Fully Asynchronous Stochastic Coordinate Descent: A Tight Lower Bound on the Parallelism Achieving Linear Speedup *†

Yun Kuen Cheung
Singapore University of
Technology and Design

Richard Cole   Yixin Tao
Courant Institute, NYU

Abstract

We seek tight bounds on the viable parallelism in asynchronous implementations of coordinate descent that achieves linear speedup. We focus on asynchronous coordinate descent (ACD) algorithms on convex functions which consist of the sum of a smooth convex part and a possibly non-smooth separable convex part.

We quantify the shortfall in progress compared to the standard sequential stochastic gradient descent. This leads to a simple yet tight analysis of the standard stochastic ACD in a partially asynchronous environment, generalizing and improving the bounds in prior work. We also give a considerably more involved analysis for general asynchronous environments in which the only constraint is that each update can overlap with at most q others. The new lower bound on the maximum degree of parallelism attaining linear speedup is tight and improves the best prior bound almost quadratically.

*Part of the work done while Yun Kuen Cheung held positions at Courant Institute, NYU, at Faculty of Computer Science, University of Vienna and at Max Planck Institute for Informatics, Saarland Informatics Campus. He was supported in part by NSF Grant CCF-1217989, the Vienna Science and Technology Fund (WWTF) project ICT10-002, Singapore NRF 2018 Fellowship NRF-NRFF2018-07 and MOE AcRF Tier 2 Grant 2016-T2-1-170. Additionally the research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no. 340506.

†Richard Cole and Yixin Tao's work was supported in part by NSF Grants CCF-1217989, CCF-1527568 and CCF-1909538.

arXiv:1811.03254v4 [math.OC] 2 Aug 2020

1   Introduction

We consider the problem of finding an (approximate) minimum point of a convex function $F : \mathbb{R}^n \to \mathbb{R}$ of the form

$$F(x) = f(x) + \sum_{k=1}^{n} \Psi_k(x_k),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function[1], and each $\Psi_k : \mathbb{R} \to \mathbb{R}$ is a univariate convex function, but may be non-smooth. Such functions occur in many data analysis and machine learning problems, such as linear regression (e.g., the Lasso approach to regularized least squares [28]) where $\Psi_k(x_k) = |x_k|$, logistic regression [21], ridge regression [26] where $\Psi_k(x_k)$ is a quadratic function, and Support Vector Machines [12] where $\Psi_k(x_k)$ is often a quadratic function or a hinge loss (essentially, $\max\{0, x_k\}$).

Due to the enormous size of modern problems, there has been considerable interest in parallel algorithms for the problem in order to achieve speedup, ideally in proportion to the number of processors or cores at hand, called linear speedup. One of the most natural parallel algorithms is to simply have each of the multiple cores perform coordinate descent in an (almost) uncoordinated way. In this work, we analyze the natural parallel version of the standard stochastic version of coordinate descent (SCD): each core, at each of its iterations, chooses the next coordinate to update uniformly at random[2].

One important issue in parallel implementations is whether the different cores are all using up-to-date information for their computations. To ensure this requires considerable synchronization, locking, and consequent waiting. Avoiding the need for the up-to-date requirement, i.e., enabling asynchronous updating, was a significant advance. The advantage of asynchronous updating is to reduce and potentially eliminate the need for waiting. At the same time, as some of the data being used in calculating updates will be out of date, one has to ensure that the out-of-datedness is bounded in some fashion. This is captured by the assumption of q-bounded asynchrony: each update can overlap with at most q others; q is at most the number of cores times the ratio of the lengths of the longest and shortest updates.

The performance of an asynchronous algorithm is typically measured against its sequential counterpart by the *linear speedup* benchmark: if p cores are used in the asynchronous algorithm, the running time is a factor of $\Theta(p)$ faster than the sequential counterpart. In the context of minimizing a convex function, the running time is measured by the convergence rate towards the minimum point.

The asynchronous version of SCD is called *Stochastic Asynchronous Coordinate Descent* (SACD). The question we address in this paper is:

> *What is the maximum possible value of $\widetilde{q}$ such that whenever $q \leq \widetilde{q}$,*
> *SACD is guaranteed to achieve linear speedup*
> *under the q-bounded asynchrony assumption?*

In some prior analyses, in addition to q-bounded asynchrony, several other seemingly natural assumptions were (implicitly) made, but they are unlikely to hold in practice. Several works have successfully avoided the use of some or all of these assumptions, but at the cost of having a substantially smaller $\widetilde{q}$. The main contribution of this paper is to derive the asymptotically best possible value of $\widetilde{q}$, while avoiding the use of every one of these assumptions. We now state our result for

---

[1]In fact, having a continuous gradient suffices.
[2]There are also versions of the sequential algorithm in which different coordinates can be selected with different probabilities.

strongly convex functions informally.

Theorem 1 (Informal). *Let q be an upper bound on how many other updates a single update can overlap. $L_{\overline{\mathrm{res}}}$ and $L_{\max}$ are Lipschitz parameters defined in Section 2. Let $F(x) = f(x) + \sum_{k=1}^{n} \Psi_k(x_k)$ be a strongly convex function with strongly convex parameter μ_F, and suppose $f(x)$ has strongly convex parameter μ_f. Without using any additional assumption, we have: if $q = O\left(\frac{\sqrt{n} L_{\max}}{L_{\overline{\mathrm{res}}}}\right)$, then $\mathbb{E}\left[F(x^{T+1}) - F^*\right] \leq \left(1 - \frac{1}{3}\frac{\mu_F}{n(\mu_F - \mu_f + L_{\max})}\right)^T \cdot \left(F(x^1) - F^*\right)$.*

Standard sequential analyses [19, 25] achieve similar bounds with the $\frac{1}{3}$ replaced by 2; i.e., up to a factor of 6, this is the same rate of convergence. Furthermore, the bound on q is asymptotically tight, as we show in a companion work [10].

Next, we discuss the assumptions which were used or avoided by the prior works concerning SACD. We will also compare their bounds on $\widetilde{q}$ with ours.

Three Assumptions in Prior Work The first analyses to prove rate of convergence bounds for stochastic asynchronous computations were those by Avron et al. [1] (for the Gauss-Seidel algorithm), and by Liu et al. [18] and Liu and Wright [17] (for coordinate descent). Liu et al. [18] imposed a "consistent read" constraint on the asynchrony; the other two works considered a more general "Inconsistent Read" model. Subsequent to Liu and Wright's work, several implicit assumptions, discussed below, were identified by Mania et al. [20] and Sun et al. [27].

Next, we give precise descriptions of the three assumptions used in prior work, and explain why they might not hold in practice. Before doing so, we note that when a core makes an update, it typically comprises four steps: (1) choose a coordinate k to update uniformly at random; (2) read the values of the coordinates needed for Step 3 from the main memory; (3) use the coordinate values read in Step 2, denoted by $\tilde{x}$, to compute the gradient $\nabla_k f(\tilde{x})$; and (4) use the computed gradient to make an update to the value of coordinate k in the main memory. In the sequential case, the values read in Step 2 are the most updated, so the t-th update will read the values from right after the $(t-1)$-st update. But in an asynchronous setting, the values read by each update can be outdated; Assumption 1 was used by Liu et al. [18] to constrain the form of this datedness. In this setting, we let x^t denote the coordinate values in memory right after the $(t-1)$-st update.

Assumption 1. *[Consistent Read (CR)] All the coordinate values read by an update computation may have some delay, but they must appear* simultaneously *at some moment. Precisely, the values read by the t-th update must be of the form $x^{t-\tau}$ for some $\tau \geq 0$.*

It is not hard to see why Assumption 1 does not hold in practice: in Step 2, the values of different coordinates are read one-by-one, not simultaneously. This is why all the later work, including ours, uses the inconsistent read model: the values read by the t-th update can be any of the $(x_1^{t-\tau_1}, \cdots, x_n^{t-\tau_n})$, where each $\tau_j \geq 0$ and some or all of the τ_j's can be distinct.

To describe Assumption 2, note that each update takes a non-trivial amount of time to finish, which we call the *timespan* of the update. Moreover, the timespan of different updates are typically not the same: in an experimental study, Sun et al. [27] showed that iteration lengths in coordinate descent problem instances varied by factors of 2 to 10. Thus, in general, the ordering of the updates based on their starting times (the ST order) is not the same as the ordering of the updates based on their commit times (the CT order). It is clear that in the ST order the random choice of coordinate for each update is independent of the other updates, and thus it is uniformly random, which is a helpful property we desire when analyzing SACD. In contrast, as illustrated in Example 1 in Section 2, in the CT order the choice of coordinate for one update can be influenced by other recently committed updates, and therefore, conditioned on the history of previous updates, the

choice need not be uniformly random; indeed, it is unclear what the distribution of choices of the coordinate to update becomes. We call this the *Undoing of Uniformity*. However, as first pointed out in Mania et al. [20] (see their Section 3.1), several earlier works implicitly made the following Assumption 2, which states that the CT order enjoys the same favorable property as the ST order.

Assumption 2. *[Uniformity Preservation (UP)] When the updates are enumerated using the CT order, the random choice of coordinate for each update is independent of the other updates, and thus it is uniformly random.*

Avron et al. [1] also raised a similar issue w.r.t. their asynchronous Gauss-Seidel algorithm.

To avoid using Assumption 2, one simple solution is to use the ST order instead of the CT order, as was done in [20, 27] for the analysis of SACD on smooth functions. For non-smooth functions, we need a slight twist to the ST order which we call the *Single Coordinate Consistent* (SCC) order; see Section 2 for its definition and justification. However, both the ST and SCC orders create several subtle challenges in the analysis of SACD. Note that just before the t-th update makes its random choice of coordinate, denoted by k_t, some earlier updates might not have committed yet. We remark that *the choice of k_t might affect the updated values computed by those earlier updates.* To see why, suppose that $k_t = 1$, and the t-th update timespan is short. Further, assume no nearby updates pick coordinate 1. Then it is possible that the t-th update commits earlier than the $(t-1)$-st update, and therefore the $(t-1)$-st update might read the value of coordinate 1 computed by the t-th update. For any other random choice of k_t, i.e., if $k_t \neq 1$, then as coordinate 1 has not been updated recently, the $(t-1)$-st update will read an earlier value of coordinate 1. As the reads by the $(t-1)$-st update can differ due to different choices of k_t, the change made by the $(t-1)$-st update is influenced by the choice of k_t. Moreover, due to analogous reasoning, for the t-th update, the coordinate values it reads when $k_t = 1$ can differ from those it reads when $k_t \neq 1$.

The subtlety here is: when we use the ST order, *the "future" (an update which appears later in the ST order) can influence the "past" (an update which appears earlier).* This apparent confusion of causality creates substantial challenges in obtaining a complete and rigorous analysis; several prior work chose to bypass the issue with Assumption 3 or the stronger Assumption 3* below. Again, the fact that Assumption 3 had been used in earlier work was first pointed out in Mania et al. [20] (see their Assumption 5.1).

Assumption 3. *[Common Value (CV)] The random choice of coordinate for an update does not affect the values read by the update.*

Assumption 3* *[Strong Common Value (SCV)] In addition to Assumption 3, the values read by an update are independent of subsequent choices of coordinate.*

Yet another order, named *After Read* (AR), was proposed by Leblond et al. [16], albeit for a different problem. Translated to the SACD algorithm, it would require swapping the order of Steps 1 and 2; i.e., in Step 1 all coordinate values are read, and then in Step 2 a random coordinate is chosen to be updated. Clearly, this will be highly inefficient if the problem is sparse. The AR order would use the times at which Step 2 is started to order the updates; there is no Undoing of Uniformity in this order. However, it will not suffice for non-smooth functions (see our justification of the SCC order in Section 2).

Table 1 provides a comparison of our results with prior work.

Related Work Convex optimization is one of the most widely used methodologies in applications across multiple disciplines; we refer readers to Nesterov's text [22] for an excellent overview. Coordinate Descent is a method that has been widely studied; see Wright [31] for a recent survey.

	Step Size	Maximum Parallelism q with linear speedup	Non Smooth Ψ_k	Avoiding Assumption		
				CR?	UP?	SCV?
Liu et al. [18]	$\Gamma \geq L_{\max}$	$\Theta\left(\frac{L_{\max}\sqrt{n}}{L_{\mathrm{res}}}\right)$	NO	NO	NO	NO
Liu and Wright [17]	$\Gamma \geq 2L_{\max}$	$\Theta\left(\frac{L_{\max}\sqrt{n}}{L_{\mathrm{res}}}\right)^{1/2}$	**YES**	**YES**	NO	NO
Mania et al. [20]	$\Gamma \geq \Theta\left(\frac{L^2}{\mu_f}\right)$	See caption	NO	**YES**	**YES**	NO
Sun et al. [27]	$\Gamma \geq \Theta(qL)$	1	NO	**YES**	**YES**	**YES**
Our Result	$\Gamma \geq L_{\max}$	$\Theta\left(\frac{L_{\max}\sqrt{n}}{L_{\overline{\mathrm{res}}}}\right)$	**YES**	**YES**	**YES**	**YES**

Table 1: Comparisons of the analyses of SACD. See Definition 1 for the specifications of Lipschitz parameters L, $L_{\max}$, L_{res} and $L_{\overline{\mathrm{res}}}$; μ_f is the strong convexity parameter. When there is no non-smooth Ψ_k, the update increment is the computed gradient divided by Γ. Thus, the larger the Γ, the less aggressive the update. Mania et al. achieve linear speedup compared to the case $q = 1$ for $q = O(n^{1/6})$; however, the case $q = 1$ is slower by a factor of $\Theta(L^2/(\mu_f L_{\max}))$ compared to a standard stochastic algorithm. In [17], Liu and Wright implicitly used the Strong Common Value (SCV) assumption, namely that the choice of coordinate for update t does not affect the value of $\tilde{x}^t$ read by update t nor the values read by earlier updates. This is the reason they can use the parameter L_{res} to bound gradient differences. To avoid using the SCV assumption, we have introduced a new but similar parameter $L_{\overline{\mathrm{res}}}$.

Relevant works concerning sequential stochastic coordinate descent include Nesterov [23], Richtárik and Takác [25], and Lu and Xiao [19].

Distributed and asynchronous computation has a long history in optimization, going back at least to the work of Chazan and Miranker [6] in 1969, with subsequent milestones in the work of Baudet [2], and of Tsitsiklis, Bertsekas and Athans [30, 3]; subsequent results include [5, 4]. For a survey formalizing pre-2000 work, see Frommer and Szyld [14]. Also see Avron et al. [1] for an informative discussion of asynchronous linear system solvers.

In the last few years, there have been multiple analyses of various asynchronous parallel implementations of stochastic coordinate descent [18, 17, 20, 27]. We have already mentioned the results of Liu et al. [18] and Liu and Wright [17]. Both obtained bounds for both convex and "optimally" strongly convex functions[3], attaining linear speedup so long as there are not too many cores. Liu et al. [18] obtained bounds similar to ours (see their Corollary 2 and our Section 2), but the version they analyzed is more restricted than ours in two respects: first, they imposed the strong assumption of consistent reads, and second, they considered only smooth functions (i.e., no non-smooth univariate components Ψ_k). The version analyzed by Liu and Wright [17] is the same as ours, but their result requires both the UP and SCV assumptions. Their bound degrades when the parallelism exceeds $\Theta(n^{1/4})$.[4] Our bound has a similar flavor but with a limit of $\Theta(n^{1/2})$.

