distributed setting.

Acknowledgments

BW is supported by a Google PhD Research Fellowship. This work is also partially supported by NSF-CCF/BSF award 1718970/2016741, and NS is also supported by a Google Faculty Research Award.

References

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. Citeseer, 2010.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 1756–1764, 2015.

[4] It could perhaps be approximately solved using a small number of passes over the K data points, which would put us back within the scope what we study in this paper, but that is not how these method are generally analyzed.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Yossi Arjevani, Amit Daniely, Stefanie Jegelka, and Hongzhou Lin. On the complexity of minimizing convex finite sums without using the indices of the individual functions. *arXiv preprint arXiv:2002.03273*, 2020.

Francis Bach et al. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31: 1–58, 1997.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020, 2016.

Yair Carmon. *The Complexity of Optimization Beyond Convexity*. Stanford University, 2020.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017. URL https://arxiv.org/abs/1710.11606.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *arXiv preprint arXiv:2003.08078*, 2020.

Karan Chadha, Gary Cheng, and John C Duchi. Accelerated, optimal, and parallel: Some results on model-based stochastic optimization. *arXiv preprint arXiv:2101.02696*, 2021.

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1647–1655. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4432-better-mini-batch-algorithms-via-accelerated-gradient-methods.pdf.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

John Duchi, Feng Ruan, and Chulhee Yun. Minimax bounds on stochastic batched convex optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning*

Research, pages 3065–3162. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/duchi18a.html.

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.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, pages 11080–11092, 2019.

Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *International Conference on Machine Learning*, pages 4203–4227. PMLR, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013. URL https://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduct pdf.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurlien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adri Gascn, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konen, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrde Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer zgr, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019.

Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Global linear convergence of newton's method without strong-convexity or lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Better communication complexity for local sgd. *arXiv preprint arXiv:1909.04746*, 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012. URL https://pdfs.semanticscholar.org/1621/f05894ad5fd6a8fcb8827a8c7aca36c81775.pdf.

Jason D Lee, Qihang Lin, Tengyu Ma, and Tianbao Yang. Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *The Journal of Machine Learning Research*, 18(1):4404–4446, 2017.

Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

Yurii Nesterov. Introductory lectures on convex optimization: a basic course. 2004.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Lam M Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, Martin Takác, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019.

Sashank J Reddi, Jakub Konečnỳ, Peter Richtárik, Barnabás Póczós, and Alex Smola. Aide: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

O. Shamir and N. Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 850–857, 2014. doi: 10.1109/ALLERTON.2014.7028543.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 46–54, 2016.

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.

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.

Max Simchowitz. On the randomized complexity of minimizing a convex quadratic function. *arXiv preprint arXiv:1807.09386*, 2018.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018. URL https://arxiv.org/abs/1805.09767.

Jialei Wang, Weiran Wang, and Nathan Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Conference on Learning Theory*, pages 1882–1919. PMLR, 2017.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Blake Woodworth, Jialei Wang, Brendan McMahan, and Nathan Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *arXiv preprint arXiv:1805.10222*, 2018. URL https://arxiv.org/abs/1805.10222.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020a.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020b.

Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3639–3647. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6058-tight-complexity-bounds-for-optimizing-composite-objectives.pdf.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *arXiv preprint arXiv:2006.08950*, 2020.

Yuchen Zhang and Lin Xiao. Disco: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning*, pages 362–370. PMLR, 2015.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617, 2013a.

Yuchen Zhang, John C Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, pages 2328–2336. Citeseer, 2013b.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013c.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Appendix A. A Framework for Proving Lower Bounds for Randomized Algorithms

Our lower bounds are based on the idea of showing that in sufficiently high dimension, any randomized algorithms behaves almost as if it is zero-respecting, meaning that its oracle queries are close to the subspace spanned by the previously-seen oracle responses. To formalize this, for a vector x, we define its α-progress as

$$\mathrm{prog}_\alpha(x) := \max\{i : |x_i| > \alpha\} \tag{22}$$

Our lower bound proceeds by showing that any intermittent communication algorithm will fail to achieve a high amount of progress, even for randomized algorithms that leave the span of previous stochastic gradient queries. We now define the properties that we will require from our oracle construction:

Definition 5 *A stochastic zeroth- and first-order oracle that returns $(f(x;z), g(x;z))$ for $z \sim \mathcal{D}$ is an (α, p, δ)-robust-zero-chain if there exist sets $\mathcal{Z}_0, \mathcal{Z}_1$ such that*

1. $\mathbb{P}(z \in \mathcal{Z}_0 \cup \mathcal{Z}_1) \geq 1 - \delta$
2. $\mathbb{P}(z \in \mathcal{Z}_0 | z \in \mathcal{Z}_0 \cup \mathcal{Z}_1) \geq 1 - p$
3. *For all $z \in \mathcal{Z}_0$ and all x, $\operatorname{prog}_0(g(x;z)) \leq \operatorname{prog}_\alpha(x)$ and there exist functions $f_1, f_2, \dots$ and $g_1, g_2, \dots$ such that*
$$\operatorname{prog}_\alpha(x) \leq i \implies \begin{cases} f(x;z) = f_i(x_1, x_2, \dots, x_i; z) \\ g(x;z) = g_i(x_1, x_2, \dots, x_i; z) \end{cases}$$
4. *For all $z \in \mathcal{Z}_1$ and all x, $\operatorname{prog}_0(g(x;z)) \leq \operatorname{prog}_\alpha(x) + 1$ and there exist functions $f_1, f_2, \dots$ and $g_1, g_2, \dots$ such that*
$$\operatorname{prog}_\alpha(x) \leq i \implies \begin{cases} f(x;z) = f_i(x_1, x_2, \dots, x_{i+1}; z) \\ g(x;z) = g_i(x_1, x_2, \dots, x_{i+1}; z) \end{cases}$$

In this section, our main result is to show that any algorithm that interacts with a robust zero chain will have low progress:

Lemma 6 *Let (f, g) be an (α, p, δ)-robust-zero-chain, let $U \in \mathbb{R}^{D \times d}$ be a uniformly random orthogonal matrix with $U^\top U = I_{d \times d}$ for $D \geq d + \frac{2\gamma^2}{\alpha^2} \log(32MKRd)$, and let $x_{k,r}^m$ be the k^{th} oracle query on the m^{th} machine during the r^{th} round of communication for an intermittent communication algorithm that interacts with the oracle (f_U, g_U) for $f_U(x;z) = f(U^\top x; z)$ and $g_U(x;z) := Ug(U^\top x; z)$. Then if $\max_{m,k,r} \|x_{k,r}^m\| \leq \gamma$, then the algorithm's output $\hat{x}$ will have progress at most*

$$\mathbb{P}\Big(\operatorname{prog}_\alpha(U^\top \hat{x}) \leq \min\{KR,\ 8KRp + 12R \log M + 12R\}\Big) \geq \frac{5}{8} - 2MKR\delta$$

The main ideas leading to Lemma 6 stem from Woodworth and Srebro (2016) and Carmon et al. (2017), who show that when a random rotation is applied to the objective and the dimension is sufficiently large, every algorithm behaves essentially as if its queries remained in the span of previously seen gradients. In the original arguments, the proof of this claim was extremely complicated and required a great deal of care due to subtleties with conditioning on the stochastic gradient oracle queries. Since then, the argument has gradually be refined and simplified, culminating in Carmon (2020) who presents the simplest argument to date. The proof of 6 therefore resembles the proof of (Proposition 2.4 Carmon, 2020), however, the arguments must be extended to accomodate the intermittent communication setting.

To facilitate our proofs, we introduce some notation. Recalling $\mathcal{Z}_0$ and $\mathcal{Z}_1$ from Definition 5, we define

$$S^m_{k,r} = \min\left\{d,\ \sum_{k'=1}^{k-1} \mathbb{1}\{z^m_{k',r} \in \mathcal{Z}_1\} + \sum_{r'=1}^{r-1} \max_{1\leq m'\leq M} \sum_{k'=1}^{K} \mathbb{1}\{z^{m'}_{k',r'} \in \mathcal{Z}_1\}\right\} \tag{23}$$

We will show that this quantity $S^m_{k,r}$ essentially upper bounds the α-progress of the k^{th} query during the r^{th} round of communication on the m^{th} machine. We also define the following "good events" where the progress of the algorithm's oracle queries remains small

$$\mathcal{G}^m_{k,r} = \left\{\operatorname{prog}_\alpha(U^\top x^m_{k,r}) \leq S^m_{k,r}\right\} \tag{24}$$

$$\bar{\mathcal{G}}^m_{k,r} = \bigcap_{k'<k} \mathcal{G}^m_{k',r} \cap \bigcap_{r'<r} \bigcap_{m',k'} \mathcal{G}^{m'}_{k',r'} \tag{25}$$

We also define the event

$$Z = \left\{\forall_{m,k,r}\ z^m_{k,r} \in \mathcal{Z}_0 \cup \mathcal{Z}_1\right\} \tag{26}$$

Finally, we use

$$U_{\leq i} = [U_1, U_2, \dots, U_i, 0, \dots, 0] \tag{27}$$

to denote the matrix U with the $(i+1)^{\text{th}}$ through d^{th} columns replaced by zeros.

We begin by showing that when the good events $\bar{\mathcal{G}}^m_{k,r}$ happen, the algorithm's queries are determined by only a subset of the columns of U.

Lemma 7 *Let (f, g) be an (α, p, δ)-robust-zero-chain, let $U \in \mathbb{R}^{D\times d}$ be a uniformly random orthogonal matrix with $U^\top U = I_{d\times d}$, and let $x^m_{k,r}$ be the k^{th} oracle query on the m^{th} machine during the r^{th} round of communication for an intermittent communication algorithm that interacts with the oracle (f_U, g_U) for $f_U(x; z) = f(U^\top x; z)$ and $g_U(x; z) := Ug(U^\top x; z)$. Then conditioned on $\mathcal{S}$—the σ-algebra generated by $\{S^m_{k,r}\}_{m,k,r}$—the events $\bar{G}^m_{k,r}$ and Z, and the query $x^m_{k,r}$ is a measurable function of ξ—the algorithm's random coins—and $U_{\leq S^m_{k,r}}$. Similarly, conditioned on $\mathcal{S}$, Z, and $\bigcap_{m=1}^M \bar{G}^m_{K,R}$, the output of the algorithm, $\hat{x}$ is a measurable function of ξ and $U_{\leq \max_m S^m_{K,R}}$.*

Proof The dependence structure of an intermittent communication algorithm's queries is determined by the communication between the machines. Specifically, the m^{th} machine's k^{th} query during the r^{th} round of communication can only depend on oracle queries that have been communicated to that machine at that time. In other words, $x^m_{k,r}$ may depend on oracle responses

$$\left(f_U(x^m_{k',r}; z^m_{k',r}), g_U(x^m_{k',r}; z^m_{k',r})\right) \tag{28}$$

for $k' < k$—i.e., oracle queries made on that machine earlier in the current round of communication, or on oracle responses

$$\left(f_U(x^{m'}_{k',r'}, z^{m'}_{k',r'}), g_U(x^{m'}_{k',r'}, z^{m'}_{k',r'})\right) \tag{29}$$

for $r' < r$—i.e., oracle queries made on any machine in earlier rounds of communication. Letting ξ denote the random coins of the algorithm, there are thus query functions $\mathcal{Q}^m_{k,r}$ such that

$$\begin{aligned} x^m_{k,r} = \mathcal{Q}^m_{k,r}\Bigg(&\left\{x^m_{k',r}, f_U(x^m_{k',r}, z^m_{k',r}), g_U(x^m_{k',r}, z^m_{k',r}) : k' < k\right\} \\ &\cup \left\{x^{m'}_{k',r'}, f_U(x^{m'}_{k',r'}, z^{m'}_{k',r'}), g_U(x^{m'}_{k',r'}, z^{m'}_{k',r'}) : r' < r\right\}, \xi\Bigg) \end{aligned} \tag{30}$$

The key question is: upon which columns of U does the righthand side of this equation depend when we condition on $\mathcal{S}$ and $\bar{\mathcal{G}}_{k,r}^m$? To answer this, we note that by Definition 5, if $z \in \mathcal{Z}_0$ then for any x with $\mathrm{prog}_\alpha(U^\top x) \leq i$, we have

$$f_U(x; z) = f(U^\top x; z) = f_i(U_{\leq i}^\top x; z) \tag{31}$$

$$g_U(x; z) = g(U^\top x; z) = U g_i(U_{\leq i}^\top x; z) \tag{32}$$

and furthermore, $\mathrm{prog}_0(g(U^\top x; z))$ so $U g_i(U_{\leq i}^\top x; z) = U_{\leq i} g_i(U_{\leq i}^\top x; z)$. Therefore, for $z \in \mathcal{Z}_0$ and $\mathrm{prog}_\alpha(U^\top x) \leq i$, the oracle response $(f_U(x; z), g_U(x; z))$ depends only on $U_{\leq i}$. By essentially the same argument, for $z \in \mathcal{Z}_1$ and $\mathrm{prog}_\alpha(U^\top x) \leq i$, the oracle response $(f_U(x; z), g_U(x; z))$ depends only on $U_{\leq i+1}$.

Therefore, conditioned on the events Z and $\bar{\mathcal{G}}_{k,r}^m$, for each m', k', r' such that $m' = m$, $r' = r$, and $k' < k$ or $r' < r$, let $i_{k',r'}^{m'} = \min\Big\{d,\ S_{k',r'}^{m'} + \mathbb{1}\{z_{k',r'}^{m'} \in \mathcal{Z}_1\}\Big\} \leq S_{k,r}^m$ then

$$f_U(x_{k',r'}^{m'}; z_{k',r'}^{m'}) = f_{U_{\leq i_{k',r'}^{m'}}}(x_{k',r'}^{m'}; z_{k',r'}^{m'}) = f_{U_{\leq S_{k,r}^m}}(x_{k',r'}^{m'}; z_{k',r'}^{m'}) \tag{33}$$

$$g_U(x_{k',r'}^{m'}; z_{k',r'}^{m'}) = g_{U_{\leq i_{k',r'}^{m'}}}(x_{k',r'}^{m'}; z_{k',r'}^{m'}) = g_{U_{\leq S_{k,r}^m}}(x_{k',r'}^{m'}; z_{k',r'}^{m'}) \tag{34}$$

We conclude that conditioned on Z and $\bar{\mathcal{G}}_{k,r}^m$

$$\begin{aligned} x_{k,r}^m = \mathcal{Q}_{k,r}^m\Bigg(&\Big\{x_{k',r}^m, f_{U_{\leq S_{k,r}^m}}(x_{k',r}^m, z_{k',r}^m), g_{U_{\leq S_{k,r}^m}}(x_{k',r}^m, z_{k',r}^m) : k' < k\Big\} \\ &\cup \Big\{x_{k',r'}^{m'}, f_{U_{\leq S_{k,r}^m}}(x_{k',r'}^{m'}, z_{k',r'}^{m'}), g_{U_{\leq S_{k,r}^m}}(x_{k',r'}^{m'}, z_{k',r'}^{m'}) : r' < r\Big\}, \xi\Bigg) \end{aligned} \tag{35}$$

so conditioned on $\mathcal{S}$, Z, and $\bar{\mathcal{G}}_{k,r}^m$, $x_{k,r}^m$ is a measurable function of $U_{\leq S_{k,r}^m}$ and ξ.