The analysis by Mania et al. [20] removed the UP assumption and needs only the SCV assumption. However, the maximum parallelism was much reduced (to at most $n^{1/6}$), and their results applied only to smooth strongly convex functions, and furthermore is efficient only on non-sparse problem instances. We note that a major focus of their work concerned a simple analysis of

---

[3]This is a weakening of the standard strong convexity.

[4]This is expressed in terms of a parameter τ, renamed q in this paper, which is essentially the possible parallelism; the connection between them depends on the relative times to calculate different updates.

HOGWILD!, an asynchronous stochastic gradient descent algorithm used in data-intensive machine learning tasks, namely to learn functions of the form $\sum_{e=1}^{N} f_e(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$, and each f_e is convex and corresponds to a loss function for one training data instance. HOGWILD! is due to Niu et al. [24]; it was the first asynchronous and lock-free SGD algorithm, and it achieves linear speedup on sparse problems.

The analysis in Sun et al. [27] removed the CV assumption and partially removed the UP assumption. However, this came at the cost of achieving no parallel speedup. They also noted that a hard bound on the parameter q could be replaced by a probabilistic bound, which in practice seems more plausible.

As already mentioned, a companion work [10] shows the bound on q in this paper is tight.

Another widely studied approach to speeding up gradient and coordinate descent is the use of acceleration. Recently, attempts have been made to combine acceleration and parallelism [15, 13, 11]. But at this point, these results do not extend to non-smooth functions.

In a companion work, Cheung and Cole [8] analyzed asynchronous tatonnement in a class of economies for which tatonnement is equivalent to gradient descent. They gave worst-case analyses for a special family of convex functions arising in these settings [9], while this work focuses on stochastic analyses.

Our Technical Contributions There are two key contributions in our work. First, we identify an amortization approach for demonstrating convergence amid asynchrony. Briefly, each update yields a progress term, modulo an error cost which occurs due to asynchrony. A fraction of the progress per update is used to demonstrate overall progress, while in expectation the remaining fraction of the total progress can be shown to compensate for the error costs of all the updates. In short, it is the amortization of progress against errors that leads to our convergence analysis. With this perspective, it is intuitively clear why we need the bounded asynchrony assumption and the Lipschitz parameter bounds: the former to control how error blows up with the datedness of information being used, and the latter to control how one update affects the gradient measurements of other updates. When we use the SCV assumption as was done by Liu and Wright [17], the amortization approach leads to a clean and fairly short analysis, and also improves the parallelism bound given in [17]; see Section 3.2.

While there is no short answer as to why our approach improves the parallel bound (partly because our analysis is substantially different from the one in [17]), we point out a notable difference between our analysis and those in [17] and [20]. In the two prior works, error bounds are *global* in the sense that they involve distance terms between the current point and the optimal point (see equation (A.18) in [17], and all the lemmas in Appendix A.1 of [20]). In contrast, all our error bounds can be kept *local*, i.e., they can be expressed only in terms of the magnitude of an update and its range of variation, and also of gradient changes due to updates, but the optimal point is not involved in the error bounds at all.

The second key contribution is to provide a rigorous analysis that removes the UP and SCV assumptions. We give a brief explanation of why this is technically challenging. The standard stochastic analysis relies on showing an inequality of the following form: $\mathbb{E}\left[F(x^{t+1}) - F(x^*) \,|\, x^t\right] \leq (1 - \delta^t) \cdot [F(x^t) - F(x^*)]$ for some positive δ^t. To remove the UP assumption, Mania et al. [20] used the ST order, while we use a slight twist (the SCC order); but with either of these orders, a direct use of the standard stochastic analysis is not possible, since with these orders the "future" can affect the "past".

Fundamentally, this apparent confusion in causality occurs because the standard choice of timing notation, i.e., a single integer parameter for ordering all updates, is inherently insufficient to represent the wide range of causality patterns in the asynchronous setting. Consequently, we need to

develop a more sophisticated notation which allows us to conveniently capture all possible causality patterns and derive useful error bounds. The SCV assumption removes the possibility of the future affecting the past, and thus guarantees that x^t is the same regardless the choice of coordinate at time t, which is why it can lead to the aforementioned simple analysis.

One key idea is to judiciously overestimate the error terms affecting the t-th update so that they do not depend on the choice of coordinate by the t-th update, which then allows averaging of the error over this choice. A second observation is that these errors can be expressed in terms of a mutual recursion, which, with the right bounds on q, remain bounded. Very briefly, the mutual recursion provides a way of capturing the maximum possible errors among all possible causality patterns. We will explain how in Section 4.

Organization of the Paper In Section 2, we describe our model of asynchronous coordinate descent and state our results. In Section 3, we give a high-level sketch of the structure of our analysis, and show that with the Strong Common Value assumption we can obtain a simple analysis of SACD; this analysis achieves the maximum possible speedup (i.e., linear speedup with up to $\Theta(\sqrt{n})$ cores). Note that this is the same assumption as in Mania et al.'s result [20] and less restrictive than the assumptions in Liu and Wright's analysis [17]. Then, in Section 4, we give the full analysis of SACD. All omitted proofs can be found in the appendix. Also, for the reader's convenience, at the end of this paper, we provide a table of the notation and parameters we use.

2   Model and Main Results

Recall that we are considering convex functions $F : \mathbb{R}^n \to \mathbb{R}$ of the form $F(x) = f(x) + \sum_{k=1}^n \Psi_k(x_k)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function, and each $\Psi_k : \mathbb{R} \to \mathbb{R}$ is a univariate and possibly non-smooth convex function. We let x^* denote a minimum point of F and X^* denote the set of all minimum points of F. Without loss of generality, we assume that F^*, the minimum value of F, is 0.

We review some standard terminology. Let e_j denote the unit vector along coordinate j.

Definition 1. *The function f is L-Lipschitz-smooth if for any $x, \Delta x \in \mathbb{R}^n$, $\|\nabla f(x + \Delta x) - \nabla f(x)\| \leq L \cdot \|\Delta x\|$. For any coordinates j, k, the function f is L_{jk}-Lipschitz-smooth if for any $x \in \mathbb{R}^n$ and $r \in \mathbb{R}$, $|\nabla_k f(x + r \cdot e_j) - \nabla_k f(x)| \leq L_{jk} \cdot |r|$; as is conventional, we write $L_k \triangleq L_{kk}$. f is L_{res}-Lipschitz-smooth if, for all j, $\|\nabla f(x + r \cdot e_j) - \nabla f(x)\| \leq L_{\text{res}} \cdot |r|$. Let $L_{\max} \triangleq \max_{j,k} L_{jk}$; we note that if f is twice differentiable, then $L_{\max} = \max_j L_{jj}$. Let $L_{\overline{\text{res}}} \triangleq \max_k \left(\sum_{j=1}^n (L_{kj})^2 \right)^{1/2}$.*

Note that if the convex function is s-sparse, meaning that each term $\nabla_k f(x)$ depends on at most s variables, then $L_{\overline{\text{res}}} \leq \sqrt{s} L_{\max}$. When n is huge, it seems plausible that the only feasible problems are going to be sparse ones.

The Difference Between L_{res} and $L_{\overline{\text{res}}}$ In general, $L_{\overline{\text{res}}} \geq L_{\text{res}}$. $L_{\text{res}} = L_{\overline{\text{res}}}$ when the rates of change of the gradient are constant, as for example in quadratic functions such as $x^\mathsf{T} A x + bx + c$. We need $L_{\overline{\text{res}}}$ because we do not make the Common Value assumption, as we explain at the end of the simple analysis in Section 3.

By a suitable rescaling of variables, we may assume that L_{jj} is the same for all j and equals $L_{\max}$. This is equivalent to using step sizes proportional to L_{jj} without rescaling, a common practice.

Next, we define strong convexity.

Definition 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. f is strongly convex with parameter $\mu_f > 0$, if for all x, y, $f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{1}{2} \mu_f \|y - x\|^2$.*

The Update Rule Recall that in a standard coordinate descent, be it sequential or parallel and synchronous, the update rule, applied to coordinate j, first computes the *accurate* gradient $g_j^t \triangleq \nabla_j f(x^t)$, and then performs the update given below.

$$W_j(d,g,x) \triangleq -gd - \Gamma d^2/2 - \Psi_j(x+d) + \Psi_j(x);$$
$$x_j^{t+1} \leftarrow x_j^t + \arg\max_d W_j(d, g_j^t, x_j^t) \triangleq x_j^t + \widehat{d}_j(g_j^t, x_j^t),$$

and $\forall k \neq j$, $x_k^{t+1} \leftarrow x_k^t$, where $\Gamma \geq L_{\max}$ is a parameter controlling the step size. As is well known, if $\Psi_j \equiv 0$, then $\widehat{d}_j(g_j^t, x_j^t) = -g_j^t/\Gamma$, i.e., it is simply an update in proportion to the gradient.

However, in an asynchronous environment, an updating core (or processor) might retrieve outdated information $\tilde{x}^t$ instead of x^t, so the gradient the core computes will be $\tilde{g}_j^t = \nabla_j f(\tilde{x}^t)$, instead of the accurate value $\nabla_j f(x^t)$. Our update rule, which is naturally motivated by its sequential counterpart, is

$$x_j^{t+1} \leftarrow x_j^t + \widehat{d}_j(\tilde{g}_j, x_j^t) \equiv x_j^t + \Delta x_j^t \qquad \text{and} \qquad \forall k \neq j, \ x_k^{t+1} \leftarrow x_k^t. \tag{1}$$

We call this the the t-th update (in the SCC order), and denote it by $\mathcal{U}_t$.

$$\text{We let} \qquad \widehat{W}_j(g,x) \triangleq \max_d W(d,g,x) \equiv W_j(\widehat{d}_j(g,x), g, x).$$

Note that $W_j(0,g,x) = 0$; thus $\widehat{W}_j(g,x) \geq 0$ always. It is well known that in the synchronous case, $\widehat{W}_j(\nabla_j f(x^t), x_j^t)$ is a lower bound on the reduction in the value of F, which we treat as the *progress*. Finally, we let k_t denote the coordinate being updated at time t.

Algorithm 1 SACD Algorithm.

Input: The initial point $x^1 = (x_1^1, x_2^1, \cdots, x_n^1)$.

Multiple processors use a shared memory. Each processor iteratively repeats the following six-step procedure, without any global coordination:

    Step 1: Choose a coordinate $j \in \{1, 2, \cdots, n\}$ uniformly at random.
    Step 2: Retrieve coordinate values $\tilde{x}^t$ from the shared memory.
    Step 3: Compute the gradient $\nabla_j f(\tilde{x}^t)$.
    Step 4: Request a write lock on the memory that stores the (up-to-date) value of the
             j-th coordinate.[5]
    Step 5: Retrieve the j-th coordinate value, then update it using rule (1).[6]
    Step 6: Release the lock acquired in Step 4.

The SACD Algorithm The coordinate descent process starts at an initial point $x^1 = (x_1^1, x_2^1, \cdots, x_n^1)$. Multiple cores then iteratively update the coordinate values. We assume that at each time, there is exactly one coordinate update which is being written (in Step 5 of the SACD algorithm). In practice, since there will be little coordination between cores, it is possible that multiple coordinate values are updated at the same *moment*; but by using an arbitrary tie-breaking rule, we can immediately extend our analyses to these scenarios.

---

[5]Instead of having a lock in lines 4–6, a compare-and-swap operation can be used to perform the update in Line 5. This has the effect of using the hardware lock that is part of the compare-and-swap operation.

[6]Even if the processor had retrieved the value of the j-th coordinate from the shared memory in Step 2, the processor needs to retrieve it again here, because it needs the most updated value when applying update rule (1).

In Algorithm 1, we provide the complete description of SACD. The retrieval times for Step 2 plus the gradient-computation time for Step 3 can be non-trivial, and also in Step 4 a core might need to wait if the coordinate it wants to update is locked by another core. Thus, during this period of time other coordinates are likely to be updated. For each update, we call the period of time spent performing the six-step procedure the *span* of the update. We say that update A *interferes with* update B if the commit time of update A lies in the span of update B.

Later in this section, we discuss why locking is needed and when it can be avoided; we also explain why the random choice of coordinate should be made before retrieving coordinate values.

Managing the Undoing of Uniformity: The Single Coordinate Consistent Order Before stating our result formally, we need to disambiguate our timing scheme. In every asynchronous iterative system, including our SACD algorithm, each procedure runs over a span of time rather than atomically. Generally, these spans are *not consistent* — it is possible for one update to start later than another one but to commit earlier. To create an analysis, we need a scheme that orders the updates in a consistent manner.

Using the commit times of the updates for the ordering seems the natural choice, since this ensures that future updates do not interfere with the current update. This is the choice made in many prior works. However, as discussed by Mania et al. [20], this causes uniformity to be undone, as shown in the following example.

Example 1. *Suppose there are three cores and four coordinates, suppose that the workload for updating x_1 is 2.99 time units, the workloads for updating x_2, x_3, x_4 are 1 time unit (the 2.99 is to avoid ties), and suppose every update takes the same 0.5 time units for Steps 4–6. Assuming the cores all start at the same time, then $\mathbb{P}[k_1 = 1] = 1/4^3$, which is the probability that all cores choose to update the first coordinate. Contrariwise, $\mathbb{P}[k_2 = 1 \mid k_1 = 1] = 1$. And, in general, the probability distribution which the random variable k_t follows is strongly dependent on the recent history.*

When there are many more cores and coordinates than the simple case we just considered, and when the other asynchronous effects[7] are taken into account, it is highly uncertain what is the exact or even an approximate distribution for k_{t+1} conditioned on knowledge of the history of $k_1, \cdots, k_t$.

To bypass the above issue, we introduce the *Single Coordinate Consistent Order*, *SCC* for short, defined as follows. We begin from the updates ordered by start time. Then, for each coordinate separately, we rearrange the updates to this coordinate so that they are in commit order, while collectively occupying the same places in the start ordering. The next example illustrates all three orders. The start times given by the ST order correspond to actual times; but henceforth, the index t will refer to the position of an update in the SCC order, and to the values computed by these updates.

Example 2. *In Figure 1 we show six updates to two variables, x_1 and x_2, starting at times $t = 1$ to 6, and ending at times 7–12. The updates are named U_1–U_6. In the order listings below, to facilitate comparisons, we give each update an argument comprising the variable it updates.*
    Update Orders:

$$ST:\ U_1(x_2), U_2(x_1), U_3(x_1), U_4(x_2), U_5(x_1), U_6(x_2)$$
$$CT:\ U_1(x_2), U_3(x_1), U_6(x_2), U_2(x_1), U_5(x_1), U_4(x_2)$$
$$SCC:\ U_1(x_2), U_3(x_1), U_2(x_1), U_6(x_2), U_5(x_1), U_4(x_2)$$

The updates to x_1 are in the same positions in the ST and SCC orders, in the same order in the CT and SCC orders. Likewise for x_2.

Figure 1: Illustration of the ST, CT and SCC orders

We can also understand the SCC order in terms of start times $1, 2, \ldots, T$. The t-th update in the ST order starts at time t and commits at some integer time in the range $[t+1, t+q+1]$ (this follows from our assumption that the asynchrony is q-bounded). Remember that the "times" are simply providing an ordering; they are not measured in a common unit. $\mathcal{U}_t$, the t-th update in the SCC order, has an integer start time in the range $[\max\{1, t-q+1\}, t+q-1]$ and commits at an integer time in the range $[t+1, t+q+1]$ (see Lemma 1).

Clearly the history has no influence on the choice of k_{t+1}. However, there is a new issue: *future* updates can interfere with the current update. Here the term future is used w.r.t. the SCC order; recall that an update U_a to one coordinate with an earlier starting time can commit later than another later starting update U_b to a different coordinate, and therefore U_b could interfere with U_a.

Further Remarks about the SACD Algorithm In many optimization problems, e.g., those involving sparse matrices, the number of coordinate values needed for computing the gradient in Step 3 of Algorithm 1 is much smaller than n, i.e., in Step 2, the core needs to retrieve only a tiny portion of the full set of coordinate values. Also, the sets of coordinate values needed for computing the gradients along different coordinates can be very different. Therefore, the random choice of coordinate (in Step 1) should be made ahead of the process of retrieving required information from the shared memory.

If the convex function F does not have the univariate non-smooth components, each update simply adds a number, which depends only on the computed gradient, to the current value in the memory. Then the update can be done atomically (e.g., by fetch-and-add[8]), and no lock is required.

However, for general scenarios with univariate non-smooth components, the update to x_j must depend on the value of x_j in memory right before the update (see (1)). Then the update cannot be done atomically, and a lock is necessary. We note that when the number of cores is far fewer than n, say when it is $\epsilon\sqrt{n}$ for some $\epsilon < 1$, delays due to locking can occur, but are unlikely to be

---

[7]E.g., communication delays, interference from other computations (say due to mutual exclusion when multiple cores commit updates to the same coordinate), interference from the operating system and CPU scheduling.

[8]The fetch-and-add CPU instruction atomically increments the contents of a memory location by a specified value.

significant.[9] As already mentioned, even if the update is carried out using a Compare-and-Swap operation, the lock is still present within the hardware implementation of this operation.

Justifying the SCC Order We begin by justifying why Step 5 in the update algorithm needs to use the most up-to-date value of x_1 (or more generally, of x_j), via the the following convex function example.

Example 3. *Let $\overline{F}$ be a convex function on $n - 1$ variables. Then define the n variable convex function F as follows.*

$$F(x) = \frac{1}{2}x_1^2 + \Psi_1(x_1) + \overline{F}(x_2, \ldots, x_n)$$

$$\text{where } \Psi_1(x_1) = \begin{cases} 0 & \text{if } -1 \leq x_1 \leq \frac{1}{2} \\ \infty & \text{otherwise} \end{cases}$$

Suppose $\Gamma = 1$, and suppose $x_1^t = -1$. Further suppose the t-th and $(t+1)$-st updates are both to x_1, and suppose they both read the value x_1^t for their gradient computation. Then they compute increments $\arg\max\{d - d^2/2 - \Psi(x_1 + d) + \Psi(x_1)\}$. If they both used the value -1 for x_1 they would both increment x_1 by $+1$; the two updates would result in x_1 being set to 1, a value for which $F = \infty$.

Note that update rule (1) implies that the sequence $x^1, x^2, \ldots, x^{T+1}$ is obtained by applying the computed increments $\Delta x^1, \Delta x^2, \ldots \Delta x^T$, one at a time, and in this order. For this order to be consistent with Step 5 using the most up-to-date value, we need that in this order, for each individual coordinate, the updates be in their up-to-date order, i.e. in their commit order. This is why we use the SCC order for our analysis. In the next example, we will show that an analysis based on the ST order need not work when F has a non-smooth part Ψ.

Example 4. *Let $\overline{F}$ be a convex function on $n - 1$ variables. Then define the n variable convex function F as follows.*

$$F(x) = \frac{1}{2}x_1^2 + \Psi_1(x_1) + \overline{F}(x_2, \ldots, x_n)$$

$$\text{where } \Psi_1(x_1) = \begin{cases} 0 & \text{if } x_1 \geq -1 \\ \infty & \text{otherwise} \end{cases}$$

Suppose $\Gamma = 1$, and suppose $x_1^0 = -1$. Further suppose there are three consecutive updates to x_1:

- *Updates 1–3 start at times 1–3 respectively.*

- *At time 4, Updates 2 and 3 read the value of x_1 (which equals -1) and calculate the gradient w.r.t. x_1 ($\nabla_{x_1} f = -1$).*

- *At times 5 and 6 respectively, Updates 2 and 3 apply the update on x_1 ($x_1' \leftarrow x_1 - \arg\max_d\{\frac{1}{\Gamma}\nabla_{x_1}f \cdot d - \Gamma d^2/2 - \Psi_1(x_1 + d) + \Psi_1(x_1)\}$). A simple calculation shows that both these updates increments x_1 by 1. Therefore, after time 6, the most up to date value of x_1 is $-1 + 1 + 1 = 1$.*

- *At time 7, Update 1 reads the value of x_1 (which now equals 1 after applying Updates 2 and 3) and calculates the gradient w.r.t. x_1 ($\nabla_{x_1} f = 1$);*

---

[9]The standard *birthday paradox* result states that if $\epsilon\sqrt{n}$ cores each chooses a random coordinate among $[n]$ uniformly, the probability of a collision is $\Theta(\epsilon^2)$.

- *Finally, at time 8, Update 1 applies the update on x_1. After this, the most up to date value of x_1 is 0.*

In this example, the values of Δx for Updates 1–3 are respectively -1, 1, and 1. If we use the ST order and apply Update 1 first, then after this update, the value of x_1 becomes $x_1^0 - 1 = -2$, which is less than -1 and thus $F(x) = \infty$. In contrast, with the SCC order, as these updates are to the same coordinate, we will apply these Δx based on the commit time order, and then $F(x)$ will never be ∞.

2.1 Results

We assume that our algorithms are run until exactly T coordinates are selected and then updated for some pre-specified T. The initial value of x is denoted by x^1, and the first update in each order is said to be at time $t = 1$ w.r.t. that order. The commit times are constrained by the following assumption.

Assumption 4. *There exists a non-negative integer q such that the only updates that might interfere with the update at time t in the ST order are those that commit at times $t+1, t+2, \ldots, t+q$.*

When asynchronous effects are moderate, and if the various gradients have a similar computational cost, the parameter q will typically be bounded above by a small constant times the number of cores.

As we are using the SCC order, we need to express the constraint in terms of the latter ordering.

Lemma 1. *Let $\mathcal{U}_t$ be the t-th update in the SCC order. Its (integer) start time lies in the range $[\max\{1, t - q + 1\}, t + q - 1]$ and its commit time is in the range $[t+1, t+q+1]$. Also, update $\mathcal{U}_s$ in the SCC order might interfere with $\mathcal{U}_t$ only if $s \in [t - 2q + 1, t + q - 1]$.*

For simplicity, we relax the first range to $[t - 2q, t + q]$. Also the earlier an update starts, the greater the variation in values in might read, and so for our analysis, we will assume the start time is $\max\{1, t - q + 1\}$. We cannot set $\mathcal{U}_t$'s commit time in a similar way, however, as its commit time could affect which other updates might read its committed value.

Theorem 2 (SACD Upper Bound). *Given initial point x^1, Algorithm 1 is run for exactly T iterations by multiple cores. Suppose that Assumption 4 holds, $\Gamma \geq L_{\max}$, and $q \leq \min\left\{\frac{\sqrt{n}}{270}, \frac{\Gamma\sqrt{n}}{270 L_{\overline{\mathrm{res}}}}\right\}$.*
(i) If F is strongly convex with parameter μ_F, and f is strongly convex with parameter μ_f, then

$$\mathbb{E}\left[F(x^{T+1})\right] \leq \left[1 - \frac{1}{3n} \cdot \frac{\mu_F}{\mu_F + \Gamma - \mu_f}\right]^T \cdot F(x^1).$$

(ii) Now suppose that F is convex. Let R be the radius of the level set for x^1, $\mathrm{Level}(x^1) = \{x \mid f(x) \leq f(x^1)\}$. Then

$$\mathbb{E}\left[F(x^{T+1})\right] \leq \frac{1}{1 + \min\left\{\frac{1}{12n}, \frac{F(x^1)}{24n\Gamma R^2}\right\} \cdot T} \cdot F(x^1).$$

In a companion paper [10], we show that the first bound is tight up to constant factors. Specifically, for any constant $c \geq 1$, for $q \geq \frac{74\Gamma\sqrt{n}}{L_{\overline{\mathrm{res}}}} + 96c\ln n + 435$, we give a family of convex functions for which, with probability at least $1 - 1/n^c$, the first n^c updates make essentially no progress toward the optimum. This result holds even for smooth convex functions. There remains a constant factor separating the upper and lower bounds, and in this range we do not know how much if any parallel speed-up is possible.

Problem Instances with large L_{res} and $L_{\overline{\mathrm{res}}}$ Both L_{res} and $L_{\overline{\mathrm{res}}}$ can be as large as $\sqrt{n} \cdot L_{\max}$. For problem instances of this type, the bound on q becomes $\Theta(1)$; i.e., it does not demonstrate any parallel speedup.

3   The Basic Framework

Recall that k_t denotes the index of the coordinate that is updated at time t. We let $g_{k_t}^t := \nabla_{k_t} f(x^t)$ denote the value of the gradient along coordinate k_t computed at time t using up-to-date values of the coordinates, and $\tilde{g}_{k_t}^t$ denote the actual value computed, which may use some out-of-date coordinate values.

3.1   Classical Analysis of Stochastic Sequential Coordinate Descent

This classical analysis proceeds by first showing that for any chosen k_t, $F(x^t) - F(x^{t+1}) \geq \widehat{W}_{k_t}(g_{k_t}^t, x_{k_t}^t)$. Taking the expectation yields

$$
\begin{aligned}
\mathbb{E}\left[F(x^t) - F(x^{t+1})\right] &\geq \frac{1}{n} \sum_{j=1}^{n} \widehat{W}_j(g_j^t, x_j^t) \\
&\geq \frac{1}{n} \cdot \frac{\mu_F}{\mu_F + \Gamma - \mu_f} \cdot F(x^t) \qquad \text{(by [25, Lemmas 4,6])} \\
&\triangleq \frac{\alpha}{n} \cdot F(x^t).
\end{aligned}
\tag{2}
$$

Note that in strongly convex case, we have defined $\alpha \triangleq \frac{\mu_F}{\mu_F + \Gamma - \mu_f}$. It follows that $\mathbb{E}\left[F(x^{t+1})\right] \leq (1 - \frac{\alpha}{n}) \cdot \mathbb{E}\left[F(x^t)\right]$; iterating this inequality yields $\mathbb{E}\left[F(x^{t+1})\right] \leq (1 - \frac{\alpha}{n})^t \cdot F(x^1)$.

3.2   Warm-up: A Simple Analysis for the Strongly Convex Case with the Strong Common Value Assumption

The following analysis already generalizes and improves the results shown in Liu et al. [18] and Liu and Wright [17].

Suppose there are a total of T updates. We view the whole stochastic process as a branching tree of height T. Each node in the tree corresponds to the *moment* when some core randomly picks a coordinate to update, and each edge corresponds to a possible choice of coordinate. We use π to denote a path from the root down to some leaf of this tree. A superscript of π on a variable will denote the instance of the variable on path π. Note that for each path π we reorder the coordinate instances so that they are in the SCC order. For each path π and for each coordinate k, this simply reorders the instances of x_k on path π.

Contrary to intuition, in general we cannot associate a single value of x with each node of the tree because future choices of coordinate to update can affect the recent past; thus we need to specify the path in order to know a coordinate value. In contrast, the SCV assumption ensures there is a single value of x for each node. A double superscript of (π, t) will denote the instance of the variable at time t on path π, i.e., right before the t-th update.

As we will be computing expected values by averaging over the n random coordinate choices for the t-th update $\mathcal{U}_t$, we introduce a notation to capture this choice: $\pi(k, t)$ will denote the path with the time t coordinate on path π replaced by coordinate k. Note that $\pi(k_t, t) = \pi$. (Recall that k_t is the coordinate chosen by update $\mathcal{U}_t$ on path π.)

Recall that $x^{\pi,t}$ denotes the value of x on path π when precisely the first $t - 1$ updates in the SCC order have been applied; however, $x^{\pi,t}$ may or may not actually be present in memory at any time. Also, recall that $x_{k_t}^{\pi,t+1} = x_{k_t}^{\pi,t} + \Delta x_{k_t}^{\pi,t}$, and $x_k^{\pi,t+1} = x_k^{\pi,t}$ for $k \neq k_t$, where $\Delta x_{k_t}^{\pi,t}$ is the increment computed by $\mathcal{U}_t$. So $x_{k_t}^{\pi(k,t),t}$ denotes the value of x_{k_t} on path $\pi(k, t)$ immediately prior to update $\mathcal{U}_t$, and $x_{k_s}^{\pi(k,t),s}$ denotes the value of x_{k_s} on path $\pi(k, t)$ immediately prior to update $\mathcal{U}_s$.

Similarly, $g_{k_t}^{\pi,t} \triangleq \nabla_{k_t} f(x^{\pi,t})$ denotes the true (accurate) gradient on path π immediately prior to update $\mathcal{U}_t$, $\tilde{g}_{k_t}^{\pi,t}$ the *inaccurate* gradient used by $\mathcal{U}_t$ on path π, and $\tilde{g}_k^{\pi(k,t),t}$ the *inaccurate* gradient used by $\mathcal{U}_t$ on path $\pi(k,t)$.

To handle the case where *inaccurate* gradients are used, we employ the following two lemmas.

Lemma 2. *If* $\Gamma \geq L_{\max}$, $F(x^{\pi,t}) - F(x^{\pi,t+1}) \geq \widehat{W}_{k_t}(g_{k_t}^{\pi,t}, x_{k_t}^{\pi,t}) - \frac{1}{\Gamma} \cdot (g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t})^2$.

Lemma 3. *If* $\Gamma \geq L_{\max}$, $F(x^{\pi,t}) - F(x^{\pi,t+1}) \geq \frac{\Gamma}{4}\left(\Delta x_{k_t}^{\pi,t}\right)^2 - \frac{1}{\Gamma} \cdot (g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t})^2$, where $\Delta x_{k_t}^{\pi,t}$ denotes the increment computed by update $\mathcal{U}_t$.

Proving these results for smooth functions is straightforward. The version for non-smooth functions is less simple, and makes use of the SCC order; it follows from Lemma 17 in Appendix A.

Combining Lemmas 2 and 3 yields

$$F(x^{\pi,t}) - F(x^{\pi,t+1}) \geq \frac{1}{2} \cdot \widehat{W}_{k_t}(g_{k_t}^{\pi,t}, x_{k_t}^{\pi,t}) + \frac{\Gamma}{8}\left(\Delta x_{k_t}^{\pi,t}\right)^2 - \frac{1}{\Gamma} \cdot (g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t})^2. \tag{3}$$

As we will see, the following claim is one reason why this analysis, which uses the SCV assumption, is much simpler than that for the fully asynchronous setting.

Claim 1. *With the SCV assumption, (i) for any $\tau \leq t$, $\tilde{x}^{\pi(k,t),\tau}$ is the same for any coordinate k, and thus equals $\tilde{x}^{\pi,\tau}$; (ii) for any $\tau \leq t$, $x^{\pi(k,t),\tau}$ is the same for any coordinate k, and thus equals $x^{\pi,\tau}$.*

Proof. Part (i) follows directly from the SCV assumption.