We can apply the same argument to the algorithm's output

$$\hat{x} = \hat{X}\Bigg(\Big\{x_{k,r}^m, g_{U_{\leq \max_m S_{K,R}^m}}(x_{k,r}^m, z_{k,r}^m)\Big\}_{m,k,r}, \xi\Bigg) \tag{36}$$

which is measurable with respect to $U_{\leq \max_m S_{K,R}^m}$ and ξ conditioned on $\mathcal{S}$, Z, and $\cap_m \bar{\mathcal{G}}_{K,R}^m$. ■

Next, we show a constant-probability upper bound on the random variables $S_{k,r}^m$:

Lemma 8 *For any (α, p, δ)-robust-zero-chain,*

$$\mathbb{P}\Bigg(\max_{m,k,r} S_{k,r}^m \geq \min\{KR,\ 8KRp + 12R\log M + 12R\}\ \Bigg|\ Z\Bigg) \leq \frac{1}{4}$$

Proof The claim is implied by

$$\mathbb{P}\Bigg(\sum_{r=1}^R \max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z_{k,r}^m \in \mathcal{Z}_1\} \geq \min\{KR,\ 8KRp + 12R\log M + 12R\}\ \Bigg|\ Z\Bigg) \leq \frac{1}{4} \tag{37}$$

The seed, z, for the oracle queries are independent, so, conditioned on Z the indicators $\mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\}$ are independent Bernoulli random variables with success probability at most p. It follows that for each m and r, $\sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\}$ are independent Binomial(K, p) random variables.

Therefore, for each r, by the union bound and then the Chernoff bound, for any $c \geq 0$

$$\mathbb{P}\left(\max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \geq (1+c)Kp \,\middle|\, Z\right) \leq M \exp\left(-\frac{c^2 Kp}{2+c}\right) \tag{38}$$

Furthermore, for any random variable $X \in [0, K]$, $\mathbb{E}X = \int_0^K \mathbb{P}(X \geq x)dx$. Therefore, for any $\epsilon > 0$

$$\mathbb{E}\left[\max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \,\middle|\, Z\right] = \int_0^K \mathbb{P}\left(\max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \geq x \,\middle|\, Z\right)dx \tag{39}$$

$$= Kp \int_{-1}^{\frac{1-p}{p}} \mathbb{P}\left(\max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \geq (1+c)Kp \,\middle|\, Z\right)dc \tag{40}$$

$$\leq (1+\epsilon)Kp + MKp \int_{\epsilon}^{\frac{1-p}{p}} \exp\left(-\frac{c^2 Kp}{2+c}\right)dc \tag{41}$$

$$\leq (1+\epsilon)Kp + MKp \int_{\epsilon}^{\infty} \exp\left(-\frac{c\epsilon Kp}{2+\epsilon}\right)dc \tag{42}$$

$$= (1+\epsilon)Kp + \frac{M(2+\epsilon)}{\epsilon} \exp\left(-\frac{\epsilon^2 Kp}{2+\epsilon}\right) \tag{43}$$

For the second line we used the change of variables $x \to (1+c)Kp$. We take $\epsilon = 1 + \frac{3}{Kp}\log M$ to conclude

$$\mathbb{E}\left[\max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \,\middle|\, Z\right] \leq (1+\epsilon)Kp + \frac{M(2+\epsilon)}{\epsilon} \exp\left(-\frac{\epsilon^2 Kp}{2+\epsilon}\right) \leq 2Kp + 3\log M + 3 \tag{44}$$

It follows that

$$\mathbb{E}\left[\sum_{r=1}^R \max_{1\leq m\leq M} \sum_{k=1}^K \mathbb{1}\{z^m_{k,r} \in \mathcal{Z}_1\} \,\middle|\, Z\right] \leq 2KRp + 3R\log M + 3R \tag{45}$$

We conclude using Markov's inequality plus the observation that $S^m_{k,r} \leq KR$ for all m, k, r. ■

Using the previous lemmas, we prove the main result:

Lemma 9 *Let (f, g) be an (α, p, δ)-robust-zero-chain, let $U \in \mathbb{R}^{D\times d}$ be a uniformly random orthogonal matrix with $U^\top U = I_{d\times d}$ for $D \geq d + \frac{2\gamma^2}{\alpha^2}\log(32MKRd)$, and let $x^m_{k,r}$ be the k^{th} oracle query on the m^{th} machine during the r^{th} round of communication for an intermittent communication algorithm that interacts with the oracle (f_U, g_U) for $f_U(x; z) = f(U^\top x; z)$ and $g_U(x; z) := Ug(U^\top x; z)$. Then if $\max_{m,k,r}\|x^m_{k,r}\| \leq \gamma$, then the algorithm's output $\hat{x}$ will have progress at most*

$$\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top \hat{x}) \leq \min\{KR,\ 8KRp + 12R\log M + 12R\}\Big) \geq \frac{5}{8} - 2MKR\delta$$

Proof We begin by conditioning on Z and $\mathcal{S}$, the σ-algebra generated by $\{S^m_{k,r}\}_{m,k,r}$ and bounding

$$\begin{aligned}
&\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S^m_{K,R} \vee \exists_{m,k,r}\ \mathrm{prog}_\alpha(U^\top x^m_{k,r}) > S^m_{k,r} \,\Big|\, Z, \mathcal{S}\Big) \\
&\quad= \mathbb{P}\left(\left\{\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S^m_{K,R}\right\} \cup \bigcup_{m,k,r}\left\{\mathrm{prog}_\alpha(U^\top x^m_{k,r}) > S^m_{k,r}\right\} \,\middle|\, Z, \mathcal{S}\right) \qquad (46) \\
&\quad= \mathbb{P}\left(\left\{\left\{\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S^m_{K,R}\right\} \cap \bigcap_{m=1}^{M} \bar{\mathcal{G}}^m_{K,R}\right\} \cup \bigcup_{m,k,r}\left\{\mathrm{prog}_\alpha(U^\top x^m_{k,r}) > S^m_{k,r}\right\} \cap \bar{\mathcal{G}}^m_{k,r} \,\middle|\, Z, \mathcal{S}\right) \qquad (47) \\
&\quad\leq \mathbb{P}\left[\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S^m_{K,R},\ \bigcap_{m=1}^{M} \bar{\mathcal{G}}^m_{K,R} \,\middle|\, Z, \mathcal{S}\right] + \sum_{m,k,r} \mathbb{P}\Big(\left\{\mathrm{prog}_\alpha(U^\top x^m_{k,r}) > S^m_{k,r}\right\} \cap \bar{\mathcal{G}}^m_{k,r} \,\Big|\, Z, \mathcal{S}\Big) \qquad (48) \\
&\quad\leq \sum_{i > \max_m S^m_{K,R}} \mathbb{P}\left(|\langle U_i,\, \hat{x}\rangle| > \alpha,\ \bigcap_{m=1}^{M} \bar{\mathcal{G}}^m_{K,R} \,\middle|\, Z, \mathcal{S}\right) + \sum_{m,k,r}\sum_{i > S^m_{k,r}} \mathbb{P}\big(\big|\langle U_i,\, x^m_{k,r}\rangle\big| > \alpha, \bar{\mathcal{G}}^m_{k,r} \,\big|\, Z, \mathcal{S}\big) \qquad (49)
\end{aligned}$$

By Lemma 7, there exist measurable functions $\mathsf{A}^m_{k,r}$ and $\mathsf{B}^m_{k,r}$ such that

$$x^m_{k,r} = \mathsf{A}^m_{k,r}(U_{\leq S^m_{k,r}}, \xi)\mathbb{1}\{Z, \bar{\mathcal{G}}^m_{k,r}\} + \mathsf{B}^m_{k,r}(U, \xi)\mathbb{1}\{\neg Z \vee \neg \bar{\mathcal{G}}^m_{k,r}\} \qquad (50)$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S^m_{K,R} \vee \exists_{m,k,r}\ \mathrm{prog}_\alpha(U^\top x^m_{k,r}) > S^m_{k,r} \,\Big|\, Z, \mathcal{S}\Big) \\
&\quad\leq \sum_{i > \max_m S^m_{K,R}} \mathbb{P}\left(\left|\left\langle U_i,\, \hat{\mathsf{A}}(U_{\leq \max_m S^m_{K,R}}, \xi)\right\rangle\right| > \alpha,\ \bigcap_{m=1}^{M} \bar{\mathcal{G}}^m_{K,R} \,\middle|\, Z, \mathcal{S}\right) \\
&\qquad + \sum_{m,k,r}\sum_{i > S^m_{k,r}} \mathbb{P}\Big(\Big|\Big\langle U_i,\, \mathsf{A}^m_{k,r}(U_{\leq S^m_{k,r}}, \xi)\Big\rangle\Big| > \alpha, \bar{\mathcal{G}}^m_{k,r} \,\Big|\, Z, \mathcal{S}\Big) \qquad (51) \\
&\quad\leq \sum_{i > \max_m S^m_{K,R}} \mathbb{P}\Big(\Big|\Big\langle U_i,\, \hat{\mathsf{A}}(U_{\leq \max_m S^m_{K,R}}, \xi)\Big\rangle\Big| > \alpha \,\Big|\, Z, \mathcal{S}\Big) \\
&\qquad + \sum_{m,k,r}\sum_{i > S^m_{k,r}} \mathbb{P}\Big(\Big|\Big\langle U_i,\, \mathsf{A}^m_{k,r}(U_{\leq S^m_{k,r}}, \xi)\Big\rangle\Big| > \alpha \,\Big|\, Z, \mathcal{S}\Big) \qquad (52)
\end{aligned}$$

The algorithm's random coins, ξ, and the stochastic gradient oracles' random coins, $\{z^m_{k,r}\}_{m,k,r}$ which determine Z and $\mathcal{S}$, are independent of the random rotation U. Furthermore, for $i > S^m_{k,r}$, U_i conditioned on $U_{\leq S^m_{k,r}}$ is a uniformly random vector on the $(D - S^m_{k,r})$-dimensional unit sphere orthogonal to the range of $U_{\leq S^m_{k,r}}$. Furthermore, by assumption $\|\mathsf{A}^m_{k,r}(U_{\leq S^m_{k,r}}, \xi)\| \leq \gamma$. Therefore, following Carmon (2020) concentration of measure on the sphere implies (Ball et al., 1997)

$$\mathbb{P}\Big(\Big|\Big\langle U_i,\, \mathsf{A}^m_{k,r}(U_{\leq S^m_{k,r}}, \xi)\Big\rangle\Big| > \alpha \,\Big|\, Z, \mathcal{S}\Big) \leq 2\exp\left(-\frac{(D - S^m_{k,r} + 1)\alpha^2}{2\gamma^2}\right) \qquad (53)$$

Using the fact that $S_{k,r}^m \leq d$ and $D \geq d + \frac{2\gamma^2}{\alpha^2}\log(32MKRd)$, we conclude that

$$\begin{aligned}
&\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_m S_{K,R}^m \vee \exists_{m,k,r}\ \mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m \,\Big|\, Z, \mathcal{S}\Big) \\
&\quad\leq 2(MKR+1)d\exp\left(-\frac{(D-d+1)\alpha^2}{2\gamma^2}\right) && (54) \\
&\quad\leq 2(MKR+1)d\exp\left(-\frac{(d+\frac{2\gamma^2}{\alpha^2}\log(32MKRd)-d+1)\alpha^2}{2\gamma^2}\right) \leq \frac{1}{8} && (55)
\end{aligned}$$

To complete the proof of the lemma, we note that for $T = \min\{KR,\ 8KRp + 12R\log M + 12R\}$ we can upper bound

$$\begin{aligned}
&\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top \hat{x}) > T\Big) \\
&\leq \mathbb{P}\left(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_{m,k,r} S_{k,r}^m \vee \exists_{m,k,r}\,\mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m,\ \max_{m,k,r} S_{k,r}^m \leq T\right) \\
&\qquad + \mathbb{P}\left(\max_{m,k,r} S_{k,r}^m > T\right) && (56) \\
&\leq \mathbb{P}\left(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_{m,k,r} S_{k,r}^m \vee \exists_{m,k,r}\,\mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m\right) + \mathbb{P}\left(\max_{m,k,r} S_{k,r}^m > T\right) && (57) \\
&= \mathbb{P}\left(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_{m,k,r} S_{k,r}^m \vee \exists_{m,k,r}\,\mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m \,\middle|\, Z\right)\mathbb{P}(Z) \\
&\qquad + \mathbb{P}\left(\max_{m,k,r} S_{k,r}^m > T \,\middle|\, Z\right)\mathbb{P}(Z) \\
&+ \mathbb{P}\left(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_{m,k,r} S_{k,r}^m \vee \exists_{m,k,r}\,\mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m \,\middle|\, \neg Z\right)\mathbb{P}(\neg Z) \\
&\qquad + \mathbb{P}\left(\max_{m,k,r} S_{k,r}^m > T \,\middle|\, \neg Z\right)\mathbb{P}(\neg Z) && (58) \\
&\leq \mathbb{P}\left(\mathrm{prog}_\alpha(U^\top \hat{x}) > \max_{m,k,r} S_{k,r}^m \vee \exists_{m,k,r}\,\mathrm{prog}_\alpha(U^\top x_{k,r}^m) > S_{k,r}^m \,\middle|\, Z\right) \\
&\qquad + \mathbb{P}\left(\max_{m,k,r} S_{k,r}^m > T \,\middle|\, Z\right) + 2(1-\mathbb{P}(Z)) && (59)
\end{aligned}$$

By (55), the first term is bounded by $\frac{1}{8}$, by Lemma 8 the second term is at most $\frac{1}{4}$, and by the union bound,

$$\mathbb{P}(Z) \geq (1-\delta)^{MKR} \geq 1 - MKR\delta \tag{60}$$

This completes the proof. ■

Appendix B. An Extension to Large-Norm Queries

In the previous section, we introduce a tool for proving lower bounds for algorithm's whose oracle queries have norm bounded by γ. Here, we show how to modify the hard instances so that the lower bound applies for any algorithm, even with unboundedly large oracle queries.

Lemma 10 *The scalar function*

$$\tilde{\Gamma}(t) = \begin{cases} 0 & t \leq 0 \\ \frac{\int_0^t \exp\left(-\frac{1}{s(1-s)}\right) ds}{\int_0^t \exp\left(-\frac{1}{s(1-s)}\right) ds} & t \in (0,1) \\ 1 & t \geq 1 \end{cases}$$

is twice differentiable, and for all t*:* $|\tilde{\Gamma}'(t)| \leq 4$ *and* $|\tilde{\Gamma}''(t)| \leq 60$*.*

Proof Let $C = \int_0^t \exp\left(-\frac{1}{s(1-s)}\right) ds$. It is straightforward to confirm numerically that $C \geq \frac{1}{200}$. First, we compute the derivatives of $\tilde{\Gamma}(t)$ for $t \in (0,1)$:

$$\tilde{\Gamma}'(t) = \frac{1}{C} \exp\left(-\frac{1}{t(1-t)}\right) \tag{61}$$

$$\tilde{\Gamma}''(t) = \frac{1}{C} \exp\left(-\frac{1}{t(1-t)}\right) \frac{1-2t}{t^2(1-t)^2} \tag{62}$$

Because

$$\lim_{t \nearrow 1} \tilde{\Gamma}'(t) = \lim_{t \searrow 0} \tilde{\Gamma}'(t) = \lim_{t \nearrow 1} \tilde{\Gamma}''(t) = \lim_{t \searrow 0} \tilde{\Gamma}''(t) = 0 \tag{63}$$

we conclude that $\tilde{\Gamma}$ is twice differentiable on $\mathbb{R}$. Furthermore,

$$\sup_t |\tilde{\Gamma}'(t)| = \frac{1}{C} \sup_{s \leq \frac{1}{4}} \exp\left(-\frac{1}{s}\right) = \frac{1}{C} e^{-4} \leq 200 e^{-4} \leq 4 \tag{64}$$

and

$$\sup_t |\tilde{\Gamma}''(t)| = \frac{1}{C} \sup_{t \in (0,1)} \exp\left(-\frac{1}{t(1-t)}\right) \left| \frac{1-2t}{t^2(1-t)^2} \right| \tag{65}$$

$$\leq \frac{1}{C} \sup_{t \in (0,1)} \exp\left(-\frac{1}{t(1-t)}\right) \frac{1}{t^2(1-t)^2} \tag{66}$$

$$= \frac{1}{C} \sup_{s \geq 4} s^2 \exp(-s) \tag{67}$$

Finally, $\frac{d}{ds} s^2 \exp(-s) = (2-s)s\exp(-s)$, which is negative for all $s \geq 4$, so we conclude

$$\sup_t |\tilde{\Gamma}''(t)| \leq \frac{1}{C} 4^2 e^{-4} \leq 200 \cdot 16 e^{-4} \leq 60 \tag{68}$$

This completes the proof. ■

Lemma 11 *For* $\tilde{\Gamma}$ *as defined in Lemma 10 and any* $a, b > 0$*, we define*

$$\Gamma(x) = \tilde{\Gamma}(a(\|x\| - b))$$

This function is twice differentiable, $4a$*-Lipschitz, and* $60a^2$*-smooth.*

Proof We recall from Lemma 10 that $|\tilde{\Gamma}'| \leq 4$ and $|\tilde{\Gamma}''| \leq 60$. We now compute the gradient and Hessian of Γ:

$$\nabla\Gamma(x) = a\tilde{\Gamma}'(a(\|x\| - b))\frac{x}{\|x\|} \tag{69}$$

$$\nabla^2\Gamma(x) = a^2\tilde{\Gamma}''(a(\|x\| - b))\frac{xx^\top}{\|x\|^2} + \frac{a\tilde{\Gamma}'(a(\|x\| - b))}{\|x\|}\left(I - \frac{xx^\top}{\|x\|^2}\right) \tag{70}$$

We note that since $b > 0$, $x = 0 \implies \tilde{\Gamma}'(a(\|x\| - b)) = 0$, so the gradient and Hessian are well-defined (and equal to zero) at the point $x = 0$.

It is easy to see that for any x

$$\|\nabla\Gamma(x)\| = \left\|a\tilde{\Gamma}'(a(\|x\| - b))\frac{x}{\|x\|}\right\| = a\left|\tilde{\Gamma}'(a(\|x\| - b))\right| \leq 4a \tag{71}$$

Similarly, the eigenvalues of $\nabla^2\Gamma(x)$ are $a^2\tilde{\Gamma}''(a(\|x\| - b))$ (with multiplicity 1) and $\frac{a\tilde{\Gamma}'(a(\|x\|-b))}{\|x\|}$ (with multiplicity dimension $-$ 1). Therefore,

$$\left\|\nabla^2\Gamma(x)\right\| \leq \max\left\{a^2|\tilde{\Gamma}''(a(\|x\| - b))|,\ \frac{a|\tilde{\Gamma}'(a(\|x\| - b))|}{\|x\|}\right\} \tag{72}$$

Furthermore, since $\tilde{\Gamma}'(0) = 0$ and $|\tilde{\Gamma}''| \leq 60$, we have $|\tilde{\Gamma}'(a(\|x\| - b))| \leq 60\max\{a(\|x\| - b), 0\}$ so

$$\left\|\nabla^2\Gamma(x)\right\| \leq \max\left\{60a^2,\ \frac{60a^2(\|x\| - b)}{\|x\|}\right\} = 60a^2 \tag{73}$$

This completes the proof. ■

Lemma 12 *Let F be convex and H-smooth with $F^* = 0$ and $\min_{x:F(x)=F^*}\|x\| \leq B$, and let Γ be defined as in Lemma 11 for*

$$a = \min\left\{\frac{1}{408B}, \sqrt{\frac{\sigma^2}{32\rho^2}}\right\} \qquad \text{and} \qquad b = 2B$$

Then

$$\tilde{F}(x) := (1 - \Gamma(x))F(x) + 42H\max\{0,\ \|x\| - B\}^2$$

satisfies the following:

1. *$\tilde{F}$ is convex and $124H$-smooth*
2. *$\tilde{F}(x) = F(x)$ for all x with $\|x\| \leq B$*
3. *$\tilde{F}(x) = 42H\max\{0,\ \|x\| - B\}^2$ for all x with $\|x\| \geq 2B + \max\left\{408B,\ \sqrt{\frac{32\rho^2}{\sigma^2}}\right\}$.*
4. *For any x, $\tilde{F}(x) - \min_x \tilde{F}(x) \geq F(x) - F^*$.*

Furthermore, if $f(x)$ and $g(x)$ are stochastic zeroth-order and first-order oracles for F with variance

$$\mathbb{E}(f(x) - F(x))^2 \leq \rho^2$$
$$\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$$

Then,

$$\tilde{g}(x) = (1 - \Gamma(x))g(x) - f(x)\nabla\Gamma(x) + 42H\max\{0,\ \|x\| - B\}\frac{x}{\|x\|}$$

is an unbiased stochastic gradient oracle for $\tilde{F}$ with variance at most

$$\mathbb{E}\|\tilde{g}(x) - \nabla\tilde{F}(x)\|^2 \leq 3\sigma^2$$

Proof Let $c = 84H$. We first note that the second derivative of the second term in the definition of $\tilde{F}$ is

$$\begin{aligned}
&\frac{d^2}{dx^2}\frac{c}{2}\max\left\{0,\ \|x\| - \frac{b}{2}\right\}^2 \\
&\qquad = \frac{d}{dx}c\max\left\{0,\ \|x\| - \frac{b}{2}\right\}\frac{x}{\|x\|} && (74)\\
&\qquad = \begin{cases} 0 & \|x\| < \frac{b}{2} \\ c\frac{xx^\top}{\|x\|^2} + c\max\left\{0,\ 1 - \frac{b}{2\|x\|}\right\}\left(I - \frac{xx^\top}{\|x\|^2}\right) & \|x\| \geq \frac{b}{2}\end{cases} && (75)
\end{aligned}$$

The eigenvalues of this matrix are c (with multiplicity 1) and $c\max\left\{0,\ 1 - \frac{b}{2\|x\|}\right\} \in [0, c)$ (with multiplicity dimension $- 1$), and therefore this term is convex and $c = 84H$-smooth. Furthermore, for $\|x\| \geq b$,

$$c\max\left\{0,\ 1 - \frac{b}{2\|x\|}\right\} \geq \frac{c}{2} \tag{76}$$

and therefore this term is actually $\frac{c}{2} = 42H$-strongly convex on $\{x : \|x\| \geq b\}$.

From here, we define $\varphi(x) = (1 - \Gamma(x))F(x)$ to be the first term of $\tilde{F}$. We will now show that φ is convex and H-smooth on $\{x : \|x\| \leq b\}$ and is $40H$-smooth on $\{x : \|x\| \geq b\}$ which, together with the previous results, implies that $\tilde{F}$ is convex and $124H$-smooth everywhere.

The first piece is simple: for $x \in \{x : \|x\| \leq b\}$, $\Gamma(x) = 0$ so $\varphi(x) = F(x)$, which is convex and H-smooth. For the second part, we fix arbitrary x, y with $\|x\| \leq \|y\|$ and upper bound $\|\nabla\varphi(x) - \nabla\varphi(y)\|$. If $1 \leq a(\|x\| - b) \leq a(\|y\| - b)$, then $(1 - \Gamma(x)) = (1 - \Gamma(y)) = \Gamma'(x) = \Gamma'(y) = 0$, so $\nabla\varphi(x) = \nabla\varphi(y) = 0$, so φ is 0-smooth on the set $\{x : a(\|x\| - b) \geq 1\}$. Otherwise, when $a(\|x\| - b) < 1$ we have

$$\begin{aligned}
&\|\nabla\varphi(x) - \nabla\varphi(y)\| \\
&= \|(1 - \Gamma(x))\nabla F(x) - F(x)\nabla\Gamma(x) - (1 - \Gamma(y))\nabla F(y) + F(y)\nabla\Gamma(y)\| && (77)\\
&\leq |\Gamma(y) - \Gamma(x)|\|\nabla F(x)\| + |1 - \Gamma(y)|\|\nabla F(x) - \nabla F(y)\| \\
&\qquad + |F(x)|\|\nabla\Gamma(y) - \nabla\Gamma(x)\| + |F(y) - F(x)|\|\nabla\Gamma(y)\| && (78)
\end{aligned}$$

From here, we note that the fact that F is H-smooth and has a minimizer with norm at most B implies that F is $H(B + b + \frac{1}{a})$-Lipschitz on the set $\{x : a(\|x\| - b) < 1\} \subseteq \{x : \|x - x^*\| \leq$

$B + b + \frac{1}{a}\}$. Also, from Lemma 11, Γ is $4a$-Lipschitz and $60a^2$-smooth. Therefore, from (78), we can upper bound:

$$\begin{aligned} &\|\nabla\varphi(x) - \nabla\varphi(y)\| \\ &\leq H\left(4a\left(B+b+\frac{1}{a}\right) + 1 + 30a^2\left(B+b+\frac{1}{a}\right)^2 + 4a\left(B+b+\frac{1}{a}\right)\right)\|x-y\| && (79) \\ &= H\big(270a^2B^2 + 204aB + 39\big)\|x-y\| && (80) \end{aligned}$$

With our choice $a \leq \frac{1}{408B}$, this means $\|\nabla\varphi(x) - \nabla\varphi(y)\| \leq 40H\|x-y\|$, so φ is $40H$-smooth on $\{x : \|x\| \geq b\}$. This concludes the proof of points 1 and 2, and the third point follows immediately from the fact that $a(\|x\| - b) \geq 1 \implies \Gamma(x) = 1$.

For the fourth point, we begin by observing that since F has a minimizer x^* with $\|x^*\| \leq B = \frac{b}{2}$,

$$\min_x \tilde{F}(x) \leq \tilde{F}(x^*) = (1 - \Gamma(x^*))F(x^*) + \frac{c}{2}\max\left\{0,\ \|x^*\| - \frac{b}{2}\right\}^2 = F^* \qquad (81)$$

Furthermore, since $F \geq F^* = 0$ and $\frac{c}{2}\max\{0,\ \|x^*\| - \frac{b}{2}\} \geq 0$, we have $\min_x \tilde{F}(x) \geq 0 = F^*$, so $\min_x \tilde{F}(x) = F^* = 0$. Now, all that remains is to show that $\tilde{F}(x) \geq F(x)$.

Let x be a point with $\|x\| \leq b$, then $\Gamma(x) = 0$ so

$$\tilde{F}(x) = F(x) + \frac{c}{2}\max\left\{0,\ \|x\| - \frac{b}{2}\right\}^2 \geq F(x) \qquad (82)$$

Otherwise, if x has norm $\|x\| > b$, we already showed that on the set $\{y : \|y\| > b\}$, $\varphi(y) = (1-\Gamma(y))F(y)$ is $40H = \frac{10c}{21}$-smooth and the second term $\frac{c}{2}\max\{0,\ \|y\| - \frac{b}{2}\}^2$ is $\frac{c}{2}$-strongly convex. Therefore, $\tilde{F}$ is $\frac{c}{42} = 2H$-strongly convex on $\{y : \|y\| > b\}$. Let $x_b = b\frac{x}{\|x\|}$ be the projection of x onto the set $\{y : \|y\| \leq b\}$. By the H-smoothness of F, we have

$$F(x) \leq F(x_b) + \langle \nabla F(x_b),\ x - x_b\rangle + \frac{H}{2}\|x - x_b\|^2 \qquad (83)$$

On the other hand, by the $2H$-strong convexity of $\tilde{F}$ on $\{y : \|y\| > b\}$, we have

$$\begin{aligned} \tilde{F}(x) &\geq \tilde{F}(x_b) + \left\langle \nabla\tilde{F}(x_b),\ x - x_b\right\rangle + H\|x - x_b\|^2 && (84) \\ &= \tilde{F}(x_b) + \left\langle \nabla F(x_b) + c\max\left\{0,\ \|x\| - \frac{b}{2}\right\}\frac{x}{\|x\|},\ x - x_b\right\rangle + H\|x - x_b\|^2 && (85) \\ &> \tilde{F}(x_b) + \langle \nabla F(x_b),\ x - x_b\rangle + H\|x - x_b\|^2 && (86) \\ &\geq F(x) + \frac{H}{2}\|x - x_b\|^2 > F(x) && (87) \end{aligned}$$

Finally, for the point about the stochastic gradient oracle, it is easy to see that $\tilde{g}$ is unbiased because

$$\mathbb{E}\tilde{g}(x) = (1 - \Gamma(x))\mathbb{E}g(x) - \mathbb{E}f(x)\nabla\Gamma(x) + cH\max\{0,\ \|x\| - B\}\frac{x}{\|x\|} = \nabla\tilde{F}(x) \qquad (88)$$

For the variance, we also have that

$$\mathbb{E}\left\|\tilde{g}(x) - \nabla\tilde{F}(x)\right\|^2 = \mathbb{E}\|(1-\Gamma(x))(g(x) - \nabla F(x)) - (f(x) - F(x))\nabla\Gamma(x)\|^2 \tag{89}$$

$$\leq 2(1-\Gamma(x))^2\mathbb{E}\|g(x) - \nabla F(x)\| + 2\mathbb{E}(f(x) - F(x))^2\|\nabla\Gamma(x)\|^2 \tag{90}$$

$$\leq 2\sigma^2 + 32a^2\rho^2 \tag{91}$$

$$\leq 3\sigma^2 \tag{92}$$

This completes the proof. ■

Lemma 13 *Fix any $H, B, \sigma, \rho, K, R > 0$ and $M \geq 2$, and let*

$$\gamma = 2B + \max\left\{408B,\ \sqrt{\frac{32\rho^2}{\sigma^2}}\right\}$$

Suppose that there is a family of objectives on $\mathbb{R}^d$ that are convex, $\frac{H}{124}$-smooth, and have a minimizer with norm less than B, and that each objective, F, in the family is equipped with an oracle $(f(x), g(x))$ such that $\mathbb{E}(f(x), g(x)) = (F(x), \nabla F(x))$, $\mathbb{E}(f(x) - F(x))^2 \leq \frac{\rho^2}{3}$, and $\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \frac{\sigma^2}{3}$, and suppose that for any intermittent communication algorithm whose oracle queries are guaranteed to have norm less than γ, their output will have error at least $\mathbb{E}F(\hat{x}) - F^* \geq \epsilon$ for at least one function in the family.

Then, there exists another family of objectives on $\mathbb{R}^d$ that are convex, H-smooth, and have a minimizer with norm less than B, and each objective, $\tilde{F}$, in the family is equipped with a stochastic gradient oracle $\tilde{g}$ such that $\mathbb{E}\tilde{g}(x) = \nabla\tilde{F}(x)$ and $\mathbb{E}\|\tilde{g}(x) - \nabla\tilde{F}(x)\|^2 \leq \sigma^2$, and such that the output of any intermittent communication algorithm (whose oracle queries may have arbitrarily large norm) will have error at least $\mathbb{E}\tilde{F}(\hat{x}) - \tilde{F}^* \geq \epsilon$.