For part (ii), we argue inductively on τ as follows. Suppose the claim holds for earlier times. By part (i), for any two coordinates k, k', $\tilde{x}^{\pi(k',t),\tau-1} = \tilde{x}^{\pi(k,t),\tau-1}$ for $\tau \leq t$. Thus the computed gradients for update $\mathcal{U}_{\tau-1}$ are the same on paths $\pi(k',t)$ and $\pi(k,t)$. Also, as the claim holds for earlier times, the values read on both paths for Step 5 of update $\mathcal{U}_{\tau-1}$ are the same, meaning that these updates are identical and hence so are the outcome of these updates; i.e., $x^{\pi(k',t),\tau} = x^{\pi(k,t),\tau}$. As this is true for all $\tau \leq t$, the claim follows. $\square$

We take expectations over all paths π on both sides of inequality (3). We compute the expectation of $\frac{1}{2} \cdot \widehat{W}_{k_t}(g_{k_t}^{\pi,t}, x_{k_t}^{\pi,t})$ as follows. We group each collection of n paths which differ only on their t-th coordinate choice; in other words, if π is a path in a group, then the n paths in the group are $\pi(1,t), \pi(2,t), \ldots, \pi(n,t)$. We first take the expectation within each group, which is the summation $\frac{1}{2n} \cdot \sum_{k=1}^n \widehat{W}_k(g_k^{\pi(k,t),t}, x_k^{\pi(k,t),t})$. By Claim 1(ii), $x^{\pi(k,t),t} = x^{\pi,t}$ for any coordinate k, and hence $g_k^{\pi(k,t),t} = \nabla_k f(x^{\pi(k,t),t}) = \nabla_k f(x^{\pi,t}) = g_k^{\pi,t}$. Thus the summation simplifies to $\frac{1}{2n} \cdot \sum_{k=1}^n \widehat{W}_k(g_k^{\pi,t}, x_k^{\pi,t})$, which is at least $\frac{\alpha}{2n} \cdot F(x^{\pi,t})$ by inequality (2). Then we take the expectation over all groups to obtain

$$\mathbb{E}\left[F(x^{\pi,t+1})\right] \leq \left(1 - \frac{\alpha}{2n}\right) \cdot \mathbb{E}\left[F(x^{\pi,t})\right]$$
$$- \mathbb{E}\left[\frac{\Gamma}{8} \cdot \left(\Delta x_{k_t}^{\pi,t}\right)^2 - \frac{1}{\Gamma} \cdot (g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t})^2\right]. \tag{4}$$

To obtain $\mathbb{E}\left[F(x^{T+1})\right] \leq \left(1 - \frac{\alpha}{2n}\right)^T \cdot F(x^1)$, it suffices to show that

$$\sum_{t=1}^T \frac{\Gamma}{8} \cdot \mathbb{E}\left[\left(\Delta x_{k_t}^{\pi,t}\right)^2\right]\left(1 - \frac{\alpha}{2n}\right)^{T-t} \geq \sum_{t=1}^T \frac{1}{\Gamma} \cdot \mathbb{E}\left[(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t})^2\right]\left(1 - \frac{\alpha}{2n}\right)^{T-t}. \tag{5}$$

13

In the remainder of this section we give a simple proof of the above inequality. We first prove Lemma 4 below, which bounds the expectation, within each group of n paths, of the gradient differences squared.

Lemma 4. *With the Strong Common Value assumption,*

$$\mathbb{E}_k[(\tilde{g}_k^{\pi(k,t),t} - g_k^{\pi(k,t),t})^2] \leq \frac{3qL_{\text{res}}^2}{n} \sum_{s \in [t-2q,t+q]\setminus\{t\}} \mathbb{E}_k[(\Delta x_{k_s}^{\pi(k,t),s})^2].$$

Proof. By definition, $g_k^{\pi(k,t),t} = \nabla_k f(x^{\pi(k,t),t})$, the gradient of up-to-date point $x^{\pi(k,t),t}$, and $\tilde{g}_k^{\pi(k,t),t} = \nabla_k f(\tilde{x}^{\pi(k,t),t})$, the gradient of the point actually read from main memory, out-of-date point $\tilde{x}^{\pi(k,t),t}$. By Lemma 1, the updates $\mathcal{U}_s$, for $s < t - 2q$, have been written into memory before update $\mathcal{U}_t$ starts. Thus, the difference between $x^{\pi(k,t),t}$ and $\tilde{x}^{\pi(k,t),t}$ is due to a subset U of the updates $\mathcal{U}_s$ with $s \in [t - 2q, t + q] \setminus \{t\}$. Let

$$U = \{t_1, t_2, ..., t_{|U|}\}.$$

Viewing $\Delta x_{k_{t_i}}^{\pi(k,t),t_i}$ as the n-vector with a non-zero entry for coordinate k_{t_i} and zero elsewhere, we have:

$$x^{\pi(k,t),t} = \tilde{x}^{\pi(k,t),t} + \sum_{i=1}^{|U|} \begin{cases} \Delta x_{k_{t_i}}^{\pi(k,t),t_i} & \text{if } t_i < t; \\ -\Delta x_{k_{t_i}}^{\pi(k,t),t_i} & \text{if } t_i > t. \end{cases}$$

We define: $\quad x^{\pi(k,t),t}[j] = \tilde{x}^{\pi(k,t),t} + \sum_{i=1}^{j} \begin{cases} \Delta x_{k_{t_i}}^{\pi(k,t),t_i} & \text{if } t_i < t; \\ -\Delta x_{k_{t_i}}^{\pi(k,t),t_i} & \text{if } t_i > t. \end{cases}$

Then, $x^{\pi(k,t),t}[0] = \tilde{x}^{\pi(k,t),t}$ and $x^{\pi(k,t),t}[|U|] = x^{\pi(k,t),t}$. By the definition of L_{res} and the triangle inequality, we obtain

$$\left\| \nabla f(\tilde{x}^{\pi(k,t),t}) - \nabla f(x^{\pi(k,t),t}) \right\|^2$$

$$\leq \left(\sum_{j=0}^{|U|-1} \left\| \nabla f(x^{\pi(k,t),t}[j+1]) - \nabla f(x^{\pi(k,t),t}[j]) \right\| \right)^2$$

$$\leq \left(\sum_{i=1}^{|U|} L_{\text{res}} \left| \Delta x_{k_{t_i}}^{\pi(k,t),t_i} \right| \right)^2 \leq 3q \sum_{s \in [t-2q,t+q]\setminus\{t\}} L_{\text{res}}^2 \left(\Delta x_{k_s}^{\pi(k,t),s} \right)^2. \tag{6}$$

The last inequality followed from applying the Cauchy-Schwarz inequality to the RHS, and relaxing U to $[t - 2q, t + q] \setminus \{t\}$.

By Claim 1(i), $\tilde{x}^{\pi(k',t),t} = \tilde{x}^{\pi(k,t),t}$. By Claim 1(ii), $x^{\pi(k',t),t} = x^{\pi(k,t),t}$. Thus,

$$\mathbb{E}_k\left[(\tilde{g}_k^{\pi(k,t),t} - g_k^{\pi(k,t),t})^2\right] = \mathbb{E}_k\left[|\nabla_k f(\tilde{x}^{\pi(k,t),t}) - \nabla_k f(x^{\pi(k,t),t})|^2\right]$$

$$= \frac{1}{n}\sum_{k'}|\nabla_{k'}f(\tilde{x}^{\pi(k',t),t}) - \nabla_{k'}f(x^{\pi(k',t),t})|^2$$

$$= \frac{1}{n}\sum_{k'}|\nabla_{k'}f(\tilde{x}^{\pi(k,t),t}) - \nabla_{k'}f(x^{\pi(k,t),t})|^2$$

$$= \frac{1}{n}\cdot\|\nabla f(\tilde{x}^{\pi(k,t),t}) - \nabla f(x^{\pi(k,t),t})\|^2$$

$$\leq \frac{3qL_{\text{res}}^2}{n}\sum_{s\in[t-2q,t+q]\setminus\{t\}}(\Delta x_{k_s}^{\pi(k,t),s})^2 \quad \text{(by 6)}. \tag{7}$$

$\square$

To obtain the bound in (5), it suffices to have

$$\sum_{t=1}^{T}\frac{\Gamma}{8}\cdot\mathbb{E}_k\left[\left(\Delta x_k^{\pi(k,t),t}\right)^2\right]\left(1-\frac{\alpha}{2n}\right)^{T-t}$$

$$\geq \frac{3qL_{\text{res}}^2}{n\Gamma}\sum_{t=1}^{T}\left(1-\frac{\alpha}{2n}\right)^{T-t}\sum_{s\in[t-2q,t+q]\setminus\{t\}}\mathbb{E}_k\left[(\Delta x_{k_s}^{\pi(k,t),s})^2\right]$$

and in turn it suffices that $\frac{9q^2L_{\text{res}}^2}{n}/(1-\frac{\alpha}{2n})^{2q} \leq \frac{\Gamma^2}{8}$. Since $\frac{\alpha}{2n} \leq \frac{1}{2n}$ and $q \ll n$, it suffices that $\frac{9q^2L_{\text{res}}^2}{n}/\frac{1}{2} \leq \frac{\Gamma^2}{8}$, or $q \leq \frac{\sqrt{n}\Gamma}{12L_{\text{res}}}$. The bound in Theorem 1 then follows readily (with $L_{\overline{\text{res}}}$ replaced by L_{res}, and setting $\Gamma = L_{\max}$).

Why $L_{\overline{\text{res}}}$ is needed in general In (7), we are seeking to bound $\text{Diff} = \sum_{k'}\|\nabla_{k'}f(\tilde{x}^{\pi(k',t),t}) - \nabla_{k'}f(x^{\pi(k',t),t})\|^2$. The SCV assumption ensures that $\tilde{x}^{\pi(k',t),t}$ and $x^{\pi(k',t),t}$ are independent of k', but this need not hold in the fully asynchronous setting. As it happens, in the fully asynchronous setting, we will be able to obtain bounds of the form $|\tilde{x}^{\pi(k',t),t} - x^{\pi(k',t),t}| \leq \sum_s \Delta_s$, i.e., independent of k', where the sum is over s with $t-2q \leq s \leq t+q$ and $s \neq t$ (the bounds Δ_s are larger than analogous terms $|\Delta x_{k_s}^{\pi(k,t),s}|$ when the SCV assumption holds). On using the Lipschitz parameters, this gives a bound of the form $\text{Diff} \leq 3q\sum_{s,k'}L_{sk'}^2\Delta_s^2 \leq 3q\sum_s L_{\overline{\text{res}}}^2\Delta_s^2$, which is weaker than the corresponding bound of $3q\sum_s L_{\text{res}}^2(\Delta x_{k_s}^{\pi(k,t),s})^2$ when the SCV assumption holds.

4 The Framework for the General Analysis

At a high level, the new framework has the same general structure as the basic framework described in Section 3.2. It consists of three parts. In the first part, we obtain the following variant of (4) without using the SCV assumption.

$$\mathbb{E}\left[F(x^{t+1})\right] \leq \left(1-\frac{\alpha}{3n}\right)\cdot\mathbb{E}\left[F(x^t)\right] + \mathbb{E}\left[\frac{\Gamma}{8}\left(\Delta x_{k_t}^{\pi,t}\right)^2 - \text{Err}_t\right].$$

Err_t will be specified in Lemma 5 below.

The second part, which is the heart of the analysis, bounds $\mathbb{E}[\text{Err}_t]$ in terms of $\mathbb{E}\left(\Delta x_{k_s}^{\pi,s}\right)^2$, for a suitable range of s values, and other terms $(\mathcal{D}_s)^2$, which we will define later, and which are themselves bounded in terms of $\mathbb{E}\left[\left(\Delta x_{k_u}^{\pi,u}\right)^2\right]$ and $(\mathcal{D}_u)^2$ for a suitable range of u values.

15

The third part deduces the bounds in Theorem 2, by means of a suitable potential function (a.k.a. a Lyapunov function) and an amortized analysis.

4.1 Part 1: Demonstrating Substantial Progress

Recall that $\pi(k, t)$ denotes the path in which coordinate k_t at time t is replaced by coordinate k; to reduce clutter we now abbreviate this as $\pi(k)$. Note that $\pi(k_t) = \pi$. We let $\text{prev}(t, k)$ denote the time of the most recent update to coordinate k, if any, in the time range $[t - 2q, t - 1]$; otherwise, we set it to t.

Lemma 5.

$$\mathbb{E}\left[F(x^t) - F(x^{t+1}) \right]$$

$$\geq \frac{1}{3n^2} \mathbb{E}\left[\sum_{k=1}^{n} \sum_{k_t=1}^{n} \widehat{W}_k(g_k^{\pi(k_t),t}, x_k^{\pi(k_t),t}) \right] + \mathbb{E}\left[\frac{\Gamma}{8} \left(\Delta x_{k_t}^{\pi,t} \right)^2 - \text{Err}_t \right],$$

$$where \quad \text{Err}_t = \frac{1}{3n^2} \Bigg[\sum_{\substack{t-2q \leq s < t \\ \& s = \text{prev}(t,k_s)}} \sum_{k_t=1}^{n} \left(\frac{3}{2\Gamma} \underbrace{\left(\tilde{g}_{k_s}^{\pi(k_t),s} - g_{k_s}^{\pi(k_t),t} \right)^2}_{A} \right.$$

$$\left. + 2\Gamma \underbrace{\left(x_{k_s}^{\pi(k_s),t} - x_{k_s}^{\pi(k_t),t} \right)^2}_{B} + \frac{3\Gamma}{2} \underbrace{\left(\Delta x_{k_s}^{\pi(k_t),s} \right)^2}_{C} \right) \Bigg]$$

$$+ \frac{1}{n^2} \sum_{k=1}^{n} \sum_{k_t=1}^{n} \frac{2}{3\Gamma} \underbrace{\left(g_k^{\pi(k),t} - g_k^{\pi(k_t),t} \right)^2}_{D} + \frac{1}{\Gamma} \underbrace{\left(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t} \right)^2}_{E}.$$

$$By\ (2), \quad \sum_{k=1}^{n} \sum_{k_t=1}^{n} \widehat{W}_k(g_k^{\pi(k_t),t}, x_k^{\pi(k_t),t}) \geq \sum_{k_t=1}^{n} \alpha F(x^{\pi(k_t),t}),$$

$$which\ gives \quad \mathbb{E}\left[F(x^{t+1}) \right] \leq \left(1 - \frac{\alpha}{3n} \right) \mathbb{E}\left[F(x^t) \right] + \mathbb{E}\left[\frac{\Gamma}{8} \left(\Delta x_{k_t}^{\pi,t} \right)^2 - \text{Err}_t \right].$$

To prove Lemma 5, we start from (3), and then apply the following two lemmas regarding shifting the parameters in $\widehat{W}$. (See Appendix A for proofs.)