Proof For each F, f, g in the original family of objectives, we define $\tilde{F}, \tilde{g}$ in terms of the function Γ as in Lemma 12, and we consider the family of all of these $\tilde{F}$'s. As shown in Lemma 12, these $\tilde{F}$'s are each convex, H-smooth, and have a minimizer with norm at most B, and $\tilde{g}$ is an unbiased estimate of $\nabla\tilde{F}$ with variance at most σ^2.

Now, suppose towards contradiction that some optimization algorithm can ensure that its output satisfies $\mathbb{E}\tilde{F}(\hat{x}) - \tilde{F}^* < \epsilon$ for every objective $\tilde{F}$ in the modified family. By point 4 in Lemma 12, this means $\epsilon > \mathbb{E}\tilde{F}(x) - \tilde{F}^* \geq \mathbb{E}F(x) - F^*$, so this algorithm could optimize all of the objectives in the original family to error less than ϵ. Furthermore, although this algorithm might query $\tilde{g}$ at points with norm greater than γ, these queries could actually be simulated via queries to the oracle $(f(x), g(x))$ using points of norm *less* than γ. In particular, by point 3 of Lemma 12, for $\|x\| \geq \gamma$

$$\tilde{g}(x) = \nabla\tilde{F}(x) = 84H\max\{0, \|x\| - B\}\frac{x}{\|x\|} \tag{93}$$

and therefore no oracle queries at all are needed to simulate queries to $\tilde{g}$ with large norm. At the same time, for $\|x\| < \gamma$,

$$\tilde{g}(x) = (1-\Gamma(x))g(x) - f(x)\nabla\Gamma(x) + 84H\max\{0, \|x\| - B\}\frac{x}{\|x\|} \tag{94}$$

can be calculated using a single query to $(f(x), g(x))$ at x with norm $\|x\| < \gamma$. Therefore, the existence of such an algorithm that can optimize $\tilde{F}$ to accuracy less than ϵ would imply the existence of an algorithm whose queries to $(f(x), g(x))$ all have norm less than γ that can optimize F to accuracy less than ϵ, which is a contradiction. We therefore conclude that no such algorithm can exists. ■

Appendix C. Proof of Theorem 1

Now, we are ready to prove our lower bounds. We construct a hard instance for the lower bound using the scalar functions $\psi : \mathbb{R} \to \mathbb{R}$:

$$\psi(x) = \frac{\sqrt{H}x}{2\beta} \arctan\left(\frac{\sqrt{H}\beta x}{2}\right) - \frac{1}{2\beta^2} \log\left(1 + \frac{H\beta^2 x^2}{4}\right) \tag{95}$$

where H is the parameter of smoothness, and $\beta > 0$ is another parameter that controls the third derivative of ψ which we will set later. The hard instance is then

$$F(x) = -\psi'(\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{N-1} \psi(x_{i+1} - x_i) \tag{96}$$

where ζ and N are additional parameters that will be chosen later. Lemma 14 below summarizes the relevant properties of F

Lemma 14 *For any $H > 0$, $\beta > 0$, $B > 0$, and $N \geq 2$, we set $\zeta = \frac{B}{N^{3/2}}$. Then, F is convex, H-smooth, and $\nabla^2 F(x)$ is $\frac{H^{3/2}\beta}{3}$-Lipschitz; there exists $x^* \in \arg\min_x F(x)$ with $\|x^*\| \leq B$; and for any x with $\mathrm{prog}_\alpha(x) \leq \frac{N}{2}$ for $\alpha \leq \min\left\{\frac{N}{12\beta\sqrt{H}}, \frac{\sqrt{H}\beta B^2}{64N^2}\right\}$,*

$$F(x) - F^* \geq \begin{cases} \frac{N}{12\beta^2} & \beta^2 > \frac{4N^3}{HB^2} \\ \frac{HB^2}{64N^2} & \beta^2 \leq \frac{4N^3}{HB^2} \end{cases}$$

Proof First, we note that $0 \leq \psi''(x) = \frac{H}{4+H\beta^2x^2} \leq \frac{H}{4}$. Therefore, F is the sum of convex functions and is thus convex itself. We now compute the Hessian of F:

$$\nabla^2 F(x) = \psi''(x_N)e_N e_N^\top + \sum_{i=1}^{N-1} \psi''(x_{i+1} - x_i)(e_{i+1} - e_i)(e_{i+1} - e_i)^\top \tag{97}$$

Therefore, for any $u \in \mathbb{R}$,

$$\begin{aligned} u^\top \nabla^2 F(x) u &\leq \psi''(x_N)u_N^2 + \sum_{i=1}^{N-1} \psi''(x_{i+1} - x_i)(u_{i+1} - u_i)^2 && (98) \\ &\leq \frac{H}{4}\left[u_N^2 + \sum_{i=1}^{N-1} 2u_{i+1}^2 + 2u_i^2\right] && (99) \\ &\leq H\|u\|^2 && (100) \end{aligned}$$

We conclude that $\nabla^2 F(x) \preceq H \cdot I$ and thus F is H-smooth.

Next, we compute the tensor of 3rd derivatives of F:

$$\nabla^3 F(x) = \psi'''(x_N) e_N^{\otimes 3} + \sum_{i=1}^{N-1} \psi'''(x_{i+1} - x_i)(e_{i+1} - e_i)^{\otimes 3} \tag{101}$$

where

$$\psi'''(x) = \frac{-2H^2\beta^2 x}{(4 + H\beta^2 x^2)^2} \tag{102}$$

Therefore, for any $u, v \in \mathbb{R}$,

$$\left|\nabla^3 F(x)[u,u,v]\right| \leq \left|\psi'''(x_N) u_N^2 v_N\right| + \sum_{i=1}^{N-1} \left|\psi'''(x_{i+1} - x_i)(u_{i+1} - u_i)^2 (v_{i+1} - v_i)\right| \tag{103}$$

We can bound this in several different ways using Lemma 19:

$$|\psi'''(x)| \leq \frac{H^{3/2}\beta}{12} \tag{104}$$

$$|\psi'''(x)| \leq 2\beta \psi''(x)^{3/2} \tag{105}$$

$$|\psi'''(x)| \leq \frac{\sqrt{H}\beta}{2} \psi''(x) \tag{106}$$

Therefore,

$$\begin{aligned} \left|\nabla^3 F(x)[u,u,v]\right| &\leq \left|\psi'''(x_N) u_N^2 v_N\right| + \sum_{i=1}^{N-1} \left|\psi'''(x_{i+1} - x_i)(u_{i+1} - u_i)^2 (v_{i+1} - v_i)\right| && (107) \\ &\leq \sup_x |\psi'''(x)| \|v\|_\infty \left(|u_N|^2 + \sum_{i=1}^{N-1} (u_{i+1} - u_i)^2 \right) && (108) \\ &\leq \sup_x |\psi'''(x)| 4 \|u\|^2 \|v\| && (109) \end{aligned}$$

Above, we used the Hölder inequality $\sum_i |a_i b_i| \leq \|a\|_1 \|b\|_\infty$. We conclude that F is β-self-concordant. We conclude that $\nabla^2 F(x)$ is $4 \sup_x |\psi'''(x)|$-Lipschitz. To conclude, we upper bound

$$\sup_x |\psi'''(x)| = \sup_x \left| \frac{-2H^2\beta^2 x}{(4 + H\beta^2 x^2)^2} \right| \tag{110}$$

To do so, we maximize the simpler function $x \mapsto \frac{x}{(1+x^2)^2}$. We note that

$$\frac{d}{dx} \frac{x}{(1+x^2)^2} = \frac{1 - 3x^2}{(1+x^2)^3} \tag{111}$$

$$\frac{d^2}{dx^2} \frac{x}{(1+x^2)^2} = \frac{12x(x^2 - 1)}{(1+x^2)^4} \tag{112}$$

Therefore, the derivative is zero at $\pm 1/\sqrt{3}$ and as $x \to \pm\infty$ and the second derivative is negative only for $+1/\sqrt{3}$, and $\lim_{x\to\pm\infty} \frac{x}{(1+x^2)^2} = 0$. Therefore, we conclude that

$$\sup_x \frac{x}{(1+x^2)^2} = \sup_x \frac{|x|}{(1+x^2)^2} = \frac{\sqrt{\frac{1}{3}}}{(1+\frac{1}{3})^2} = \frac{3\sqrt{3}}{16} \tag{113}$$

By rescaling, we conclude that

$$\sup_x \left|\psi'''(x)\right| = \sup_x \frac{2H^2\beta^2|x|}{(4+H\beta^2x^2)^2} = \frac{H^{3/2}\beta}{4}\sup_x \frac{\left|\frac{\sqrt{H}\beta x}{2}\right|}{\left(1+\left(\frac{\sqrt{H}\beta x}{2}\right)^2\right)^2} = \frac{3\sqrt{3}H^{3/2}\beta}{64} < \frac{H^{3/2}\beta}{12} \tag{114}$$

Combining this with (109) completes the upper bound on the Hessian Lipschitz parameter.

We now bound the norm of the minimizer of F. The first-order optimality condition $\nabla F(x^*) = 0$ indicates

$$\begin{aligned} [\nabla F(x^*)]_1 &= 0 = \psi'(-\zeta) - \psi'(x_2^* - x_1^*) \\ [\nabla F(x^*)]_i &= 0 = \psi'(x_i^* - x_{i-1}^*) - \psi'(x_{i+1}^* - x_i^*) \qquad 1 < i < N \\ [\nabla F(x^*)]_N &= 0 = \psi'(x_N^* - x_{N-1}^*) + \psi'(x_N^*) \end{aligned} \tag{115}$$

Because $\psi'(x) = \arctan(x)$ is invertible on its range, we conclude that $x_i^* - x_{i+1}^* = \zeta$ for $i < N$, and $x_N^* = \zeta$. So, the following point minimizes F:

$$x^* = \zeta \sum_{i=1}^{N} (N-i+1)e_i \tag{116}$$

This point has norm

$$\|x^*\|^2 = \zeta^2 \sum_{i=1}^{N} (N-i+1)^2 = \frac{\zeta^2}{6}\left(2N^3 + 3N^2 + N\right) \leq \zeta^2 N^3 \tag{117}$$

Therefore, setting $\zeta^2 = \frac{B^2}{N^3}$ ensures the existence of a minimizer with norm less than B.

At this point,

$$F(x^*) = -N\zeta\psi'(\zeta) + \psi(\zeta) + \sum_{i=1}^{N} \psi(-\zeta) = N(\psi(\zeta) - \zeta\psi'(\zeta)) \tag{118}$$

Finally, by Jensen's inequality and the convexity of ψ, for any I

$$\sum_{i=1}^{I} \psi(x_{i+1} - x_i) = I \cdot \frac{1}{I}\sum_{i=1}^{I} \psi(x_{i+1} - x_i) \geq I\psi\left(\frac{x_{I+1} - x_1}{I}\right) \tag{119}$$

Therefore, for any x with $\operatorname{prog}_\alpha(x) = I \leq \frac{N}{2}$,

$$\begin{aligned} F(x) &= -\psi'(\zeta)x_1 + \psi(x_N) + \sum_{n=1}^{N-1} \psi(x_{i+1} - x_i) && (120) \\ &\geq -\psi'(\zeta)x_1 + I\psi\left(\frac{x_{I+1} - x_1}{I}\right) && (121) \\ &\geq -\psi'(\zeta)x_1 + I\psi\left(\frac{x_1 - \alpha}{I}\right) && (122) \\ &\geq \inf_y -\psi'(\zeta)y + I\psi\left(\frac{y - \alpha}{I}\right) && (123) \end{aligned}$$

The minimizing y above satisfies

$$0 = -\psi'(\zeta) + \psi'\left(\frac{y-\alpha}{I}\right) \implies y = \alpha + I\zeta \tag{124}$$

so

$$F(x) \geq -\psi'(\zeta)(\alpha + I\zeta) + I\psi(\zeta) \tag{125}$$

Finally, we conclude that

$$\begin{aligned} F(x) - F^* &\geq I\big(\psi(\zeta) - \zeta\psi'(\zeta)\big) - \alpha\psi'(\zeta) - N\big(\psi(\zeta) - \zeta\psi'(\zeta)\big) && (126) \\ &= (N - I)\big(\zeta\psi'(\zeta) - \psi(\zeta)\big) - \alpha\psi'(\zeta) && (127) \\ &\geq \frac{N}{4\beta^2}\log\left(1 + \frac{H\beta^2\zeta^2}{4}\right) - \frac{\alpha\sqrt{H}}{\beta} && (128) \end{aligned}$$

Where we used that

$$\begin{aligned} \psi(\zeta) &= \frac{\sqrt{H}\zeta}{2\beta}\arctan\left(\frac{\sqrt{H}\beta\zeta}{2}\right) - \frac{1}{2\beta^2}\log\left(1 + \frac{H\beta^2\zeta^2}{4}\right) && (129) \\ \psi'(\zeta) &= \frac{\sqrt{H}}{2\beta}\arctan\left(\frac{\sqrt{H}\beta\zeta}{2}\right) \leq \frac{\pi\sqrt{H}}{4\beta} < \frac{\sqrt{H}}{\beta} && (130) \end{aligned}$$

From here, we consider two cases, if $\beta^2 > \frac{4}{H\zeta^2}$, then

$$F(x) - F^* > \frac{N}{4\beta^2}\log(2) - \frac{\alpha\sqrt{H}}{\beta} > \frac{N}{6\beta^2} - \frac{\alpha\sqrt{H}}{\beta} \geq \frac{N}{12\beta^2} \tag{131}$$

Otherwise, if $\beta^2 \leq \frac{4}{H\zeta^2}$ then we use that for $x \leq 1$, $\log(1+x) \geq \frac{x}{2}$ and conclude

$$F(x) - F^* > \frac{N}{4\beta^2} \cdot \frac{H\beta^2\zeta^2}{8} - \frac{\alpha\sqrt{H}}{\beta} = \frac{HB^2}{32N^2} - \frac{\alpha\sqrt{H}}{\beta} \geq \frac{HB^2}{64N^2} \tag{132}$$

This completes the proof. ■

Now, we define a stochastic zeroth- and first-order oracle for F. First, we specify the distribution over z:

$$z = \begin{cases} 0 & \text{with probability } (1-p)(1-\delta) \\ 1 & \text{with probability } p(1-\delta) \\ 2 & \text{with probability } \delta \end{cases} \tag{133}$$

Then, f and g are defined as

$$\begin{aligned} f(x;0) &= F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)}, 0, \dots, 0\big]) \\ f(x;1) &= F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)+1}, 0, \dots, 0\big]) \\ f(x;2) &= \frac{1}{\delta}F(x) - \frac{1-\delta}{\delta}((1-p)f(x;0) + pf(x;1)) \end{aligned} \tag{134}$$

and

$$\begin{aligned} g(x;0) &= \sum_{i=1}^{\mathrm{prog}_\alpha(x)} e_i e_i^\top \nabla F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)}, 0, \dots, 0\big]) \\ g(x;1) &= \frac{1}{p}\sum_{i=1}^{\mathrm{prog}_\alpha(x)+1} e_i e_i^\top \nabla F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)+1}, 0, \dots, 0\big]) - \frac{1-p}{p}g(x;0) \\ g(x;2) &= \frac{1}{\delta}\nabla F(x) - \frac{1-\delta}{\delta}\sum_{i=1}^{\mathrm{prog}_\alpha(x)+1} e_i e_i^\top \nabla F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)+1}, 0, \dots, 0\big]) \end{aligned} \tag{135}$$

The following lemma relates the properties of ψ to those of F and g:

Lemma 15 *For F defined as in (96) and the oracle (f,g)*

1. $\mathbb{E}_z f(x;z) = F(x)$ *and* $\mathbb{E}_z g(x;z) = \nabla F(x)$
2. $\sup_x \mathbb{E}_z (f(x;z) - F(x))^2 \leq \frac{12H\alpha^2}{\beta^2\delta} + \frac{3N^2H^2\alpha^4}{\delta}$
3. $\sup_x \mathbb{E}_z \|g(x;z) - \nabla F(x)\|^2 \leq \frac{6H(1-p)}{\beta^2 p} + 6NH^2\alpha^2\left(\frac{1}{p} + \frac{1}{\delta}\right)$
4. (f,g) *is an* (α,p,δ)*-robust-zero-chain.*

Proof We will prove each property one by one.
1) This is a simple calculation. For f, we have

$$\begin{aligned} \mathbb{E}_z f(x;z) &= (1-p)(1-\delta)f(x;0) + p(1-\delta)f(x;1) + \delta f(x;2) && (136) \\ &= (1-p)(1-\delta)f(x;0) + p(1-\delta)f(x;1) \\ &\quad + \delta\left(\frac{1}{\delta}F(x) - \frac{1-\delta}{\delta}((1-p)f(x;0) + pf(x;1))\right) && (137) \\ &= F(x) && (138) \end{aligned}$$

For the gradient, let

$$\nabla_0 = \sum_{i=1}^{\mathrm{prog}_\alpha(x)} e_i e_i^\top \nabla F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)}, 0, \dots, 0\big]) \tag{139}$$

$$\nabla_1 = \sum_{i=1}^{\mathrm{prog}_\alpha(x)+1} e_i e_i^\top \nabla F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)+1}, 0, \dots, 0\big]) \tag{140}$$

Then

$$\begin{aligned}
&\mathbb{E}_z g(x; z) \\
&\quad = (1-p)(1-\delta) g(x; 0) + p(1-\delta) g(x; 1) + \delta g(x; 2) && (141) \\
&\quad = (1-p)(1-\delta)\nabla_0 + p(1-\delta)\left(\frac{1}{p}\nabla_1 - \frac{1-p}{p}\nabla_0\right) + \delta\left(\frac{1}{\delta}\nabla F(x) - \frac{1-\delta}{\delta}\nabla_1\right) && (142) \\
&\quad = \nabla F(x) && (143)
\end{aligned}$$

2) For any x

$$\begin{aligned}
&\mathbb{E}_z (f(x; z) - F(x))^2 \\
&= (1-p)(1-\delta)(f(x;0) - F(x))^2 + p(1-\delta)(f(x;1) - F(x))^2 + \delta(f(x;2) - F(x))^2 && (144) \\
&= (1-p)(1-\delta)(f(x;0) - F(x))^2 + p(1-\delta)(f(x;1) - F(x))^2 \\
&\qquad + \delta\left(\frac{1}{\delta}F(x) - \frac{1-\delta}{\delta}((1-p)f(x;0) + pf(x;1)) - F(x)\right)^2 && (145) \\
&= (1-p)(1-\delta)(f(x;0) - F(x))^2 + p(1-\delta)(f(x;1) - F(x))^2 \\
&\qquad + \frac{(1-\delta)^2}{\delta}((1-p)f(x;0) + pf(x;1) - F(x))^2 && (146) \\
&\le \left((1-p)(1-\delta) + \frac{2(1-\delta)^2(1-p)^2}{\delta}\right)(f(x;0) - F(x))^2 \\
&\qquad + \left(p(1-\delta) + \frac{2(1-\delta)^2 p^2}{\delta}\right)(f(x;1) - F(x))^2 && (147) \\
&\le \frac{3}{\delta}\big((f(x;0) - F(x))^2 + (f(x;1) - F(x))^2\big) && (148)
\end{aligned}$$

We will address each term separately.

$$\begin{aligned}
(f(x;0) - F(x))^2 &= (F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)}, 0, \dots, 0\big]) - F(x))^2 && (149)\\
&= \left(\psi(-x_{\mathrm{prog}_\alpha(x)}) - \sum_{i=\mathrm{prog}_\alpha(x)}^{N-1} \psi(x_{i+1} - x_i) - \psi(x_N)\right)^2 && (150)\\
&\leq 2\big(\psi(-x_{\mathrm{prog}_\alpha(x)}) - \psi(x_{\mathrm{prog}_\alpha(x)+1} - x_{\mathrm{prog}_\alpha(x)})\big)^2 \\
&\qquad + 2\left(\sum_{i=\mathrm{prog}_\alpha(x)+1}^{N-1} \psi(x_{i+1} - x_i) + \psi(x_N)\right)^2 && (151)\\
&\leq 2\left(\frac{\sqrt{H}}{\beta}\alpha\right)^2 + 2\left(\sum_{i=\mathrm{prog}_\alpha(x)+1}^{N} \psi(2\alpha)\right)^2 && (152)\\
&\leq \frac{2H\alpha^2}{\beta^2} + \frac{N^2H^2\alpha^4}{2} && (153)
\end{aligned}$$

Above, we used that $|\psi'(x)| = \left|\frac{\sqrt{H}}{2\beta}\arctan\left(\frac{\sqrt{H}\beta x}{2}\right)\right| \leq \frac{\sqrt{H}}{\beta}$so ψ is $\frac{\sqrt{H}}{\beta}$-Lipschitz, and also $|\psi''(x)| = \frac{H}{4+H\beta^2x^2} \leq \frac{H}{4}$ so ψ is $\frac{H}{4}$-smooth. Similarly,

$$\begin{aligned}
(f(x;1) - F(x))^2 &= (F(\big[x_1, \dots, x_{\mathrm{prog}_\alpha(x)+1}, 0, \dots, 0\big]) - F(x))^2 && (154)\\
&= \left(\psi(-x_{\mathrm{prog}_\alpha(x)+1}) - \sum_{i=\mathrm{prog}_\alpha(x)+1}^{N-1} \psi(x_{i+1} - x_i) - \psi(x_N)\right)^2 && (155)\\
&\leq 2\big(\psi(-x_{\mathrm{prog}_\alpha(x)+1}) - \psi(x_{\mathrm{prog}_\alpha(x)+2} - x_{\mathrm{prog}_\alpha(x)+1})\big)^2 \\
&\qquad + 2\left(\sum_{i=\mathrm{prog}_\alpha(x)+1}^{N-1} \psi(x_{i+1} - x_i) + \psi(x_N)\right)^2 && (156)\\
&\leq 2\left(\frac{\sqrt{H}}{\beta}\alpha\right)^2 + 2\left(\sum_{i=\mathrm{prog}_\alpha(x)+1}^{N} \psi(2\alpha)\right)^2 && (157)\\
&\leq \frac{2H\alpha^2}{\beta^2} + \frac{N^2H^2\alpha^4}{2} && (158)
\end{aligned}$$

Therefore,

$$\mathbb{E}_z(f(x;z) - F(x))^2 \leq \frac{12H\alpha^2}{\beta^2\delta} + \frac{3N^2H^2\alpha^4}{\delta} \qquad (159)$$

3) Using the same ∇_0 and ∇_1 as above, we first expand

$$\begin{aligned}
&\mathbb{E}_z\|g(x;z) - \nabla F(x)\|^2 \\
&\leq (1-p)\|\nabla_0 - \nabla F(x)\|^2 + \frac{1}{p}\|\nabla_1 - (1-p)\nabla_0 - p\nabla F(x)\|^2 + \frac{(1-\delta)^2}{\delta}\|\nabla_1 - \nabla F(x)\|^2 \qquad (160) \\
&\leq \frac{3(1-p)}{p}\|\nabla_0 - \nabla F(x)\|^2 + \frac{3}{\delta}\|\nabla_1 - \nabla F(x)\|^2 \qquad (161)
\end{aligned}$$

Before proceeding, we recall that $\psi''(x) \leq \frac{H}{4}$ and $|\psi'(x)| \leq \frac{\sqrt{H}}{\beta}$. Therefore, for $j = \mathrm{prog}_\alpha(x)$,

$$\begin{aligned}
&\|\nabla_0 - \nabla F(x)\|^2 \\
&= \Bigg\| -\psi'(\zeta)e_1 + \sum_{i=1}^{j-1}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i) + \psi'(-x_j)(-e_j) \\
&\qquad + \psi'(\zeta)e_1 - \psi'(x_N)e_N - \sum_{i=1}^{N-1}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i)\Bigg\|^2 \qquad (162) \\
&= \Bigg\| -e_j\psi'(-x_j) - \psi'(x_N)e_N - \sum_{i=j}^{N-1}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i)\Bigg\|^2 \qquad (163) \\
&\leq 2\big\| -\psi'(-x_j)e_j - \psi'(x_{j+1}-x_j)(e_{j+1}-e_j)\big\|^2 \\
&\qquad + 2\Bigg\|\psi'(x_N)e_N + \sum_{i=j+1}^{N-1}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i)\Bigg\|^2 \qquad (164) \\
&\leq 2\left(\frac{H}{\beta^2} + \frac{H^2}{16}\alpha^2\right) + 32(N-j-1)\frac{H^2}{16}\alpha^2 \qquad (165) \\
&\leq \frac{2H}{\beta^2} + 2NH^2\alpha^2 \qquad (166)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|\nabla_1 - \nabla F(x)\|^2 \\
&= \Bigg\| -\psi'(\zeta)e_1 + \sum_{i=1}^{j} \psi'(x_{i+1} - x_i)(e_{i+1} - e_i) + \psi'(-x_{j+1})(-e_{j+1}) \\
&\qquad + \psi'(\zeta)e_1 - \psi'(x_N)e_N - \sum_{i=1}^{N-1} \psi'(x_{i+1} - x_i)(e_{i+1} - e_i) \Bigg\|^2 \qquad (167) \\
&= \Bigg\| -e_{j+1}\psi'(-x_{j+1}) - \psi'(x_N)e_N - \sum_{i=j+1}^{N-1} \psi'(x_{i+1} - x_i)(e_{i+1} - e_i) \Bigg\|^2 \qquad (168) \\
&\le 2\big\| -e_{j+1}\psi'(-x_{j+1}) - \psi'(x_{j+2} - x_{j+1})(e_{j+2} - e_{j+1}) \big\|^2 \\
&\qquad + 2\Bigg\| \psi'(x_N)e_N + \sum_{i=j+2}^{N-1} \psi'(x_{i+1} - x_i)(e_{i+1} - e_i) \Bigg\|^2 \qquad (169) \\
&\le 10\frac{H^2}{16}\alpha^2 + 32(N - j - 2)\frac{H^2}{16}\alpha^2 \qquad (170) \\
&\le 2NH^2\alpha^2 \qquad (171)
\end{aligned}$$

We conclude that

$$\begin{aligned}
\mathbb{E}_z\|g(x;z) - \nabla F(x)\|^2 &\le \frac{3(1-p)}{p}\left(\frac{2H}{\beta^2} + 2NH^2\alpha^2\right) + \frac{6NH^2\alpha^2}{\delta} \qquad (172) \\
&\le \frac{6H(1-p)}{\beta^2 p} + 6NH^2\alpha^2\left(\frac{1}{p} + \frac{1}{\delta}\right) \qquad (173)
\end{aligned}$$

4) Comparing (134) and (135) to Definition 5, it is easy to see that (f,g) is an (α,p,δ)-robust-zero-chain with $\mathcal{Z}_0 = \{0\}$ and $\mathcal{Z}_1 = \{1\}$. ■

Lemma 16 *Let $H, B, \sigma^2 > 0$ and let g be a stochastic gradient oracle with variance bounded by σ^2. Then for any algorithm that accesses the oracle T times, there exists a function in one dimension such that the algorithm's output will have error at least*

$$\mathbb{E}F(\hat{x}) - F^* \ge \frac{3}{8}\min\left\{\frac{\sigma B}{\sqrt{T}}, HB^2\right\}$$

Proof Consider the following pair of objectives:

$$\begin{aligned}
F_+(x) &= \frac{a}{2}x^2 + bx \\
F_-(x) &= \frac{a}{2}x^2 - bx
\end{aligned} \qquad (174)$$

with a stochatic gradient oracles

$$\begin{aligned}
z &\sim \mathcal{N}(0,\sigma^2) \\
g_+(x;z) &= \nabla F_+(x) + z \\
g_-(x;z) &= \nabla F_-(x) + z
\end{aligned} \qquad (175)$$

First, we note that for any x,

$$F_+(x) - \min_x F_+(x) \leq \frac{b^2}{2a} \implies x \leq 0 \implies F_-(x) - \min_x F_-(x) \geq \frac{b^2}{2a} \tag{176}$$

and vice versa. Therefore, any algorithm that succeeds in optimizing both F_+ and F_- to accuracy better than $\frac{b^2}{2a}$ with probability at least $\frac{3}{4}$ needs to determine which of the two functions it is optimizing with probability at least $\frac{3}{4}$. However, by the Pinsker inequality, the total variation distance between T queries to g_+ and g_- is at most

$$\begin{aligned}\|\mathbb{P}_+ - \mathbb{P}_-\|_{\mathrm{TV}} &\leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_+\|\mathbb{P}_-)} &(177)\\ &\leq \sqrt{\frac{T}{2}\mathrm{D}_{\mathrm{KL}}(\mathcal{N}(2b,\sigma^2)\|\mathcal{N}(0,\sigma^2))} &(178)\\ &= \frac{b}{\sigma}\sqrt{T} &(179)\end{aligned}$$

Therefore, if $b \leq \frac{3\sigma}{4\sqrt{T}}$, no algorithm can optimize to accuracy better $\frac{b^2}{2a}$ with probability greater than $\frac{3}{4}$. Finally, we note that F_+ and F_- are a-smooth, and have minimizers $\mp\frac{b}{a}$. Therefore, we take $b = \min\left\{aB, \frac{3\sigma}{4\sqrt{T}}\right\}$ and $a = \min\left\{H, \frac{3\sigma}{4B\sqrt{T}}\right\}$ so that the objectives are H-smooth and have solutions of norm B and with probability at least $\frac{1}{4}$

$$\max_{*\in\{+,-\}} \mathbb{E}F_*(\hat{x}) - \min_x F_*(x) \geq \min\left\{\frac{aB^2}{2}, \frac{9\sigma^2}{32aT}\right\} \geq \min\left\{\frac{HB^2}{2}, \frac{3\sigma B}{8\sqrt{T}}\right\} \tag{180}$$

This completes the proof. ■

Theorem 1 *For any $H, B, \sigma, K, R > 0$ and $M \geq 2$, and any intermittent communication algorithm, there exists a convex, H-smooth objective which has a minimizer with norm at most B in any dimension*

$$d \geq 2KR + \left(10^9\left(1 + KR + \left(\frac{HB}{\sigma}\right)^{3/2} M(KR)^{5/4}\right) + \frac{6144H^2B^2MKR}{\sigma^2}\right)\log(64MK^2R^2)$$

and a stochastic gradient oracle, g, with $\mathbb{E}_z\|g(x;z) - \nabla F(x)\|^2 \leq \sigma^2$ such that the algorithm's output will have error at least

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2(1+\log^2 M)}, \frac{\sigma B}{\sqrt{KR}}\right\}\right)$$

for a numerical constant c.