Lemma 6 ($\widehat{W}$ Shifting on the g parameter). *For any g_j, g_j',*

$$\widehat{W}_j(g_j, x_j) \geq \frac{2}{3} \cdot \widehat{W}_j(g_j', x_j) - \frac{4}{3\Gamma} \cdot (g_j - g_j')^2.$$

Lemma 7 ($\widehat{W}$ Shifting on the x parameter). *Suppose there are ℓ updates to coordinate k over the time interval $[t - 2q, t - 1]$. Then*

$$if\ \ell = 0, \quad \widehat{W}(g_k^{\pi,t}, x_k^{\pi(k),t}) = \widehat{W}(g_k^{\pi,t}, x_k^{\pi,t})$$

$$if\ \ell > 0, \quad \widehat{W}(g_k^{\pi,t}, x_k^{\pi(k),t}) \geq \widehat{W}(g_k^{\pi,t}, x_k^{\pi,t}) - \frac{3}{2\Gamma} \cdot (\tilde{g}_k^{\pi,\text{prev}(t,k)} - g_k^{\pi,t})^2$$

$$- 2\Gamma(x_k^{\pi,t} - x_k^{\pi(k),t})^2 - \frac{3\Gamma}{2} \cdot (\Delta x_k^{\pi,\text{prev}(t,k)})^2.$$

16

4.2 Part 2: Bounding Err_t, the Error Term

We begin by stating the following lemma, which bounds the difference in the increments computed by two updates to a coordinate x_j when the inputs to Step 5 vary. To avoid notational clutter, we write Ψ and $\widehat{d}$ in lieu of Ψ_j and $\widehat{d}_j$ (to review its definition see the update rule in Section 2); also, by x_1 and x_2 we mean two possible values of x_j, and by g_1 and g_2 two possible values of g_j.

Lemma 8. *For any $g_1, g_2, x_1, x_2 \in \mathbb{R}$ and $\Gamma \in \mathbb{R}^+$, $|\widehat{d}(g_1, x_1) - \widehat{d}(g_2, x_2)| \leq |x_1 - x_2| + \frac{1}{\Gamma} \cdot |g_1 - g_2|$, and hence*

$$\left(\widehat{d}(g_1, x_1) - \widehat{d}(g_2, x_2)\right)^2 \leq 2(x_1 - x_2)^2 + \frac{2}{\Gamma^2} \cdot (g_1 - g_2)^2.$$

If Ψ is the zero function, then $|\widehat{d}(g_1, x_1) - \widehat{d}(g_2, x_2)| = \frac{1}{\Gamma} \cdot |g_1 - g_2|$.

4.2.1 Additional Notation

In this subsection, we will be defining notation of the form $\Delta^{\bullet}_{\max} x^{\pi,s}_{k_s}$, where $\bullet$ refers to various parameters we will specify as needed, and the max refers to taking a suitable maximum. Without spelling it out, we will assume the analogous notation with $\Delta^{\bullet}_{\min}$ is also being defined. In addition, we will define $\Delta^{\bullet}_{\text{span}} x^{\pi,s}_{k_s} \triangleq \Delta^{\bullet}_{\max} x^{\pi,s}_{k_s} - \Delta^{\bullet}_{\min} x^{\pi,s}_{k_s}$, and $\Delta^{\bullet}_{\text{var}} x^{\pi,s}_{k_s} \triangleq \max\{|\Delta^{\bullet}_{\min} x^{\pi,s}_{k_s}|, |\Delta^{\bullet}_{\max} x^{\pi,s}_{k_s}|, \Delta^{\bullet}_{\text{span}} x^{\pi,s}_{k_s}\}$.

The next step in our analysis is to generalize Lemma 4 to settings in which the SCV Assumption needs not hold, so as to bound the "error" terms in Err_t. We seek to carry out an analysis analogous to (7). The first difficulty we face is that the bound we obtain is going to depend on the span of possible values of $\Delta x^{\pi,s}_{k_s}$, which we denote by $\Delta_{\text{span}} x^{\pi,s}_{k_s}$, where $\Delta_{\max} x^{\pi,s}_{k_s}$ is the maximum possible value for this increment over all asynchronous schedules on path π, assuming the first $t - 2q - 1$ updates are already fixed. Thus, in addition to bounds on the various gradient differences, we will need to bound $\Delta_{\text{span}} x^{\pi,s}_{k_s}$. We begin with this task.

Notice that we have assumed that the first $t - 2q - 1$ updates are known rather than the first $s - 2q - 1$. To reflect this, we denote the maximum possible value of the update by $\Delta^t_{\max} x^{\pi,s}_{k_s}$, and analogously, we write $\Delta^t_{\text{span}} x^{\pi,s}_{k_s}$. We are interested in (s, t) pairs with $t - 2q \leq s \leq t + q$, or equivalently, $s - q \leq t \leq s + 2q$; these are the updates $\mathcal{U}_s$ whose value may not be determined at the start of update $\mathcal{U}_t$ and which may affect update $\mathcal{U}_t$. We call t the *reference time* for update $\mathcal{U}_s$.

For notational convenience, rather than give a bound on $\Delta^t_{\text{span}} x^{\pi,s}_{k_s}$, we will bound $\Delta^u_{\text{span}} x^{\pi,t}_{k_t}$ instead. So, suppose that the first $u - 2q - 1$ updates have been fixed, for some u with $t - q \leq u \leq t + 2q$. Let $x^{\pi,u,t}_{\max,k_t}$ and $x^{\pi,u,t}_{\min,k_t}$, resp., be the largest and smallest values that $x^{\pi,t}_{k_t}$ could attain with the first $u - 2q - 1$ updates already fixed; similarly, let $\tilde{g}^{\pi,u,t}_{\max,k_t}$ and $\tilde{g}^{\pi,u,t}_{\min,k_t}$ be the largest and smallest gradient values that could be computed by update $\mathcal{U}_t$ with the first $u - 2q - 1$ updates already fixed.

Lemma 8 implies

$$\left(\Delta^u_{\text{span}} x^{\pi,t}_{k_t}\right)^2 \leq 2\left(x^{\pi,u,t}_{\max,k_t} - x^{\pi,u,t}_{\min,k_t}\right)^2 + \frac{2}{\Gamma^2}\left(\tilde{g}^{\pi,u,t}_{\max,k_t} - \tilde{g}^{\pi,u,t}_{\min,k_t}\right)^2.$$

We use a Lipschitz bound to obtain

$$\left(\tilde{g}^{\pi,u,t}_{\max,k_t} - \tilde{g}^{\pi,u,t}_{\min,k_t}\right)^2 \leq \Bigg[\sum_{\substack{t-2q \leq s \leq t+q \\ \text{and } s \neq t}} L_{k_s k_t} \max\left\{\left|\Delta^u_{\max} x^{\pi,s}_{k_s}\right|, \left|\Delta^u_{\min} x^{\pi,s}_{k_s}\right|, \Delta^u_{\text{span}} x^{\pi,s}_{k_s}\right\}\Bigg]^2.$$

17

The reason for the three terms is that in determining each gradient, for each s in the given range, the relevant update could be read or not read; so the difference due to this coordinate could stem from its maximum value, its minimum value, or their difference.

By the Cauchy-Schwartz inequality,

$$
\begin{aligned}
\left(\Delta_{\mathsf{span}}^u x_{k_t}^{\pi,t}\right)^2 \leq\ & 2\left(x_{\max,k_t}^{\pi,u,t} - x_{\min,k_t}^{\pi,u,t}\right)^2 \\
& + \frac{6q}{\Gamma^2} \sum_{\substack{t-2q \leq s \leq t+q \\ \text{and } s \neq t}} L_{k_s k_t}^2 \max\left\{\left|\Delta_{\max}^u x_{k_s}^{\pi,s}\right|^2, \left|\Delta_{\min}^u x_{k_s}^{\pi,s}\right|^2,\right. \\
& \hspace{7cm} \left.\left(\Delta_{\mathsf{span}}^u x_{k_s}^{\pi,s}\right)^2\right\}.
\end{aligned}
\tag{8}
$$

Legitimate Averaging via Exclusion Recall that in Section 3.2, a crucial step for obtaining a good parallelism bound was to perform averaging over the n paths in each group, i.e., to replace the terms $L_{k_s k_t}^2$ by $L_{\overline{\mathsf{res}}}^2$ by averaging over k_t. To do this here we would need $\Delta_{\max}^u x_{k_s}^{\pi,s}$ and $\Delta_{\min}^u x_{k_s}^{\pi,s}$ to have the same value on every path $\pi(k)$. But this need not be the case, because the computation of $\Delta_{\max}^u x_{k_s}^{\pi,s}$ could read the result of update $\mathcal{U}_t$, and therefore depend on the choice of k_t. To address this, we will create terms which upper bound $\Delta_{\max}^u x_{k_s}^{\pi,s}$ and which have the same value on every path $\pi(k)$, and similarly for $\Delta_{\min}^u x_{k_s}^{\pi,s}$.

Our first key observation, concerns $\mathcal{U}_s$ and $\mathcal{U}_t$: one of them commits first. Suppose $\mathcal{U}_s$ commits first; then the value computed by $\mathcal{U}_s$ does not use the value computed by $\mathcal{U}_t$, either directly as an input, or indirectly because its inputs do not use this value either. Otherwise, $\mathcal{U}_s$ has no impact on $\Delta_{\mathsf{span}}^u x_{k_t}^{\pi,t}$.

We introduce new terminology to capture this observation. If update $\mathcal{U}_s$ commits before $\mathcal{U}_t$, we will say that $\mathcal{U}_t$ is *excluded* from the computation of $\mathcal{U}_s$. To capture the exclusion of $\mathcal{U}_t$, we define $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi,s}$ to be the maximum value $\mathcal{U}_s$ can compute on path π, assuming that $\mathcal{U}_s$ commits before $\mathcal{U}_t$, and the first $u - 2q - 1$ updates are fixed.

The updates that cause the output of $\mathcal{U}_s$ to vary are those that might commit before or after $\mathcal{U}_s$; these are always a subset of the $\mathcal{U}_v$ with $v \in [s - 2q, s + q]$. We also observe that if update $\mathcal{U}_s$ commits before $\mathcal{U}_t$, then $\mathcal{U}_s$ commits before any update $\mathcal{U}_v$ with $v > t + q$. We can safely incorporate this constraint in the notation $\Delta_{\max}^u$ and $\Delta_{\max}^{u,\{t\}}$. The notation extends to Δ_{span} and Δ_{var} in the natural way.

Note that $\pi(k)$ and $\pi(k')$ are identical paths apart from the coordinate chosen at time t. The phrase "$\mathcal{U}_t$ is excluded from the computation of $\mathcal{U}_s$" could cause us to conjecture that $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi(k),s}$ are identical for all k, and similarly for the $\Delta_{\min}^{u,\{t\}} x_{k_s}^{\pi(k),s}$. If it were so, we could rewrite (8) as follows, and average over k_t:

$$
\left(\Delta_{\mathsf{span}}^u x_{k_t}^{\pi,t}\right)^2 \leq \dots + \frac{6q}{\Gamma^2} \sum_{\substack{t-2q \leq s \leq t+q \\ \text{and } s \neq t}} L_{\overline{\mathsf{res}}}^2 \cdot \max\left\{\dots, \left(\Delta_{\mathsf{span}}^{u,\{t\}} x_{k_s}^{\pi,s}\right)^2\right\}.
$$

However, this conjecture needs not be true with the current definition, so the above averaging is not yet valid. For, as explained in the next paragraph, a problem can arise if there is a coordinate $k_v = k_t$ with $t < v \leq t + q$; when averaging over all k_t we are certain to encounter paths with this property. (As an aside, we note that the conjecture is true when $\Psi \equiv 0$ and the ST order is used.)

Suppose $k_v = k_t$, $v \neq t$, and suppose $\mathcal{U}_t$ is excluded. Given the SCC order, it would appear $\mathcal{U}_v$ should also be excluded. But then suppose there is some other update $\mathcal{U}_s$ with $t - 2q \leq s \leq t + q$, $k_s \neq k_t$, and on some path $\pi(k)$ where $k \neq k_t$, in computing $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi,s}$, $\mathcal{U}_s$ reads the result of

update $\mathcal{U}_v$. Then to be sure the same maximum value were computed by $\mathcal{U}_s$ on path π, we would need it to read this excluded value. So this value cannot be excluded. Instead, we define the computation to act as if it were in the SCC order, but with update $\mathcal{U}_t$ simply not present, i.e., as if there were a total of $T - 1$ updates over the whole computation. (This only pertains to updates $\mathcal{U}_s$ with $t - 2q \leq s \leq t + q$ and $s \neq t$.)

This looks promising, as we would anticipate that $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi,s} \leq \Delta_{\max}^u x_{k_s}^{\pi,s}$, and so it would seem the last term on the RHS can be bounded recursively. But unfortunately, this property need not hold if there are updates to the same coordinate on π in the range $[t + 1, \min\{s, u\} + q]$. To understand the issue we revisit Example 3. Suppose updates $\mathcal{U}_t$ and $\mathcal{U}_{t+1}$ are both to coordinate x_1, with $x_1^t = -1$ as before. If both updates are present, and both read value $x_1 = -1$ for their gradient computation, then $\mathcal{U}_t$ computes an increment of 1 and $\mathcal{U}_{t+1}$ an increment of $\frac{1}{2}$. However, if $\mathcal{U}_t$ is excluded, then $\mathcal{U}_{t+1}$ computes an increment of 1; i.e., $\Delta_{\max}^{u,\{t\}} x_{k_{t+1}}^{\pi,t+1} > \Delta_{\max}^u x_{k_{t+1}}^{\pi,t+1}$, contrary to the desired property.

To avoid the difficulty illustrated by the above example, we will need to modify the definition of $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi,s}$. These modifications enable Lemma 9 below, which ensures $\Delta_{\max}^{u,\{t\}} x_{k_s}^{\pi,s}$ has the properties we need to carry out the averaging in the analysis. Recall that depending on the choice of asynchronous schedule, $\mathcal{U}_s$ may or may not read values computed by updates in the range $[s - 2q, \min\{s, u\} + q] \setminus \{t\}$. We will want to pretend that $\mathcal{U}_s$ can make this choice for an expanded range of updates, namely a subset of $[s - 4q, \min\{s, u\} + q] \setminus \{t\}$. In effect, this enlarges the set of possible asynchronous schedules. To be very precise: let l be the maximum commit time for updates $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_{u-4q-1}$; $\mathcal{U}_s$ may read or not read any values computed by updates $\mathcal{U}_r$ that commit after time l for $r \in [s - 4q, \min\{s, u\} + q] \setminus \{t\}$. We call this $\mathcal{U}_r$'s *extended computation*.

So as to identify the updates $\mathcal{U}_r$ with $r \geq u - 4q$ that commit by time l, we introduce the following set $A^{\pi,u}$ of updates: $A^{\pi,u} = \{r \mid u - 4q \leq r < u - 2q$ and $\mathcal{U}_r$ has committed before some $\mathcal{U}_p$ for $p < u - 4q\}$, which means the updates in $A^{\pi,u}$ have committed before $\mathcal{U}_u$ starts its extended computation. Also, rather than just excluding $\mathcal{U}_t$, we allow any additional subset $S = \{v \mid u - 4q \leq v \leq u + q\} \setminus A^{\pi,u}$ of updates to be excluded; we follow this by maximizing over all such S. This also implies that the range of s in which we are interested becomes $u - 4q \leq s \leq u + q$. Note that if $s \in A^{\pi,u}$ then $\Delta_{\mathsf{span}}^{u,\{t\}} x_{k_s}^{\pi,s} = 0$.