Proof We set the parameters of the objective (96) and the oracles (134) and (135) according to

$$p = \min\left\{1,\ \frac{\sqrt{HB}}{\sqrt{\sigma}K^{3/4}R^{3/4}}\right\} \tag{181}$$

$$\delta = \frac{1}{16MKR} \tag{182}$$

$$N = \min\{2KR, 16KRp + 24R(1+\log M)\} \tag{183}$$

$$\alpha = \min\left\{\frac{N}{12\beta\sqrt{H}},\ \frac{\sqrt{H}\beta B^2}{64N^2},\ \frac{\sigma\sqrt{p\delta}}{5H\sqrt{N}}\right\} \tag{184}$$

$$\beta = \frac{2N^{3/2}}{\sqrt{HB}} \tag{185}$$

By Lemma 14, the objective (96) is convex, H-smooth, and has a minimizer with norm less than B. In order to apply Lemma 6, we will also introduce a random rotation U, but since $U^\top U = I_{N\times N}$, this does not affect the convexity, smoothness, or norm of the minimizers of the objective, so we conclude that our construction satisfies the necessary conditions.

Furthermore, by Lemma 15, the stochastic zeroth- and first-order oracles defined in (134) and (135) are unbiased and, with our choice of $\alpha \le \frac{\sigma\sqrt{p\delta}}{5H\sqrt{N}}$ and (as we will show) $p \ge \frac{12H}{12H+\beta^2\sigma^2}$, they have variance bounded by

$$\sup_x \mathbb{E}_z(f(x;z) - F(x))^2 \le \frac{12H\alpha^2}{\beta^2\delta} + \frac{3N^2H^2\alpha^4}{\delta} \tag{186}$$

$$\le \frac{12\sigma^2 p}{25HN\beta^2} + \frac{3\sigma^4p^2\delta}{625H^2} =: \rho^2 \tag{187}$$

$$\sup_x \mathbb{E}_z\|g(x;z) - \nabla F(x)\|^2 \le \frac{6H(1-p)}{\beta^2 p} + 6NH^2\alpha^2\left(\frac{1}{p} + \frac{1}{\delta}\right) \le \sigma^2 \tag{188}$$

Next, on the way to applying Lemma 6, we introduce a uniformly random rotation for $U \in \mathbb{R}^{d\times N}$, and consider $F(U^\top x)$ with oracle (f_U, g_U). With our choice of d, we note that

$$d \ge N + \frac{2\gamma^2}{\alpha^2}\log(32MKRN) \tag{189}$$

for

$$\gamma = 2B + \max\left\{408B,\ \sqrt{\frac{32\rho^2}{\sigma^2}}\right\} \tag{190}$$

Therefore, by Lemma 6, for any algorithm whose queries are bounded in norm by γ, the algorithm's output $\hat{x}$ will satisfy

$$\mathbb{P}\Big(\mathrm{prog}_\alpha(U^\top\hat{x}) \le \min\{KR, 8KRp + 12R(1+\log M)\}\Big) \ge \frac{5}{8} - 2MKR\delta \ge \frac{1}{2} \tag{191}$$

Therefore, with our choice of

$$N = \min\{2KR, 16KRp + 24R(1+\log M)\} \tag{192}$$

we have that $\mathrm{prog}_\alpha(U^\top \hat{x}) \leq \frac{N}{2}$ with probability at least $\frac{1}{2}$. It therefore follows from Lemma 14 that with probability at least $\frac{1}{2}$

$$\begin{aligned} F(U^\top \hat{x}) - F^* &\geq \begin{cases} \frac{N}{12\beta^2} & \beta^2 > \frac{4N^3}{HB^2} \\ \frac{HB^2}{64N^2} & \beta^2 \leq \frac{4N^3}{HB^2} \end{cases} && (193) \\ &= \frac{HB^2}{64\min\{2KR, 16KRp + 24R(1+\log M)\}^2} && (194) \\ &\geq \frac{HB^2}{512K^2R^2} + \frac{HB^2}{32768K^2R^2p^2 + 73728R^2(1+\log M)^2} && (195) \\ &\geq \frac{1}{73728}\left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{K^2R^2p^2}, \frac{HB^2}{R^2(1+\log M)^2}\right\}\right) && (196) \end{aligned}$$

From here, what remains is to show how small p can be taken. Above, we claimed that our choice satisfies $p \geq \frac{12H}{12H+\beta^2\sigma^2}$, but this is a more complicated statement than it appears since β is defined in terms of N, which is, in turn, defined in terms of p. We have

$$\begin{aligned} \frac{12H}{12H+\beta^2\sigma^2} &= \frac{12H}{12H + \frac{4\sigma^2N^3}{HB^2}} && (197) \\ &= \frac{4H^2B^2}{4H^2B^2 + 4\sigma^2N^3} && (198) \\ &= \frac{4H^2B^2}{4H^2B^2 + 4\sigma^2\min\{2KR, 16KRp + 24R(1+\log M)\}^3} && (199) \\ &\leq \frac{4H^2B^2}{4H^2B^2 + 4\sigma^2(2KRp)^3} && (200) \end{aligned}$$

Therefore, it suffices to set

$$\begin{aligned} p &\geq \frac{4H^2B^2}{4H^2B^2 + 4\sigma^2(2KRp)^3} && (201) \\ \iff 32\sigma^2K^3R^3p^4 + 4H^2B^2p &\geq 4H^2B^2 && (202) \\ \impliedby p^4 &\geq \min\left\{1, \frac{H^2B^2}{8\sigma^2K^3R^3}\right\} && (203) \end{aligned}$$

so, our choice of p is sound. Therefore, we can lower bound

$$\begin{aligned} \mathbb{E}F(U^\top \hat{x}) - F^* &\geq \frac{1}{2}\cdot\frac{1}{73728}\left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{K^2R^2p^2}, \frac{HB^2}{R^2(1+\log M)^2}\right\}\right) && (204) \\ &\geq \frac{1}{2}\cdot\frac{1}{73728}\left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{K^2R^2\frac{HB}{\sigma K^{3/2}R^{3/2}}}, \frac{HB^2}{R^2(1+\log M)^2}\right\}\right) && (205) \\ &\geq \frac{1}{2}\cdot\frac{1}{73728}\left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{HB^2}{R^2(1+\log M)^2}\right\}\right) && (206) \end{aligned}$$

This lower bound applies to all algorithm's whose queries are bounded by γ. To extend the result to all randomized algorithms, we apply Lemma 13, which results in only a constant factor degredation in the lower bound. Finally, we apply Lemma 16 to conclude that

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{1}{4}\min\left\{HB^2,\, \frac{\sigma B}{\sqrt{MKR}}\right\} \tag{207}$$

This completes the proof. ■

Corollary 17 *There is a numerical constant, c, such that no intermittent communication algorithm can guarantee for any H-smooth, λ-strongly convex objective F and stochastic gradient oracle with variance less than σ^2 that its output will have suboptimality*

$$\begin{aligned}\mathbb{E}F(\hat{x}) - F^* \leq c\cdot\Bigg(&\frac{F(0)-F^*}{K^2R^2}\exp\left(-\sqrt{\frac{\lambda}{H}}KR\right) + \frac{\sigma^2}{\lambda MKR} \\ &+ \min\left\{\frac{F(0)-F^*}{R^2\log^2 M}\exp\left(-\sqrt{\frac{\lambda}{H}}R\log M\right),\, \frac{\sigma^2}{\lambda KR}\right\}\Bigg)\end{aligned}$$

Proof Suppose there were an algorithm which guaranteed convergence at a rate

$$\begin{aligned}G(\hat{x}) - G^* \leq c\cdot\Bigg(&\frac{G(0)-G^*}{K^2R^2}\exp\left(-\sqrt{\frac{\lambda}{H}}KR\right) + \frac{\sigma^2}{\lambda MKR} \\ &+ \min\left\{\frac{G(0)-G^*}{R^2\log^2 M}\exp\left(-\sqrt{\frac{\lambda}{H}}R\log M\right),\, \frac{\sigma^2}{\lambda KR}\right\}\Bigg)\end{aligned} \tag{208}$$

for any λ-strongly convex function G. Then, we could use this algorithm to optimize a merely convex F with $\|x^*\| \leq B$ by applying it to the λ-strongly convex $G(x) = F(x) + \frac{\lambda}{2}\|x\|^2$. Using x^*_G to denote the minimizer of G, this would ensure (for a universal constant c which may change from line to line)

$$F(\hat{x}) - F^* \leq G(\hat{x}) - F^* \tag{209}$$

$$= G(\hat{x}) - G(x^*) + \frac{\lambda}{2}\|x^*\|^2 \tag{210}$$

$$\leq G(\hat{x}) - G(x^*_G) + \frac{\lambda B^2}{2} \tag{211}$$

$$\leq c\cdot\Bigg(\frac{G(0)-G^*}{K^2R^2}\exp\left(-\sqrt{\frac{\lambda}{H}}KR\right) + \frac{\sigma^2}{\lambda MKR} \tag{212}$$

$$+ \min\left\{\frac{G(0)-G^*}{R^2\log^2 M}\exp\left(-\sqrt{\frac{\lambda}{H}}R\log M\right),\, \frac{\sigma^2}{\lambda KR}\right\}\Bigg) + \lambda B^2 \tag{213}$$

$$\leq c\cdot\Bigg(\frac{HB^2}{K^2R^2}\exp\left(-\sqrt{\frac{\lambda}{H}}KR\right) + \frac{\sigma^2}{\lambda MKR} \tag{214}$$

$$+ \min\left\{\frac{HB^2}{R^2\log^2 M}\exp\left(-\sqrt{\frac{\lambda}{H}}R\log M\right),\, \frac{\sigma^2}{\lambda KR}\right\} + \lambda B^2\Bigg) \tag{215}$$

Consequently, if we choose

$$\lambda = \max\left\{\frac{H}{K^2R^2}, \frac{\sigma}{B\sqrt{MKR}}, \frac{H}{R^2\log^2 M}, \frac{\sigma}{B\sqrt{KR}}\right\} \tag{216}$$

then this approach would guarantee

$$F(\hat{x}) - F^* \leq c\cdot\left(\frac{HB^2}{K^2R^2} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right. \tag{217}$$

$$\left. + B^2\max\left\{\frac{H}{K^2R^2}, \frac{\sigma}{B\sqrt{MKR}}, \frac{H}{R^2\log^2 M}, \frac{\sigma}{B\sqrt{KR}}\right\}\right) \tag{218}$$

$$= c\cdot\left(\frac{HB^2}{K^2R^2} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right) \tag{219}$$

In light of the lower bound in Theorem 1, we conclude that no algorithm can provide a guarantee that is more than a constant factor better than (208). ■

Appendix D. Proof of Theorem 4

In this section, we extend Theorem 1 to the case where the objective is required to exhibit higher-order smoothness. Although we are confident that similar results as Theorem 4 would apply to arbitrary randomized algorithms, we prove the lower bound here just for zero-respecting algorithms (Carmon et al., 2017):

Definition 18 (Distributed Zero-Respecting Intermittent Communication Algorithm) *We say that a parallel method is an intermittent communication algorithm if for each m, k, r, there exists a mapping $\mathcal{A}_{k,r}^m$ such that $x_{k,r}^m$, the k^{th} query on the m^{th} machine during the r^{th} round of communication, is computed as*

$$x_{k,r}^m = \mathcal{A}_{k,r}^m\left(\left[x_{k',r'}^{m'}, g\left(x_{k',r'}^{m'}\right)\right]_{m'\in[M],k'\in[K],r'<r}, \left[x_{k',r}^{m}, g\left(x_{k',r}^{m}\right)\right]_{k'<k}, \xi\right)$$

where ξ is a string of random bits that the algorithm may use for randomization. In addition, for a vector x, we define support$(x) := \{j : x_j \neq 0\}$, and we say that an intermittent communication algorithm is distributed zero-respecting

$$\textit{support}(x_{k,r}^m) \subseteq \bigcup_{m'\in[M],k'\in[K],r'<r} \textit{support}\left(g(x_{k',r'}^{m'})\right) \cup \bigcup_{k'<k} \textit{support}\left(g(x_{k',r}^{m})\right)$$

We construct a hard instance for the lower bound using the scalar functions $\psi : \mathbb{R} \to \mathbb{R}$:

$$\psi(x) = \frac{\sqrt{H}x}{2\beta}\arctan\left(\frac{\sqrt{H}\beta x}{2}\right) - \frac{1}{2\beta^2}\log\left(1 + \frac{H\beta^2x^2}{4}\right) \tag{220}$$

where H is the parameter of smoothness, and $\beta > 0$ is another parameter that controls the third derivative of ψ which we will set later. The hard instance is then

$$F(x) = -\psi'(\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{N-1}\psi(x_{i+1} - x_i) \tag{221}$$

where ζ and N are additional parameters that will be chosen later. Lemma 20 below summarizes the relevant properties of F, whose proof relies on the following bounds on ψ''':

Lemma 19 *For any $H, \beta \geq 0$,*

$$|\psi'''(x)| \leq \frac{H^{3/2}\beta}{12}$$
$$|\psi'''(x)| \leq 2\beta\psi''(x)^{3/2}$$
$$|\psi'''(x)| \leq \frac{\sqrt{H}\beta}{2}\psi''(x)$$

Proof The third derivative of ψ is

$$\psi'''(x) = \frac{-2H^2\beta^2 x}{(4+H\beta^2x^2)^2} \tag{222}$$

For the first claim, we first maximize the simpler function $x \mapsto \frac{x}{(1+x^2)^2}$. We note that

$$\frac{d}{dx}\frac{x}{(1+x^2)^2} = \frac{1-3x^2}{(1+x^2)^3} \tag{223}$$
$$\frac{d^2}{dx^2}\frac{x}{(1+x^2)^2} = \frac{12x(x^2-1)}{(1+x^2)^4} \tag{224}$$

Therefore, the derivative is zero at $\pm 1/\sqrt{3}$ and the second derivative is negative only for $+1/\sqrt{3}$, furthermore, $\lim_{x\to\pm\infty}\frac{x}{(1+x^2)^2} = 0$. Therefore, we conclude that

$$\max_{x\in\mathbb{R}}\frac{x}{(1+x^2)^2} = \max_{x\in\mathbb{R}}\frac{|x|}{(1+x^2)^2} = \frac{\sqrt{\frac{1}{3}}}{(1+\sqrt{\frac{1}{3}}^2)^2} = \frac{3\sqrt{3}}{16} \tag{225}$$

By rescaling, we conclude that

$$\max_{x\in\mathbb{R}}\left|\psi'''(x)\right| = \max_{x\in\mathbb{R}}\frac{2H^2\beta^2|x|}{(4+H\beta^2x^2)^2} = \frac{H^{3/2}\beta}{4}\max_{x\in\mathbb{R}}\frac{\left|\frac{\sqrt{H}\beta x}{2}\right|}{\left(1+\left(\frac{\sqrt{H}\beta x}{2}\right)^2\right)^2} = \frac{3\sqrt{3}H^{3/2}\beta}{64} < \frac{H^{3/2}\beta}{12} \tag{226}$$

This establishes the first claim. For the second claim, we observe that

$$\left|\psi'''(x)\right| = \frac{2\sqrt{H}\beta^2|x|}{\sqrt{4+H\beta^2x^2}}\psi''(x)^{3/2} \leq \frac{2\sqrt{H}\beta^2|x|}{\sqrt{H\beta^2x^2}}\psi''(x)^{3/2} = 2\beta\psi''(x)^{3/2} \tag{227}$$

Finally, for the third claim, we start by noting

$$\left|\psi'''(x)\right| = \frac{2H\beta^2|x|}{4+H\beta^2x^2}\psi''(x) \tag{228}$$

We now consider the function $x \mapsto \frac{x}{1+x^2}$, for which

$$\frac{d}{dx}\frac{x}{1+x^2} = \frac{1-x^2}{(1+x^2)^2} \tag{229}$$

$$\frac{d^2}{dx^2}\frac{x}{1+x^2} = \frac{2x(x^2-3)}{(1+x^2)^3} \tag{230}$$

We conclude that

$$\max_{x\in\mathbb{R}} \frac{|x|}{1+x^2} = \frac{1}{1+1^2} = \frac{1}{2} \tag{231}$$

and therefore,

$$\max_{x\in\mathbb{R}} \frac{2H\beta^2|x|}{4+H\beta^2x^2} = \sqrt{H}\beta \max_{x\in\mathbb{R}} \frac{\left|\frac{\sqrt{H}\beta x}{2}\right|}{1+\left(\frac{\sqrt{H}\beta x}{2}\right)^2} = \frac{\sqrt{H}\beta}{2} \tag{232}$$

This completes the proof. ■

Lemma 20 *For any $H \geq 0$, $\beta > 0$, $\zeta > 0$, and $N \geq 2$, F is convex, H-smooth, β-self-concordant, $\frac{\sqrt{H}\beta}{2}$-quasi-self-concordant, and $\|\nabla^3 F(x)\| \leq \frac{4H^{3/2}\beta}{3}$.*

Proof First, we note that $0 \leq \psi''(x) = \frac{H}{4+H\beta^2x^2} \leq \frac{H}{4}$. Therefore, F is the sum of convex functions and is thus convex itself. We now compute the Hessian of F:

$$\nabla^2 F(x) = \psi''(x_N)e_Ne_N^\top + \sum_{i=1}^{N-1}\psi''(x_{i+1}-x_i)(e_{i+1}-e_i)(e_{i+1}-e_i)^\top \tag{233}$$

Therefore, for any $u \in \mathbb{R}$,

$$\begin{aligned} u^\top\nabla^2 F(x)u &\leq \psi''(x_N)u_N^2 + \sum_{i=1}^{N-1}\psi''(x_{i+1}-x_i)(u_{i+1}-u_i)^2 && (234)\\ &\leq \frac{H}{4}\left[u_N^2 + \sum_{i=1}^{N-1} 2u_{i+1}^2 + 2u_i^2\right] && (235)\\ &\leq H\|u\|^2 && (236)\end{aligned}$$

We conclude that $\nabla^2 F(x) \preceq H \cdot I$ and thus F is H-smooth.

Next, we compute the tensor of 3rd derivatives of F:

$$\nabla^3 F(x) = \psi'''(x_N)e_N^{\otimes 3} + \sum_{i=1}^{N-1}\psi'''(x_{i+1}-x_i)(e_{i+1}-e_i)^{\otimes 3} \tag{237}$$

where

$$\psi'''(x) = \frac{-2H^2\beta^2 x}{(4+H\beta^2x^2)^2} \tag{238}$$

Therefore, for any $u \in \mathbb{R}$,

$$\left|\nabla^3 F(x)[u,u,u]\right| \leq \left|\psi'''(x_N)u_N^3\right| + \sum_{i=1}^{N-1}\left|\psi'''(x_{i+1}-x_i)(u_{i+1}-u_i)^3\right| \tag{239}$$

We can bound this in several different ways using Lemma 19:

$$|\psi'''(x)| \leq \frac{H^{3/2}\beta}{12} \tag{240}$$
$$|\psi'''(x)| \leq 2\beta\psi''(x)^{3/2} \tag{241}$$
$$|\psi'''(x)| \leq \frac{\sqrt{H}\beta}{2}\psi''(x) \tag{242}$$

Therefore,

$$\begin{aligned}\left|\nabla^3 F(x)[u,u,u]\right| &\leq \left|\psi'''(x_N)u_N^3\right| + \sum_{i=1}^{N-1}\left|\psi'''(x_{i+1}-x_i)(u_{i+1}-u_i)^3\right| &(243)\\ &\leq \frac{H^{3/2}\beta}{12}\left[\left|u_N\right|^3 + 8\sum_{i=1}^{N-1}\left|u_{i+1}\right|^3 + \left|u_i\right|^3\right] &(244)\\ &\leq \frac{4H^{3/2}\beta}{3}\|u\|^3 &(245)\end{aligned}$$

Above, we used that $|a-b|^3 \leq (|a|+|b|)^3 \leq 8(|a|^3+|b|^3)$. We conclude that $\|\nabla^3 F(x)\| \leq \frac{4H^{3/2}\beta}{3}$.

Similarly,

$$\begin{aligned}|\nabla^3 F(x)[u,u,u]| &\leq |\psi'''(x_N)||u_N|^3 + \sum_{i=1}^{N-1}|\psi'''(x_{i+1}-x_i)||u_{i+1}-u_i|^3 &(246)\\ &\leq 2\beta\left[\psi''(x_N)^{3/2}(u_N^2)^{3/2} + \sum_{i=1}^{N-1}\psi''(x_{i+1}-x_i)^{3/2}((u_{i+1}-u_i)^2)^{3/2}\right] &(247)\\ &\leq 2\beta\left[\psi''(x_N)u_N^2 + \sum_{i=1}^{N-1}\psi''(x_{i+1}-x_i)(u_{i+1}-u_i)^2\right]^{3/2} &(248)\\ &= 2\beta\left\langle\nabla^2 F(x)u,\, u\right\rangle^{3/2} &(249)\end{aligned}$$

For the final inequality, we used that $|a|^{3/2} + |b|^{3/2} \leq (|a|+|b|)^{3/2}$. We conclude that F is β-self-concordant.

Finally,

$$\begin{aligned}
\left|\nabla^3 F(x)[u,u,u]\right| &\leq |\psi'''(x_N)||u_N|^3 + \sum_{i=1}^{N-1} |\psi'''(x_{i+1}-x_i)||u_{i+1}-u_i|^3 && (250)\\
&\leq \frac{\sqrt{H}\beta}{2}\left[\psi''(x_N)|u_N|^3 + \sum_{i=1}^{N-1} \psi''(x_{i+1}-x_i)|u_{i+1}-u_i|^3\right] && (251)\\
&\leq \frac{\sqrt{H}\beta}{2}\left[\psi''(x_N)|u_N|^2 + \sum_{i=1}^{N-1} \psi''(x_{i+1}-x_i)|u_{i+1}-u_i|^2\right] && (252)\\
&\qquad\qquad \cdot \max\left\{|u_N|, \max_{1\leq i\leq N-1}|u_{i+1}-u_i|\right\}\\
&\leq \frac{\sqrt{H}\beta}{2}\|u\|\nabla^2 F(x)[u,u] && (253)
\end{aligned}$$

For the second to last line, we applied the Hölder inequality $\sum_i |a_i b_i| \leq \|a\|_1 \|b\|_\infty$. We conclude that F is $\frac{\sqrt{H}\beta}{2}$-quasi-self-concordant. ■

We will now proceed to construct a stochastic gradient oracle for F. To do so, we define $\mathrm{prog}(x)$ to be the highest index of a non-zero coordinate of x:

$$\mathrm{prog}(x) = \mathrm{prog}_0(x) = \max\{j \,:\, x_j \neq 0\} \tag{254}$$

With this in hand, we define F^- to be equal to the objective with the $\mathrm{prog}(x)^{\text{th}}$ term removed:

$$F^-(x) = \psi'(-\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{\mathrm{prog}(x)-1} \psi(x_{i+1}-x_i) + \sum_{i=\mathrm{prog}(x)+1}^{N-1} \psi(x_{i+1}-x_i) \tag{255}$$

The stochastic gradient oracle for F is then given by

$$g(x) = \begin{cases} \nabla F^-(x) & \text{with probability } 1-p \\ \nabla F(x) + \frac{1-p}{p}\left(\nabla F(x) - \nabla F^-(x)\right) & \text{with probability } p \end{cases} \tag{256}$$

The following Lemma shows that g is a suitable stochastic gradient oracle for F:

Lemma 21 *For any $H, \beta, \sigma, \zeta, N$, if $p \geq \frac{\pi^2 H}{\pi^2 H + 8\sigma^2\beta^2}$ then for any x*

$$\begin{aligned}
\mathbb{E}g(x) &= \nabla F(x)\\
\mathbb{E}\|g(x) - \nabla F(x)\|^2 &\leq \sigma^2
\end{aligned}$$

Proof First, we compute the expectation of $g(x)$:

$$\mathbb{E}g(x) = (1-p)\nabla F^-(x) + p\left(\nabla F(x) + \frac{1-p}{p}\left(\nabla F(x) - \nabla F^-(x)\right)\right) = \nabla F(x) \tag{257}$$

Second, the variance can be bounded by

$$\mathbb{E}\|g(x)-\nabla F(x)\|^2 = (1-p)\big\|\nabla F^-(x)-\nabla F(x)\big\|^2 + p\left\|\frac{1-p}{p}\big(\nabla F(x)-\nabla F^-(x)\big)\right\|^2 \tag{258}$$

$$= \frac{1-p}{p}\big\|\psi'(x_{\mathrm{prog}(x)+1}-x_{\mathrm{prog}(x)})(e_{\mathrm{prog}(x)+1}-e_{\mathrm{prog}(x)})\big\|^2 \tag{259}$$

$$\le \frac{2(1-p)}{p}\sup_{x\in\mathbb{R}}\big(\psi'(x)\big)^2 \tag{260}$$

$$= \frac{2(1-p)}{p}\cdot\frac{\pi^2 H}{16\beta^2} \tag{261}$$

Therefore, taking $p \ge \frac{\pi^2 H}{\pi^2 H+8\sigma^2\beta^2}$ ensures the variance is bounded by σ^2. ■

In order to prove the lower bound, we will show that with constant probability, all of the iterates generated by any distributed zero-respecting intermittent communication algorithm will have progress $\mathrm{prog}(x) \le N/2$, and we will proceed to show that this implies high suboptimality. The next Lemma upper bounds the progress of the algorithm's iterates:

Lemma 22 *For any $H,\beta,\zeta,\sigma,K,R > 0$ and $N,M \ge 2$, let $p = \max\left\{\frac{2}{K},\frac{\pi^2 H}{\pi^2 H+8\sigma^2\beta^2}\right\}$. Then with probability at least $\frac{1}{2}$, all of the oracle queries made by any distributed zero-respecting intermittent communication algorithm will have progress at most*

$$\max_{m,k,r}\mathrm{prog}(x_{k,r}^m) \le \min\left\{RK,\ \max\left\{48R\log M,\ \frac{4\pi^2 HKR}{\pi^2 H+8\sigma^2\beta^2}\right\}\right\}$$

Proof To begin, fix any vector x and let $j = \mathrm{prog}(x)$. Then since $\psi'(0)=0$,

$$\nabla F^-(x) = \psi'(-\zeta)e_1 + \psi'(x_N)e_N + \sum_{i\neq j}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i) \tag{262}$$

$$= \psi'(-\zeta)e_1 + \sum_{i=1}^{j-1}\psi'(x_{i+1}-x_i)(e_{i+1}-e_i) \tag{263}$$

$$\in \mathrm{span}\{e_1,\dots,e_j\} \tag{264}$$

Therefore, $\mathrm{prog}(\nabla F^-(x)) \le \mathrm{prog}(x)$, so

$$\mathbb{P}[\mathrm{prog}(g(x)) > \mathrm{prog}(x)] = p \tag{265}$$

By a similar argument, is also easy to confirm that $\mathrm{prog}(g(x)) \le \mathrm{prog}(x)+1$.

By the definition of a distributed zero-respecting algorithm, the k^{th} oracle query on the m^{th} machine in the r^{th} round of communication has progress no greater than the highest progress of any stochastic gradient that is available, i.e. any stochastic gradients computed by any machine in rounds $1,\dots,r-1$, and the first $k-1$ gradients computed on machine m in round r. As shown above, each stochastic gradient oracle query allows the algorithm to increase its progress by at most one, and only with probability p.

Therefore, the maximum amount of progress that can be made on the m^{th} machine during the r^{th} round of communication is upper bounded by a Binomial(K,p) random variable, and the total progress made by all the machines during the r^{th} round is upper bounded by the maximum of M independent Binomial(K, p) random variables. Let $n_r^m \sim$ Binomial(K, p) denote the amount of progress made by the m^{th} machine during the r^{th}, then for any n

$$\mathbb{P}\left[\max_{m,k,r} \text{prog}(x_{k,r}^m) > n\right] \leq \mathbb{P}\left[\sum_{r=1}^{R} \max_{1\leq m\leq M} n_r^m > n\right] \tag{266}$$

To start, by the union bound and then the Chernoff bound, for each r and any $\epsilon > 0$

$$\mathbb{P}\left[\max_{1\leq m\leq M} n_r^m \geq (1+\epsilon)Kp\right] \leq M\,\mathbb{P}\left[n_r^1 \geq (1+\epsilon)Kp\right] \leq M\exp\left(-\frac{\epsilon^2 Kp}{2+\epsilon}\right) \tag{267}$$

For any random variable $X \in [0, K]$, $\mathbb{E}X = \int_0^K \mathbb{P}[X \geq x]dx$. Therefore, for each r and any $\epsilon > 0$

$$\mathbb{E}\left[\max_{1\leq m\leq M} n_r^m\right] = \int_0^{(1+\epsilon)Kp} \mathbb{P}\left[\max_{1\leq m\leq M} n_r^m \geq x\right]dx + \int_{(1+\epsilon)Kp}^{K} \mathbb{P}\left[\max_{1\leq m\leq M} n_r^m \geq x\right]dx \tag{268}$$

$$\leq (1+\epsilon)Kp + \int_{\epsilon}^{\frac{1-p}{p}} \mathbb{P}\left[\max_{1\leq m\leq M} n_r^m \geq (1+c)Kp\right]dc \tag{269}$$

$$\leq (1+\epsilon)Kp + M\int_{\epsilon}^{\infty} \exp\left(-\frac{c^2 Kp}{2+c}\right)dc \tag{270}$$

$$\leq (1+\epsilon)Kp + M\int_{\epsilon}^{\infty} \exp\left(-\frac{c\epsilon Kp}{2+\epsilon}\right)dc \tag{271}$$

$$= (1+\epsilon)Kp + \frac{M(2+\epsilon)}{\epsilon Kp}\exp\left(-\frac{\epsilon^2 Kp}{2+\epsilon}\right) \tag{272}$$

We apply this result with $\epsilon = 2 + \frac{2\log M}{Kp}$ so, recalling that $p \geq 2/K$ and $M \geq 2$,

$$\mathbb{E}\left[\max_{1\leq m\leq M} n_r^m\right] \leq (1+\epsilon)Kp + \frac{M(2+\epsilon)}{\epsilon Kp}\exp\left(-\frac{\epsilon^2 Kp}{2+\epsilon}\right) \tag{273}$$

$$\leq 4Kp + 2\log M \tag{274}$$

$$= \max\left\{8, \frac{\pi^2 HK}{\pi^2 H + 8\sigma^2\beta^2}\right\} + 2\log M \tag{275}$$

$$\leq \max\left\{24\log M, \frac{2\pi^2 HK}{\pi^2 H + 8\sigma^2\beta^2}\right\} \tag{276}$$

Therefore, in light of (266) we use Markov's inequality to conclude

$$\mathbb{P}\left[\max_{m,k,r} \text{prog}(x_{k,r}^m) > 2R\mathbb{E}\left[\max_{1\leq m\leq M} n_r^m\right]\right] \leq \mathbb{P}\left[\sum_{r=1}^{R} \max_{1\leq m\leq M} n_r^m > 2R\mathbb{E}\left[\max_{1\leq m\leq M} n_r^m\right]\right] \leq \frac{1}{2} \tag{277}$$