Later on, we will be allowing the exclusion of sets R other than just $\{t\}$, so we also incorporate this in our definitions. For R, S disjoint from $A^{\pi,u}$, we define:

$$\Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s} \triangleq \begin{array}{l}\text{the maximum value that } \Delta x_{k_s}^{\pi,s} \text{ can assume when the} \\ \text{first } (u - 4q - 1) \text{ updates and all updates in } A^{\pi,u} \text{ on} \\ \text{path } \pi \text{ have been fixed, and update } \mathcal{U}_v \text{ is excluded} \\ \text{from the computation of } \mathcal{U}_s \text{ for } v \in R \cup S \text{ and for} \\ v > u + q;\end{array}$$

$$\Delta_{\max}^{u,R} x_{k_s}^{\pi,s} \triangleq \max_{S \subseteq [u-4q,u+q] \setminus A^{\pi,u} \cup \{s\}} \Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s}.$$

$$\Delta_{\max}^u x_{k_s}^{\pi,s} \triangleq \Delta_{\max}^{u,\emptyset} x_{k_s}^{\pi,s}$$

Now we have the desired properties:

Lemma 9. *i. If $t \in R$ and $u \leq t + 2q$, then $\Delta_{\max}^{u,R} x_{k_s}^{\pi(k),s}$ is identical on every path $\pi(k)$.*

ii. If $s \leq t$ then $\Delta_{\max}^{t,R} x_{k_s}^{\pi,s} \leq \Delta_{\max}^{s,R} x_{k_s}^{\pi,s}$.

iii. If $R \subset R'$ then $\Delta_{\max}^{u,R'} x_{k_s}^{\pi,s} \leq \Delta_{\max}^{u,R} x_{k_s}^{\pi,s}$.

iv. $\Delta_{\max}^{u,R} x_{k_s}^{\pi,s} \leq \Delta_{\max}^{u,\emptyset} x_{k_s}^{\pi,s}$.

Proof. i. We begin by showing that the set $A^{\pi(k),u}$ is the same for all k. Recall that $u \leq t + 2q$. By Lemma 1, the start time for $\mathcal{U}_t$ is at least $t - q + 1 \geq u - 3q + 1$. For $p < u - 4q$, again by Lemma 1, $\mathcal{U}_p$ has commit time at most $u - 4q - 1 + q + 1 = u - 3q$. So if $\mathcal{U}_r$ commits before $\mathcal{U}_p$ for some $p \leq u - 4q$, $\mathcal{U}_r$ has commit time at most $u - 3q - 1$. Thus $\mathcal{U}_r$'s commit time is unaffected by the choice of coordinate by $\mathcal{U}_t$, and therefore $A^{\pi(k),u}$ is the same for all k.

Now, recall that $\pi(k)$ is the path in which coordinate k_t at time t on path π is replaced by coordinate k. By definition, if $t \in R$, then $\mathcal{U}_t$ is excluded from the computation of $\Delta_{\max}^{u,R} x_{k_s}^{\pi,s}$. As the paths $\pi(k)$ are identical apart from their t-th coordinate, and as the sets $A^{\pi(k),u}$ are the same for all these paths, it follows that $\Delta_{\max}^{u,R} x_{k_s}^{\pi(k),s}$ is exactly the same for every k.

ii. This follows from the next two observations: (a) the update does not read any of the variable values computed by updates $\mathcal{U}_v$ for $v > s + q$ and therefore reducing the top end of the range in going from reference time t to s does not change the possible updates computed by $\mathcal{U}_s$; (b) $A^{\pi,s} \cap [t - 4q, t - 2q] \subseteq A^{\pi,t}$, and therefore there are at least as many updates that are not yet fixed with reference time s compared to reference time t; furthermore, the fixed values in $A^{\pi,t} \setminus A^{\pi,s}$ are all values that could be computed in the computation with reference time s.

iii. Recall the definition of $\Delta_{\max}^{u,R} x_{k_s}^{\pi,s}$, and let S' be the set for which $\Delta_{\max}^{u,R'} x_{k_s}^{\pi,s} = \Delta_{\max}^{u,R',S'} x_{k_s}^{\pi,s}$. Now, we let $S = (R' \cup S') \setminus R$; thus $S \cup R = S' \cup R'$. Clearly, $\Delta_{\max}^{u,R',S'} x_{k_s}^{\pi,s} = \Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s} \leq \Delta_{\max}^{u,R} x_{k_s}^{\pi,s}$.

iv. This follows immediately from iii. □

Recall that we defined $\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s} = \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} - \Delta_{\min}^{u,R} x_{k_s}^{\pi,s}$ and $\Delta_{\mathsf{var}}^{u,R} x_{k_s}^{\pi,s} = \max \left\{ \left| \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} \right|, \left| \Delta_{\min}^{u,R} x_{k_s}^{\pi,s} \right|, \Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s} \right\}$. We want to have one term to cover every update in which $x_{k_s}^{\pi,s}$ is involved. Accordingly, we define

$$\overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s} := \max_{\substack{u:u-4q \leq s \\ \leq u+q}} \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} = \max_{\substack{u:s-q \leq u \\ \leq s+4q}} \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} = \max_{s-q \leq u \leq s} \Delta_{\max}^{u,R} x_{k_s}^{\pi,s},$$

where we use Lemma 9(ii) for the final equality.

We let $\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s} = \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} - \Delta_{\min}^{u,R} x_{k_s}^{\pi,s}$, $\overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s} = \overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s} - \overline{\Delta}_{\min}^{R} x_{k_s}^{\pi,s}$ and $\overline{\Delta}_{\mathsf{var}}^{R} x_{k_s}^{\pi,s} = \max \left\{ \left| \overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s} \right|, \left| \overline{\Delta}_{\min}^{R} x_{k_s}^{\pi,s} \right|, \overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s} \right\}$. We simplify the notation when $R = \emptyset$, defining $\overline{\Delta}_{\max} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\max}^{\emptyset} x_{k_s}^{\pi,s}$, $\overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\mathsf{span}}^{\emptyset} x_{k_s}^{\pi,s}$ and $\overline{\Delta}_{\mathsf{var}} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\mathsf{var}}^{\emptyset} x_{k_s}^{\pi,s}$. Clearly, if $u \subseteq [s - q, s]$, then

$$\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s} \leq \overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s} \leq \overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi,s} \quad \text{and} \quad \Delta_{\mathsf{var}}^{u,R} x_{k_s}^{\pi,s} \leq \overline{\Delta}_{\mathsf{var}}^{R} x_{k_s}^{\pi,s} \leq \overline{\Delta}_{\mathsf{var}} x_{k_s}^{\pi,s}. \tag{9}$$

Finally, we will want to know the expected effect of update $\mathcal{U}_t$. Thus, we define

$$(\mathcal{D}_t)^2 \triangleq \mathbb{E}\left[\left(\overline{\Delta}_{\max} x_{k_t}^{\pi,t} - \overline{\Delta}_{\min} x_{k_t}^{\pi,t} \right)^2 \right] \quad \text{and} \quad \left(\Delta_t^X \right)^2 \triangleq \mathbb{E}\left[\left(\Delta x_{k_t}^{\pi,t} \right)^2 \right].$$

Since exactly T updates are made, we assume that $(\mathcal{D}_t)^2, \left(\Delta_t^X \right)^2 \equiv 0$ for $t = 0$ and $t \geq T + 1$ throughout the analysis.

Next, we introduce analogous notation for the gradients.

$$\tilde{g}_{\max,k_s}^{u,R,S,\pi,s} \triangleq \begin{array}{l} \text{the maximum value of } \tilde{g}_{k_s}^{\pi,s} \text{ can assume when the} \\ \text{first } (u-4q-1) \text{ updates on path } \pi \text{ have been fixed, and} \\ \text{update } \mathcal{U}_v \text{ is excluded from the computation of } \mathcal{U}_s \text{ for} \\ v \in R \cup S \text{ and for } v > u+q; \text{ for all } r \in A^{\pi,u}, \text{ the} \\ \text{value of the update } \mathcal{U}_r \text{ is already fixed;} \end{array}$$

$$\tilde{g}_{\max,k_s}^{u,R,\pi,s} \triangleq \max_{S \subseteq [u-4q,u+q]\} \setminus A^{\pi,u} \cup \{s\}} \tilde{g}_{\max,k_s}^{u,R,S,\pi,s};$$

$$\overline{g}_{\max,k_s}^{R,\pi,s} \triangleq \max_{s-q \leq u \leq s} \tilde{g}_{\max,k_s}^{u,R,\pi,s};$$

$$\overline{g}_{\max,k_s}^{\pi,s} \triangleq \overline{g}_{\max,k_s}^{\emptyset,\pi,s} \quad \text{and} \quad \overline{g}_{\mathsf{span},k_s}^{\pi,s} \triangleq \overline{g}_{\max,k_s}^{\pi,s} - \overline{g}_{\min,k_s}^{\pi,s}.$$

4.2.2 Bounding $(\mathcal{D}_t)^2 = \mathbb{E}\left[\left(\overline{\Delta}_{\mathsf{span}} x_{k_t}^{\pi,t} \right)^2 \right]$

We are now ready to bound $(\mathcal{D}_t)^2$. Let $\nu_1 := \frac{20q^2}{n}$ and $\nu_2 = \frac{24q^2 L_{\overline{\mathsf{res}}}^2}{n\Gamma^2}$.

Lemma 10.

$$\Gamma \cdot (\mathcal{D}_t)^2 \quad \leq \quad \left(\frac{\nu_1}{q} + \frac{\nu_2}{q} \right) \Gamma \sum_{s \in [t-5q,t+q] \setminus \{t\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X \right)^2 \right].$$

Proof. Recall that $\overline{\Delta}_{\mathsf{span}} x_{k_t}^{\pi,t} = \max_{t-q \leq r \leq t} \Delta_{\max}^r x_{k_t}^{\pi,t} - \min_{t-q \leq r' \leq t} \Delta_{\min}^{r'} x_{k_t}^{\pi,t}$. We call r and r' the reference parameters. By Lemma 8, we obtain a bound of

$$\max_{t-q \leq r,r' \leq t} \left[2\left(x_{k_t}^{r,\pi,t} - x_{k_t}^{r',\pi,t} \right)^2 + \frac{2}{\Gamma^2} \cdot \left(\tilde{g}_{k_t}^{r,\emptyset,\pi,t} - \tilde{g}_{k_t}^{r',\emptyset,\pi,t} \right)^2 \right], \tag{10}$$

where $x_{k_t}^{r,\pi,t}$ is the value of $x_{k_t}^{\pi,t}$ right before Step 5 of update $\mathcal{U}_t$ when r is the reference parameter; the maximum is also over the maximum and minimum possible values of the four terms in the above expression.

The first difference on the RHS of the above expression is going to involve updates to coordinate x_{k_t}, i.e., updates $\mathcal{U}_{k_s}$ with $k_s = k_t$ and $\min\{r-4q, r'-4q\} \leq s < t$. We will consider the maximum and minimum possible values for these updates. This suggests a bound of $\left(\Delta_{\max}^r x_{k_s}^{\pi,s} - \Delta_{\min}^{r'} x_{k_s}^{\pi,s} \right)^2$ for each such s. But recall that the definition of $\Delta_{\max}^r x_{k_s}^{\pi,s}$ allows the exclusion of updates $\mathcal{U}_v$ for a worst case set of v when computing $\mathcal{U}_s$ (this is the effect of the maximization over S in the definition of $\Delta_{\max}^r$), and therefore the first difference in (10) may be as large as $\left(\Delta_{\max}^r x_{k_s}^{\pi,s} \right)^2$ or $\left(\Delta_{\min}^{r'} x_{k_s}^{\pi,s} \right)^2$, meaning that the actual bound is $\left(\max \left\{ \left| \Delta_{\max}^r x_{k_s}^{\pi,s} \right|, \left| \Delta_{\min}^{r'} x_{k_s}^{\pi,s} \right|, \Delta_{\max}^r x_{k_s}^{\pi,s} - \Delta_{\min}^{r'} x_{k_s}^{\pi,s} \right\} \right)^2$.

By Lemma 9(ii), we can modify the range of r, r' from $[t-q, t]$ to $[\min\{s, t-q\}, s] \subset [s-q, s]$ as $s < t$. Thus the first term in (10) is bounded by

$$\left(\sum_{\substack{t-5q \leq s < t \\ k_s = k_t}} \overline{\Delta}_{\mathsf{var}}^{\{t\}} x_{k_s}^{\pi,s} \right)^2.$$

We use a similar argument to bound the second term in (10). The value of $g_{k_t}^{r,\emptyset,\pi,t}$ can vary due to the updates $\Delta^r x_{k_s}^{\pi,s}$ for $r-4q \leq s \leq r+q$ and $s \neq t$; equivalently, $s-q \leq r \leq s+4q$. Recall also that $t-q \leq r \leq t$. Thus, for $s > t$, the relevant range of r is $[s-q, t] \subset [s-q, s]$, and for $s < t$, as before, the relevant range is also at most $[s-q, s]$. Using the Lipschitz bound for the gradients, we obtain:

$$\max_{t-q \leq r,r' \leq t} \left(\tilde{g}_{k_t}^{r,\emptyset,\pi,t} - \tilde{g}_{k_t}^{r',\emptyset,\pi,t} \right)^2 \leq \left(\sum_{\substack{t-5q \leq s \leq t+q \\ s \neq t}} L_{k_s k_t} \overline{\Delta}_{\mathsf{var}}^{\{t\}} x_{k_s}^{\pi,s} \right)^2. \tag{11}$$

Thus, using the Cauchy-Schwartz inequality for the second inequality below, we obtain

$$\left(\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}\right)^2 \leq 2\bigg(\sum_{\substack{t-5q\leq s<t \\ k_s=k_t}} \overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\bigg)^2 + \frac{2}{\Gamma^2}\bigg(\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}} L_{k_sk_t}\overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\bigg)^2$$

$$\leq 10q\sum_{\substack{t-5q\leq s<t \\ k_s=k_t}} \big(\overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\big)^2 + \frac{12q}{\Gamma^2}\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}} L_{k_sk_t}^2\big(\overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\big)^2.$$

Now we average over all n choices of k_t; consequently, π is now being viewed as a random variable where k_t on π is being chosen uniformly at random, while the coordinates at times other than t are fixed. Notice that on all the paths π being considered in the averaging, the value of each $\Delta_{\mathsf{max}}^{r,\{t\}}x_{k_s}^{\pi,s}$ is the same as their computation does not involve the update to $x_{k_t}^t$, and because at most the first $(t-4q)$ updates have been fixed in any of these terms, none of the updates that could affect the update to $x_{k_t}^t$ in its extended computation have been fixed; similarly for $\Delta_{\mathsf{min}}^{r,\{t\}}x_{k_s}^{\pi,s}$. Consequently, for each s, the term $\overline{\Delta}_{\mathsf{var}}x_{\pi,s}^{k_s}$ is the same on each path. Thus the averaging is simply averaging the values $L_{k_sk_t}$ as k_t varies. Recall that by Definition 1, $L_{\mathsf{res}}^2 = \max_k \sum_{j=1}^n (L_{kj})^2$; this yields

$$\mathbb{E}_{k_t}\Big[\big(\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}\big)^2\Big]$$

$$\leq \frac{10q}{n}\sum_{t-5q\leq s<t} \big(\overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\big)^2 + \frac{12q}{n\Gamma^2}\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}} L_{\mathsf{res}}^2 \cdot \big(\overline{\Delta}_{\mathsf{var}}^{\{t\}}x_{k_s}^{\pi,s}\big)^2$$

$$\leq \frac{10q}{n}\sum_{t-5q\leq s<t} \big(\overline{\Delta}_{\mathsf{var}}x_{k_s}^{\pi,s}\big)^2 + \frac{12q}{n\Gamma^2}\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}} L_{\mathsf{res}}^2 \cdot \big(\overline{\Delta}_{\mathsf{var}}x_{k_s}^{\pi,s}\big)^2 \quad \text{(by (9)).} \tag{12}$$

Now, $\Delta x_{k_s}^{\pi,s} \in \big[\Delta_{\mathsf{min}}^s x_{k_s}^{\pi,s},\ \Delta_{\mathsf{max}}^s x_{k_s}^{\pi,s}\big] \subseteq \big[\overline{\Delta}_{\mathsf{min}}x_{k_s}^{\pi,s},\ \overline{\Delta}_{\mathsf{max}}x_{k_s}^{\pi,s}\big]$; thus, $\big|\overline{\Delta}_{\mathsf{min}}x_{k_s}^{\pi,s}\big|, \big|\overline{\Delta}_{\mathsf{max}}x_{k_s}^{\pi,s}\big| \leq \big|\Delta x_{k_s}^{\pi,s}\big| + \big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)$. Also, $\big(\overline{\Delta}_{\mathsf{min}}x_{k_s}^{\pi,s}\big)^2$, $\big(\overline{\Delta}_{\mathsf{max}}x_{k_s}^{\pi,s}\big)^2 \leq 2\big(\Delta x_{k_s}^{\pi,s}\big)^2 + 2\big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2$. So, $\big(\overline{\Delta}_{\mathsf{var}}x_{k_s}^{\pi,s}\big)^2 \leq 2\big(\Delta x_{k_s}^{\pi,s}\big)^2 + 2\big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2$. Consequently,

$$\mathbb{E}_{k_t}\Big[\big(\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}\big)^2\Big] \leq \mathbb{E}_{k_t}\bigg[\frac{20q}{n}\sum_{t-5q\leq s<t} \big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2 + \big(\Delta x_{k_s}^{\pi,s}\big)^2$$

$$+ \frac{24qL_{\mathsf{res}}^2}{n\Gamma^2}\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}} \big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2 + \big(\Delta x_{k_s}^{\pi,s}\big)^2\bigg].$$

Taking the expectation over every group, which on the RHS amounts to taking the expectation over every path π, yields

$$(\mathcal{D}_t)^2 = \mathbb{E}\Big[\big(\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}\big)^2\Big] \leq \frac{20q}{n}\sum_{t-5q\leq s\leq t}\Big(\mathbb{E}\big[\big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2\big] + \mathbb{E}\big[\big(\Delta x_{k_s}^{\pi,s}\big)^2\big]\Big)$$

$$+ \frac{24qL_{\mathsf{res}}^2}{n\Gamma^2}\sum_{\substack{t-5q\leq s\leq t+q \\ s\neq t}}\Big(\mathbb{E}\big[\big(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\big)^2\big] + \mathbb{E}\big[\big(\Delta x_{k_s}^{\pi,s}\big)^2\big]\Big).$$

Lemma 10 follows. $\qquad\square$

4.2.3 Gradient Bounds

In the previous subsection, one of the terms being bounded was $(\tilde{g}_{k_t}^{r,\emptyset,\pi,t} - \tilde{g}_{k_t}^{r',\emptyset,\pi,t})^2 = (\overline{g}_{\max,k_t}^{\pi,t} - \overline{g}_{\min,k_t}^{\pi,t})^2$.

Here, we will bound term E, $\left(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t}\right)^2$. Unfortunately, $g_{k_t}^{\pi,t}$ might not be in $\left[\overline{g}_{\min,k_t}^{\pi,t}, \overline{g}_{\max,k_t}^{\pi,t}\right]$ and so we cannot simply apply Lemma 10. The reason is that $g_{k_t}^{\pi,t}$ is a function of $x^{\pi(k_t),t}$, and this up-to-date x value could depend on the value of update $\mathcal{U}_t$; for recall that $\mathcal{U}_t$ might finish before some updates $\mathcal{U}_s$ with $s < t$, and the latter updates could then read the updated value of the coordinate being updated by $\mathcal{U}_t$. As already explained, there are also other ways that the choice of k_t by update $\mathcal{U}_t$ could affect earlier updates.

We start by upper bounding term E by $\left(\sum_{l_0 \in [t-2q,t-1]} L_{k_{l_0},k_t} \Delta_{\mathsf{var}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0}\right)^2$. The challenge is that changing either coordinate k_{l_0} or k_t may change the value of $\Delta_{\max}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0}$ or of $\Delta_{\min}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0}$ which implies that a simple averaging of the terms $L_{k_{l_0},k_t}^2$ to obtain a term L_{res}^2 as was done to obtain (12) is not possible. Instead, we will bound this term recursively. The somewhat more general result we will need is stated in the following lemma. Its proof, which is quite involved, is deferred to Appendix A.4.

Let $\Lambda^2 = \frac{L_{\mathsf{res}}^2}{\Gamma^2} + 1$, $r = \frac{160 q^2 \Lambda^2}{n}$, $\nu_3 = \frac{3}{16}\left(\frac{r^2}{1-r} + r\right)$, and $\nu_4 = \frac{6r}{1-r}$.

Lemma 11. *For any $u \in [t-2q,t]$, if $r < 1$, then*

$$
\mathbb{E}\left[\left(\sum_{l_0 \in [t-4q,t+q]\setminus\{u\}} L_{k_{l_0},k_u} \Delta_{\mathsf{var}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0}\right)^2\right]
$$

$$
\leq \frac{\nu_3 \Gamma^2}{q} \sum_{s \in [t-7q,t+q]\setminus\{u\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X\right)^2\right] + \nu_4 \Gamma^2 \left[(\mathcal{D}_u)^2 + \left(\Delta_u^X\right)^2\right].
$$

Then, on substituting for $\mathcal{D}_t$ from Lemma 10, we obtain the following bound on term E.

Claim 2. *[Bounding Term E] If $r < 1$, term E is bounded by:*

$$
\frac{\nu_3 \Gamma}{q} \sum_{s \in [t-7q,t+q]\setminus\{t\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X\right)^2\right]
$$

$$
+ \nu_4 \Gamma\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right) \sum_{s \in [t-5q,t+q]\setminus\{t\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X\right)^2\right] + \nu_4 \Gamma \left(\Delta_t^X\right)^2.
$$

With more effort one can also bound the gradient difference terms A and D as follows, as shown in the appendix.

Claim 3. *[Bounding Term A] If $r < 1$, term A is bounded by:*

$$
\frac{2(\nu_3 + \nu_4)\Gamma}{n} \sum_{s \in [t-7q,t+q]\setminus\{t\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X\right)^2\right] + \frac{\Gamma}{n} \sum_{s \in [t-2q,t-1]} \left(\Delta_s^X\right)^2
$$

$$
+ \frac{2\nu_3 \Gamma}{n}\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right) \sum_{s \in [t-5q,t+q]\setminus\{t\}} \left[(\mathcal{D}_s)^2 + \left(\Delta_s^X\right)^2\right] + \frac{2\nu_3 \Gamma}{n} \left(\Delta_t^X\right)^2.
$$

23

Claim 4. *[Bounding Term D] If $r < 1$, term D is bounded by:*

$$\frac{2\nu_2\Gamma}{3q}\sum_{\substack{s\in[t-2q,\\t-1]}}\left[(\mathcal{D}_s)^2+\left(\Delta_s^X\right)^2\right]+\frac{4\nu_3\Gamma}{3q}\sum_{\substack{s\in[t-7q,\\t+q]\setminus\{t\}}}\left[(\mathcal{D}_s)^2+\left(\Delta_s^X\right)^2\right]$$

$$+\frac{4\nu_4\Gamma}{3}\left(\frac{\nu_1}{q}+\frac{\nu_2}{q}\right)\sum_{s\in[t-5q,t+q]\setminus\{t\}}\left[(\mathcal{D}_s)^2+\left(\Delta_s^X\right)^2\right]+\frac{4\nu_4\Gamma}{3}\left(\Delta_t^X\right)^2.$$

In Section 4.2.2, to obtain a bound on $\mathbb{E}\left[\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}\right]$, we bound $\overline{\Delta}_{\mathsf{span}}x_{k_t}^{\pi,t}$ in terms of $\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}$ for $s\in[t-5q,t+q]$. In contrast, in the proof of Lemma 11, we start by bounding $\Delta_{\mathsf{var}}x_{k_{l_0}}^{\pi,l_0}$ in terms of various $\Delta_{\mathsf{span}}x_{k_{l_1}}^{\pi,l_1}$; but then, for each l_1, we bound $\Delta_{\mathsf{span}}x_{k_{l_1}}^{\pi,l_1}$ in terms of various $\Delta_{\mathsf{span}}x_{k_{l_2}}^{\pi,l_2}$; we continue recursively until one of the following two cases occurs.
1. When the recursion reaches a term $\Delta_{\mathsf{span}}x_{k_s}^{\pi,s}$ with $s < u - q$, as this term does not depend on $\mathcal{U}_u$, one can safely average over k_u, thereby replacing the multiplier $L_{k_0 k_u}^2$ by $\frac{1}{n}L_{\mathsf{res}}^2$.
2. When the recursion reaches a term $\Delta_{\mathsf{span}}x_{k_u}^{\pi,u}$ it needs to stop. Now to remove the multiplier $L_{k_0 k_u}^2$ we simply upper bound it by $L_{\max}^2$. Unfortunately, there is no averaging and so we "lose" a factor of $\frac{1}{n}$, which could otherwise more than compensate for a growth by a factor of q^2, as in the non-recursive analysis. We save one $\Theta(q)$ factor by using an *unbalanced* Cauchy-Schwartz inequality described in the appendix. (This might not appear to be sufficient, but in fact it is. The reason is that for each time u, this error term originates from u itself. The other error terms originate from $\Theta(q)$ times in a range $u \pm \Theta(q)$, and so are $\Theta(q)$ times as numerous, so in fact we are saving a $\Theta(q^2)$ factor.)

4.2.4 Finalizing the bound on Err_t

We finish Part 2 of the analysis by expressing the bound on $\mathbb{E}[\mathrm{Err}_t]$ in terms of $(\mathcal{D}_s)^2$ and $\left(\Delta_s^X\right)^2$. Recall that $\Lambda^2 = \frac{L_{\mathsf{res}}^2}{\Gamma^2}+1$ and $r = \frac{160q^2\Lambda^2}{n}$, and we make q sufficiently small to ensure that $r < 1$.

Lemma 12.

$$\mathbb{E}\left[\mathrm{Err}_t\right] \le \left(\frac{15r}{1-r}\right)\Gamma\left(\Delta_t^X\right)^2+\varpi\Gamma\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[(\mathcal{D}_s)^2+\left(\Delta_s^X\right)^2\right]$$

$$\text{where } \varpi=\frac{1}{q}\left[\frac{2r}{3}+\frac{3r^2}{1280}+\frac{9r^3}{25600}+\frac{3r^2}{1-r}+\frac{r^3}{426(1-r)}+\frac{r^4}{2844(1-r)}\right].$$

Proof. We begin with the bound in Lemma 5. We have already bounded terms A, D, and E. Terms B and C are bounded as follows.

Claim 5. *[Bounding Term B] Term B is bounded by*

$$\mathbb{E}\left[\frac{2}{3n^2}\sum_{\substack{t-2q\le s<t\\ \&s=\mathrm{prev}(t,k_s)}}\sum_{k_t=1}^{n}\Gamma\cdot\left(x_{k_s}^{\pi(k_s),t}-x_{k_s}^{\pi(k_t),t}\right)^2\right]\le\frac{\nu_1}{15q}\sum_{s\in[t-2q,t-1]}\Gamma\cdot(\mathcal{D}_s)^2.$$

Claim 6. *[Bounding Term C] Term C is bounded by*

$$\mathbb{E}\left[\frac{\Gamma}{2n^2}\sum_{\substack{t-2q\le s<t\\ \&s=\mathrm{prev}(t,k_s)}}\sum_{k_t=1}^{n}\left(\Delta x_{k_s}^{\pi(k_t),s}\right)^2\right]\le\frac{\Gamma}{2n}\sum_{t-2q\le s\le t-1}\left(\Delta_s^X\right)^2.$$

Summing up these bounds yields

$$\mathbb{E}\left[\text{Err}_t\right] \leq \left(\frac{2\nu_3}{n} + \frac{4\nu_4}{3} + \nu_4\right)\Gamma\left(\Delta_t^X\right)^2$$

$$+ \underbrace{\left[\max\left\{\frac{\nu_1}{15q}, \frac{3}{2n}\right\} + \frac{2\nu_2}{3q} + \frac{2(\nu_3 + \nu_4)}{n} + \frac{7\nu_3}{3q} + \left(\frac{2\nu_3}{nq} + \frac{7\nu_4}{3q}\right)(\nu_1 + \nu_2)\right]}_{G}$$

$$\cdot\,\Gamma\sum_{s\in[t-7q,t+q]}\left(\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right). \tag{13}$$

Via some elementary calculations, presented in Claim 7 in Appendix A.5, we show that the quantity G above can be bounded by

$$\frac{1}{q}\left[\frac{2r}{3} + \frac{3r^2}{1280} + \frac{9r^3}{25600} + \frac{3r^2}{1-r} + \frac{r^3}{426(1-r)} + \frac{r^4}{2844(1-r)}\right]$$

and that $\frac{2\nu_3}{n} + \frac{7\nu_4}{3} \leq \frac{15r}{1-r}$ as $1 \leq q < n$ and $r \leq 1$, which implies the bounds stated in the lemma. $\qquad\square$

4.3   Part 3: The Amortization

We let $\varrho = \frac{1}{8} - \frac{15r}{1-r}$. From Lemmas 5 and 12, we obtain:

$$\mathbb{E}\left[F(x^t) - F(x^{t+1})\right] \geq \frac{1}{3n^2}\cdot\mathbb{E}\left[\sum_{k=1}^{n}\sum_{k_t=1}^{n}\widehat{W}_k(g_k^{\pi(k_t,t),t}, x_k^{\pi(k_t,t),t})\right]$$

$$+ \varrho\Gamma\left(\Delta_t^X\right)^2 - \varpi\Gamma\sum_{s\in[t-7q,t+q]}\left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right]. \tag{14}$$

The term $\left(\Delta_s^X\right)^2$ and $\left(\mathcal{D}_s\right)^2$ in (14) will be paid for by the progress terms from time s by means of an amortization. Also, we will account for the term $\left(\mathcal{D}_t\right)^2$ using the bound from Lemma 10:

$$\Gamma\cdot\left(\mathcal{D}_t\right)^2 \leq \left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)\Gamma\sum_{s\in[t-5q,t+q]\setminus\{t\}}\left(\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right). \tag{15}$$

For the purpose of amortizing the $\left(\mathcal{D}_s\right)^2$ terms, for some constant $\gamma > 0$ which we will specify later, we add terms $+\gamma\Gamma\left(\mathcal{D}_t\right)^2 - \gamma\Gamma\left(\mathcal{D}_t\right)^2$ to the bound from (14), and then we use (15) to bound $(\gamma + \varpi)\Gamma\left(\mathcal{D}_t\right)^2$, which yields

$$\mathbb{E}\left[F(x^t) - F(x^{t+1})\right] \geq \frac{1}{3n^2}\cdot\mathbb{E}\left[\sum_{k=1}^{n}\sum_{k_t=1}^{n}\widehat{W}_k(g_k^{\pi(k_t,t),t}, x_k^{\pi(k_t,t),t})\right]$$

$$+ (\varrho - \varpi)\,\Gamma\left(\Delta_t^X\right)^2 + \gamma\Gamma\left(\mathcal{D}_t\right)^2 - \varpi\Gamma\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right]$$

$$- (\varpi + \gamma)\left[\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)\Gamma\sum_{s\in[t-5q,t+q]\setminus\{t\}}\left(\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right)\right]. \tag{16}$$

In the standard convergence analysis for a sequential stochastic coordinate descent, one shows that

$$\mathbb{E}\left[F(x^t) - F(x^{t+1})\right] \geq \frac{1}{3n} \cdot \mathbb{E}\left[\sum_{k=1}^{n} \widehat{W}_k(g_k^{\pi(k_t,t),t}, x_k^{\pi(k_t,t),t})\right]$$

$$= \frac{1}{3n^2} \cdot \mathbb{E}\left[\sum_{k=1}^{n}\sum_{k_t=1}^{n} \widehat{W}_k(g_k^{\pi(k_t,t),t}, x_k^{\pi(k_t,t),t})\right].$$

To obtain an analogous bound, we have to show all the additional terms in (16) make a non-positive contribution over the T steps of the algorithm, analogous to the use of (5) in the basic framework. To this end, we apply the following theorem regarding rates of convergence. This theorem uses amortization terms A^+ and A^-, to define a potential function $H(t)$, which could also be viewed as a Lyapunov function. We note that the same result, but without the amortization terms A^+ and A^-, can be found in [25]. A^+ represents progress that has occurred, but which is being saved to pay for future errors; A^-, in contrast, represents the effect of errors in the past, which will be paid for by future progress.