We conclude by substituting for $\mathbb{E}[\max_{1\leq m\leq M} n_r^m]$ and noting that $n_r^m \leq K$ always. ■

The final piece of the proof is to show that if $\text{prog}(x) \leq N/2$ then $F(x) - F^*$ is large:

Lemma 23 *For any* $H, \beta, B > 0$ *and* $N \geq 2$*, set* $\zeta^2 = \frac{B^2}{N^3}$*. Then,* $\|x^*\| \leq B$ *and for any* x *such that* $\operatorname{prog}(x) \leq \frac{N}{2}$*,*

$$F(x) - F^* \geq \begin{cases} \frac{N}{6\beta^2} & \beta^2 > \frac{4N^3}{HB^2} \\ \frac{HB^2}{48N^2} & \beta^2 \leq \frac{4N^3}{HB^2} \end{cases}$$

Proof The first-order optimality condition $\nabla F(x^*) = 0$ indicates

$$\begin{aligned} [\nabla F(x^*)]_1 &= 0 = \psi'(-\zeta) - \psi'(x_2^* - x_1^*) \\ [\nabla F(x^*)]_i &= 0 = \psi'(x_i^* - x_{i-1}^*) - \psi'(x_{i+1}^* - x_i^*) \qquad 1 < i < N \\ [\nabla F(x^*)]_N &= 0 = \psi'(x_N^* - x_{N-1}^*) + \psi'(x_N^*) \end{aligned} \tag{278}$$

So, $x_i^* - x_{i+1}^* = \zeta$ for $i < N$, and $x_N^* = \zeta$, therefore,

$$x^* = \zeta \sum_{i=1}^{N} (N - i + 1) e_i \tag{279}$$

The minimizer has norm

$$\|x^*\|^2 = \zeta^2 \sum_{i=1}^{N} (N - i + 1)^2 = \frac{\zeta^2}{6}\left(2N^3 + 3N^2 + N\right) \tag{280}$$

We therefore choose $\zeta^2 = \frac{B^2}{N^3}$ so that $\|x^*\| \leq B$. In this case,

$$\begin{aligned} \min_{x:\|x\|\leq B} F(x) &= F(x^*) && (281) \\ &= \psi'(-\zeta)x_1^* + \psi(x_N^*) + \sum_{i=1}^{N-1} \psi\left(x_{i+1}^* - x_i^*\right) && (282) \\ &= N\zeta\psi'(-\zeta) + N\psi(\zeta) && (283) \\ &= -N\zeta\frac{\sqrt{H}}{2\beta}\arctan\left(\frac{\sqrt{H}\beta\zeta}{2}\right) && (284) \\ &\quad + N\left[\frac{\sqrt{H}\zeta}{2\beta}\arctan\left(\frac{\sqrt{H}\beta\zeta}{2}\right) - \frac{1}{2\beta^2}\log\left(1 + \frac{H\beta^2\zeta^2}{4}\right)\right] \\ &= -\frac{N}{2\beta^2}\log\left(1 + \frac{HB^2\beta^2}{4N^3}\right) && (285) \end{aligned}$$

Now, consider some x such that $\operatorname{prog}(x) = n \leq \frac{N}{2}$, and observe that

$$\begin{aligned} F(x) &= \psi'(-\zeta)x_1 + \psi(x_N) + \sum_{i=1}^{N-1} \psi(x_{i+1} - x_i) && (286) \\ &= \psi'(-\zeta)x_1 + \psi(x_n) + \sum_{i=1}^{n-1} \psi(x_{i+1} - x_i) && (287) \end{aligned}$$

Therefore, by the same argument as above,

$$F(x) \geq -\frac{n}{2\beta^2}\log\left(1 + \frac{HB^2\beta^2}{4N^3}\right) \tag{288}$$

and we conclude that

$$F(x) - F^* \geq \frac{N}{4\beta^2}\log\left(1 + \frac{HB^2\beta^2}{4N^3}\right) \tag{289}$$

From here, we consider two cases, if $\beta^2 > \frac{4N^3}{HB^2}$, then

$$F(x) - F^* > \frac{N}{4\beta^2}\log(2) > \frac{N}{6\beta^2} \tag{290}$$

Otherwise, if $\beta^2 \leq \frac{4N^3}{HB^2}$ then we use that for $x \leq 1$, $\log(1+x) \geq \frac{x}{2}$ and conclude

$$F(x) - F^* > \frac{N}{4\beta^2} > \frac{N}{6\beta^2} \cdot \frac{HB^2\beta^2}{8N^3} = \frac{HB^2}{48N^2} \tag{291}$$

This completes the proof. ■

We are now ready to prove a lower bound in terms of β, which Theorem 4 instantiates for different constraints on the objective:

Lemma 24 *For any $H, B, \sigma, K, R, \beta > 0$ and any $M \geq 2$, there exists a convex, H-smooth objective F with $\|x^*\| \leq B$ and a stochastic gradient oracle g with $\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ for all x such that with probability at least $\frac{1}{2}$, all of the oracle queries, $\{x_{k,r}^m\}$, made by any distributed zero-respecting intermittent communication algorithm have suboptimality*

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c\cdot\left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\beta^4\sigma^4B^2}{HK^2R^2}, \frac{\sigma B}{\sqrt{KR}}\right\}\right]$$

Furthermore, the objective F is simultaneously β-self-concordant, $\frac{\sqrt{H}\beta}{2}$-quasi-self-concordant, and has $\sup_{x,u}|\nabla^3 F(x)[u,u,u]| \leq \frac{4H^{3/2}\beta}{3}\|u\|^3$.

Proof By Lemma 20, F as defined in (221) is H-smooth, convex, β-self-concordant, $\frac{\sqrt{H}\beta}{2}$-quasi-self-concordant, and has $\sup_{x,u}|\nabla^3 F(x)[u,u,u]| \leq \frac{4H^{3/2}\beta}{3}\|u\|^3$. Furthermore, by Lemma 21, g as defined in (256) is unbiased and has variance bounded by σ^2 for the choice of p used in Lemma 22. Finally, when we choose $\zeta^2 = \frac{B^2}{N^3}$, the minimizer of F has norm $\|x^*\| \leq B$ by Lemma 23. Therefore, the objective F and stochastic gradient oracle g are suitable for the lower bound.

By Lemma 22, with probability at least $\frac{1}{2}$, all of the iterates of any distributed zero-respecting intermittent communication algorithm will have progress at most

$$\max_{m,k,r} \operatorname{prog}(x_{k,r}^m) \leq \min\left\{RK,\ \max\left\{48R\log M,\ \frac{4\pi^2 HKR}{\pi^2 H + 8\sigma^2\beta^2}\right\}\right\} \tag{292}$$

For the rest of the proof, we condition on this event and set

$$N = \min\left\{2KR,\ \max\left\{96R\log M,\ \frac{8\pi^2 HKR}{\pi^2 H + 8\sigma^2\beta^2}\right\}\right\} \tag{293}$$

so that $\max_{m,k,r} \operatorname{prog}(x_{k,r}^m) \leq \frac{N}{2}$. By Lemma 23, this means that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq \begin{cases} \frac{N}{6\beta^2} & \beta^2 > \frac{4N^3}{HB^2} \\ \frac{HB^2}{48N^2} & \beta^2 \leq \frac{4N^3}{HB^2} \end{cases} \tag{294}$$

Thus, if $\beta^2 \leq \frac{4N^3}{HB^2}$, then

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq \frac{HB^2}{48N^2} = \frac{HB^2}{48 \min\left\{4K^2R^2,\ \max\left\{9216R^2\log^2 M,\ \frac{64\pi^4H^2K^2R^2}{(\pi^2H+8\sigma^2\beta^2)^2}\right\}\right\}} \tag{295}$$

Therefore, there is a universal constant c such that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left(\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sigma^4\beta^4B^2}{HK^2R^2}\right\}\right) \tag{296}$$

On the other hand, if $\beta^2 > \frac{4N^3}{HB^2}$, since $N = N(\beta)$ is a non-increasing function of β, we can always instantiate the lower bound in terms of $\beta' < \beta$ such that $\beta'^2 = \frac{4N(\beta')^3}{HB^2}$. With this choice, in light of (295), there is a universal constant c such that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{(\pi^2H + 8\sigma^2\beta'^2)^2B^2}{HK^2R^2}\right\}\right] \tag{297}$$

Furthermore, with our choice of β',

$$\beta'^2 = \frac{4N(\beta')^3}{HB^2} \tag{298}$$

$$= \frac{4}{HB^2}\min\left\{8K^3R^3,\ \max\left\{96^3R^3\log^3 M,\ \frac{8^3\pi^6H^3K^3R^3}{(\pi^2H+8\sigma^2\beta'^2)^3}\right\}\right\} \tag{299}$$

$$\geq \frac{4K^3R^3}{HB^2}\min\left\{8,\ \frac{8^3\pi^6H^3}{(\pi^2H+8\sigma^2\beta'^2)^3}\right\} \tag{300}$$

$$\geq \frac{32K^3R^3}{HB^2}\frac{\pi^6H^3}{(\pi^2H+8\sigma^2\beta'^2)^3} \tag{301}$$

Therefore,

$$\frac{1}{8\sigma^2}(\pi^2H + 8\sigma^2\beta'^2)^4 \geq \beta'^2(\pi^2H+8\sigma^2\beta'^2)^3 \geq \frac{32\pi^6H^2K^3R^3}{B^2} \tag{302}$$

$$\implies (\pi^2H+8\sigma^2\beta'^2)^2 \geq \frac{16\pi^3H\sigma K^{3/2}R^{3/2}}{B} \tag{303}$$

In light of (297), we conclude that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right] \tag{304}$$

Combining Lemma 16 with (296) and (304) completes the proof. We note that the lower bound in Lemma 16 is achieved by a quadratic hard instance, which is 0-self-concordant, 0-quasi-self-concordant, and has 0-Lipschitz Hessian. ■

Theorem 4 *For any $H, B, \sigma, Q, K, R > 0$ and any $M \geq 2$, there exists a convex, H-smooth objective F with $\|x^*\| \leq B$ and with $\nabla^2 F$ being Q-Lipschitz with respect to the L2 norm, and a stochastic gradient oracle g with $\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ for all x, such that with probability at least $\frac{1}{2}$ all of the oracle queries $\{x_{k,r}^m\}$ made by any distributed-zero-respecting intermittent communication algorithm (see Definition 18 in Appendix D) will have suboptimality*

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \Bigg[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2 \log^2 M}, \frac{\sqrt{Q\sigma}B^2}{K^{1/4}R^2 \log^{7/4} M}, \frac{\sigma B}{\sqrt{KR}}\right\}\Bigg]$$

Proof By Lemma 24, the objective F and stochastic gradient oracle g satisfy the necessary conditions with F having $\|\nabla^3 F(x)\| \leq \frac{4H^{3/2}\beta}{3}$. In terms of β and the other problem parameters, the lower bound is then

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2 \log^2 M}, \frac{\beta^4\sigma^4 B^2}{HK^2R^2}, \frac{\sigma B}{\sqrt{KR}}\right\}\right] \tag{305}$$

with probability $\frac{1}{2}$. Below, we will use the fact that we can instantiate the lower bound in terms of any $H' \leq H$ without invalidating the result since an H'-smooth objective is also H-smooth. This allows us to maximize the lower bound over $H' \leq H$ to achieve a tighter result. Choosing $\beta = \frac{3Q}{4H^{3/2}}$ ensures that $\|\nabla^3 F(x)\| \leq Q$, i.e. $\nabla^2 F$ is Q-Lipschitz, so

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2 \log^2 M}, \frac{Q^4\sigma^4 B^2}{H^7K^2R^2}, \frac{\sigma B}{\sqrt{KR}}\right\}\right] \tag{306}$$

In the event that the $\frac{Q^4\sigma^4B^2}{H^7K^2R^2}$ term is the minimizer, we can take $H' = \frac{\sqrt{Q\sigma}\log^{1/4} M}{K^{1/4}} \leq H$ to conclude

$$\begin{aligned}&\min_{m,k,r} F(x_{k,r}^m) - F^* \\ &\geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2 \log^2 M}, \frac{\sqrt{Q\sigma}B^2}{K^{1/4}R^2 \log^{7/4} M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right]\end{aligned} \tag{307}$$

This completes the proof. ■

Theorem 25 *For any $H, B, \sigma, Q, K, R > 0$ and any $M \geq 2$, there exists a convex, H-smooth objective F with $\|x^*\| \leq B$ and a stochastic gradient oracle g with $\mathbb{E}\|g(x) - \nabla F(x)\|^2 \leq \sigma^2$ for all x, such with probability at least $\frac{1}{2}$ all of the oracle queries $\{x_{k,r}^m\}$ made by any distributed-zero-respecting intermittent communication algorithm will have suboptimality lower bounded as follows: When the objective, F, is required to be Q-self-concordant*

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2 \log^2 M}, \frac{Q^2\sigma^2 B^2}{KR^2 \log M}, \frac{\sigma B}{\sqrt{KR}}\right\}\right]$$

When the objective, F, is required to be Q-quasi-self-concordant

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{Q\sigma B^2}{\sqrt{K}R^2 \log^{3/2} M}, \frac{\sigma B}{\sqrt{KR}} \right\} \right]$$

Proof By Lemma 24, the objective F and stochastic gradient oracle g satisfy the necessary conditions with F being β-self concordant and $\frac{\sqrt{H}\beta}{2}$-quasi-self-concordant. In terms of β and the other problem parameters, the lower bound is then

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{\beta^4\sigma^4 B^2}{HK^2R^2}, \frac{\sigma B}{\sqrt{KR}} \right\} \right] \tag{308}$$

with probability $\frac{1}{2}$. Below, we will use the fact that we can instantiate the lower bound in terms of any $H' \leq H$ without invalidating the result since an H'-smooth objective is also H-smooth. This allows us to maximize the lower bound over $H' \leq H$ to achieve a tighter result. We proceed by considering the two cases separately.

Self-Concordance: Choosing $\beta = Q$ ensures that F is Q-self-concordant, so

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{Q^4\sigma^4 B^2}{HK^2R^2}, \frac{\sigma B}{\sqrt{KR}} \right\} \right] \tag{309}$$

In the event that the $\frac{Q^4\sigma^4B^2}{HK^2R^2}$ term is the minimizer, we can take $H' = \frac{Q^2\sigma^2 \log M}{K} \leq H$ to conclude that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{Q^2\sigma^2 B^2}{KR^2 \log M}, \frac{\sigma B}{\sqrt{KR}} \right\} \right] \tag{310}$$

Quasi-self-Concordance: Choosing $\beta = \frac{2Q}{\sqrt{H}}$ ensures that F is Q-quasi-self-concordant, so

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{Q^4\sigma^4 B^2}{H^3K^2R^2}, \frac{\sigma B}{\sqrt{KR}} \right\} \right] \tag{311}$$

In the event that the $\frac{Q^4\sigma^4B^2}{H^3K^2R^2}$ term is the minimizer, we can take $H' = \frac{Q\sigma \log^{1/2} M}{\sqrt{K}} \leq H$ to conclude that

$$\min_{m,k,r} F(x_{k,r}^m) - F^* \geq c \cdot \left[\frac{HB^2}{K^2R^2} + \min\left\{ \frac{\sigma B}{\sqrt{MKR}}, HB^2 \right\} + \min\left\{ \frac{HB^2}{R^2 \log^2 M}, \frac{Q\sigma B^2}{\sqrt{K}R^2 \log^{3/2} M}, \frac{\sigma B}{\sqrt{KR}} \right\} \right] \tag{312}$$

This completes the proof. ■