Theorem 3. *Suppose that $\Gamma \geq L_{\max}$. Let q be a fixed integer parameter. Let $A^+(t)$, $A^-(t)$ be non-negative functions with $A^+(1) = 0$, $A^-(T+1) = 0$, and let $H(t) := \mathbb{E}\left[F(x^t)\right] + A^+(t) - A^-(t)$. Suppose that*

a. $H(t) \geq 0$ for all $t \geq 1$;

b. for all $t \geq 1$, $H(t+1) \leq H(t)$, i.e., $H(t)$ is a decreasing function of t;

c. there exist constants $\alpha, \beta > 0$ such that for any $t \geq 1$,

$$H(t) - H(t+1) \geq \frac{\alpha}{n} \mathbb{E}\left[\sum_{k=1}^{n} \widehat{W}_k(\nabla_k f(x^t), x_k^t)\right] + \frac{\beta}{n} \cdot A^+(t).$$

(i) If F is strongly convex with parameter μ_F, and f has strongly convex parameter μ_f, then for all $T \geq 0$,

$$\mathbb{E}\left[F(x^{T+1})\right] \leq H(T+1) \leq \left[1 - \min\left\{\frac{\alpha}{n} \cdot \frac{\mu_F}{\mu_F + \Gamma - \mu_f} , \frac{\beta}{n}\right\}\right]^T \cdot F(x^1).$$

(ii) Now suppose that F is convex. Let $\mathcal{R}$ be the radius of the level set for x^1. Formally, let $X = \{x \mid F(x) \leq F(x^1)\}$; then $\mathcal{R} = \sup_{x \in X} \inf_{x^* \in X^*} \|x - x^*\|$. Then, for all $T \geq 0$,

$$\mathbb{E}\left[F(x^{T+1})\right] \leq H(T+1) \leq \frac{F(x^1)}{1 + \min\left\{\frac{\beta}{2n \cdot F(x^1)}, \frac{\alpha}{4n \cdot F(x^1)}, \frac{\alpha}{8n\Gamma\mathcal{R}^2}\right\} \cdot F(x^1) \cdot T}.$$

We will be applying Theorem 3 with $\alpha = \beta = \frac{1}{3n}$.

In order to obtain condition (c) of Thereom 3 from (16), it suffices to show

$$\left[\left(1 - \frac{1}{3n}\right) A^+(t) - A^-(t)\right] - \left[A^+(t+1) - A^-(t+1)\right]$$

$$\geq -(\varrho - \varpi)\Gamma\left(\Delta_t^X\right)^2 - \gamma\Gamma(\mathcal{D}_t)^2 + \varpi\Gamma \sum_{s \in [t-7q, t+q]\setminus\{t\}} \left[(\mathcal{D}_s)^2 + (\Delta_s^X)^2\right]$$

$$+ (\varpi + \gamma)\left[\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)\Gamma \sum_{s \in [t-5q, t+q]\setminus\{t\}} \left((\mathcal{D}_s)^2 + (\Delta_s^X)^2\right)\right]. \tag{17}$$

Lemma 13. *Inequality* (17) *holds if* $7q < n$, $c = \varpi + (\gamma + \varpi)\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)\Gamma$, $\gamma = \varrho - \varpi$, $\Lambda = \frac{L_{\text{res}}^2}{\Gamma^2} + 1$, $r = \frac{160q^2\Lambda}{n} \leq \frac{1}{225}$, *and*

$$A^+(t) = \sum_{s=t-7q}^{t-1} \sum_{v=t}^{s+7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t+1}} \left[c(\mathcal{D}_s)^2 + c\left(\Delta_s^X\right)^2\right],$$

$$A^-(t) = \sum_{s=t-q}^{t-1} \sum_{v=t}^{s+q} \left[c(\mathcal{D}_v)^2 + c\left(\Delta_v^X\right)^2\right].$$

We note that as $\mathcal{D}_t = 0$ and $\Delta_t^X = 0$ for $t = 0$ and for $t \geq T + 1$, with the above definition, $A^+(1) = 0$ and $A^-(T+1) = 0$.

We are now ready to conclude the proof of our main result.

Proof of Theorem 2. We set $\alpha = \beta = \frac{1}{3}$.

By Lemma 13, if $r \leq \frac{1}{225}$, the conditions for applying Theorem 3 hold: (c) holds by construction; this implies that (b) holds as the RHS of (c) is non-negative; finally, as $A^-(T+1) = 0$, $H(T+1) \geq 0$, and together with (b) this implies (a). As $A^-(T+1) = 0$, we conclude that $\mathbb{E}\left[F(x^{T+1})\right] \leq H(T+1)$; this inequality also holds in expectation, thus we are done.

We now apply Theorem 3, which yields the stated results. Recall that $r = \dfrac{160q^2\left(\frac{L_{\text{res}}^2}{\Gamma^2} + 1\right)}{n}$. Thus, to achieve $r \leq \frac{1}{225}$ it suffices to have $q \leq \min\left\{\frac{\Gamma\sqrt{n}}{270L_{\text{res}}}, \frac{\sqrt{n}}{270}\right\}$. $\qquad\square$

Note that we have not sought to fully optimize the constants.

Acknowledgment

References

[1] Haim Avron, Alex Druinsky, and Anshul Gupta. Revisiting asynchronous linear solvers: Provable convergence rate through randomization. *J. ACM*, 62(6):51:1–51:27, 2015.

[2] Gérard M. Baudet. Asynchronous iterative methods for multiprocessors. *J. ACM*, 25(2):226–244, 1978.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation:Numerical Methods*. Prentice Hall, 1989.

[4] Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optimization*, 10(3):627–642, 2000.

[5] Vivek S. Borkar. Asynchronous stochastic approximations. *SIAM J. Control and Optimization*, 36(3):662–663, 1998.

[6] D. Chazan and W. Miranker. Chaotic relaxation. *Linear Agebra Appl.*, 2(2):199–222, 1969.

[7] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman's function. *SIAM J. Optimization*, 3(3):538–543, 1993.

[8] Yun Kuen Cheung and Richard Cole. Amortized analysis of asynchronous price dynamics. In *26th Annual European Symposium on Algorithms, ESA 2018, August 20-22, 2018, Helsinki, Finland*, pages 18:1–18:15, 2018.

[9] Yun Kuen Cheung, Richard Cole, and Nikhil R. Devanur. Tatonnement beyond gross substitutes? gradient descent to the rescue. In *STOC*, pages 191–200, 2013.

[10] Yun Kuen Cheung, Richard Cole, and Yixin Tao. Parallel Stochastic Asynchronous Coordinate Descent: Tight Bounds on the Possible Parallelism. *arXiv e-prints*, page arXiv:1811.05087, November 2018.

[11] Richard Cole and Yixin Tao. An Analysis of Asynchronous Stochastic Accelerated Coordinate Descent. *arXiv e-prints*, page arXiv:1808.05156, Aug 2018.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[13] Cong Fang, Yameng Huang, and Zhouchen Lin. Accelerating asynchronous algorithms for convex optimization by momentum compensation. *arXiv preprint arXiv:1802.09747*, 2018.

[14] Andreas Frommer and Daniel B. Szyld. On asynchronous iterations. *Journal of Computational and Applied Mathematics*, 123(1-2):201–216, 2000. Numerical Analysis 2000. Vol. III: Linear Algebra.

[15] Robert Hannah, Fei Feng, and Wotao Yin. A2bcd: An asynchronous accelerated block coordinate descent algorithm with optimal complexity. 2018.

[16] Rémi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *J. Mach. Learn. Res.*, 19:81:1–81:68, 2018.

[17] Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

[18] Ji Liu, Stephen J. Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16:285–322, 2015.

[19] Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

[20] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

[21] Lukas Meier, Sara Van De Geer, and Peter Bhlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[22] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer US, 2004.

[23] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optimization*, 22(2):341–362, 2012.

[24] Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pages 693–701, 2011.

[25] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[26] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[27] Tao Sun, Robert Hannah, and Wotao Yin. Asynchronous coordinate descent under more realistic assumptions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6182–6190. Curran Associates, Inc., 2017.

[28] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[29] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117(1-2):387–423, 2009.

[30] John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[31] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

A   Omitted Proofs and Subsidiary Lemmas

We begin with the proof of Lemma 1. Next, in Appendix A.1, we prove Lemmas 2 and 3, the basic progress lemmas. In Appendix A.2, we prove Lemma 5, the progress lemma for the general analysis. Then, in Appendix A.3, we give several bounds on how much $\widehat{W}$ can change when one of its arguments is altered, leading to proofs of Lemmas 6–8. We follow this, in Appendix A.4, with the proof of the recursive bound given in Lemma 21, which is used to show Lemma 2. We continue, in Appendix A.5, with the proofs of the claims from Section 4.2.4. Finally, in Appendix A.6, we prove Theorem 3 and Lemma 13.

Proof of Lemma 1. Let $\mathcal{U}_b$ be an update to coordinate x_j with start and finish times s_b and f_b, resp. Let $\mathcal{U}_a$ be the update to coordinate x_j with the earliest start time before s_b that commits later than $\mathcal{U}_b$, if any. Let $\mathcal{U}_c$ be the update to coordinate x_j with the latest start time after s_b that commits earlier than $\mathcal{U}_b$, if any. Let the start and finish times for $\mathcal{U}_a$ and $\mathcal{U}_c$ be (s_b, f_b), and (s_c, f_c), resp. Note that the start and commit times of an update differ by at most $q+1$ as there are at most q interfering updates for each update.

Suppose the SCC order for $\mathcal{U}_b$ is s. If $\mathcal{U}_a$ does not exist then $s \geq s_b$; it is convenient to set $(s_a, f_a) = (s_b, f_b)$ in this case. Similarly, If $\mathcal{U}_c$ does not exist then $s \leq s_b$; it is convenient to set $(s_c, f_c) = (s_b, f_b)$ in this case. Then it is always the case that $s_a \leq s \leq s_c$.

Case i. $s < s_b$.
Then $s_a \leq s < s_b < f_b < f_a \leq s_a + q + 1$. It follows that $s_b \leq s_a + q - 1$ and hence $s \geq s_b - q + 1$. Also $f_b \leq s_b + q + 1 \leq s_a + q - 1 \leq s + q - 1$.

Case ii. $s > s_b$.
Then $s_b < s \leq s_c < f_c < f_b \leq s_b + q + 1$. It follows that $s \leq s_b + q - 1$ and $f_b \leq s + q$.

Case iii. $s = s_b$.
Then $f_b \leq s + q + 1$.

Therefore, for any updates, their SCC order s and their commit time f_b satisfy $s < f_b \leq s + q + 1$. Now we determine the SCC range for updates that might interfere $\mathcal{U}_b$. Note that they have commit time in the range $[s_b + 1, f_b - 1]$.

Case i. $s < s_b$.
The commit time for possibly interfering updates is in the range $[s_b + 1, f_b - 1] \subseteq [s, s + q - 2]$ and hence their SCC rank is in the range $[s - (q + 1), s + q - 3] = [s - q - 1, s + q - 3]$.

Case ii. $s > s_b$.
The commit time for possibly interfering updates is in the range $[s_b + 1, f_b - 1] \subseteq [s - q + 2, s + q]$ and hence their SCC rank is in the range $[s - q + 2 - (q + 1), s + q - 1] = [s - 2q + 1, s + q - 1]$.

Case iii. $s = s_b$.
The commit time for possibly interfering updates is in the range $[s_b + 1, f_b - 1] \subseteq [s + 1, s + q]$ and hence their SCC rank is in the range $[s + 1 - (q + 1), s + q - 1] = [s - q, s + q - 1]$.

For all $q \geq 1$, this range is contained in $[s - 2q + 1, s + q - 1]$. □

A.1   The Basic Progress Lemmas, Lemmas 2 and 3

We recall two known results.

Lemma 14 (Three-Point Property, [7, Lemma 3.2]). *For any proper, convex and lower semi-continuous function* $Y : \mathbb{R} \to \mathbb{R}$ *and for any* $d^- \in \mathbb{R}$, *let* $d^+ := \arg\max_{d \in \mathbb{R}} \{-Y(d) - \Gamma(d - d^-)^2\}$. *Then for any* $d' \in \mathbb{R}$,

$$Y(d') + \Gamma(d' - d^-)^2 \;\geq\; Y(d^+) + \Gamma(d^+ - d^-)^2 + \Gamma(d' - d^+)^2.$$

Lemma 15 ([29, Lemma 4]). *For any $g_1, g_2, x \in \mathbb{R}$ and $\Gamma \in \mathbb{R}^+$,*
$$\left| \widehat{d}(g_1, x) - \widehat{d}(g_2, x) \right| \leq \tfrac{1}{\Gamma} \cdot |g_1 - g_2| \,.$$

We can now lower bound $\widehat{W}_j(g, x)$ in terms of $\widehat{d}_j(g, x)$.

Lemma 16. *For any $g, x \in \mathbb{R}$ and $\Gamma \in \mathbb{R}^+$, $\widehat{W}_j(g, x) \geq \tfrac{\Gamma}{2} \left(\widehat{d}_j(g, x) \right)^2 .$*

Proof. We apply Lemma 14 with $d^- = d' = 0$, with Γ replaced by $\tfrac{1}{2}\Gamma$, and $Y(d) = gd - \Psi(x) + \Psi(x + d)$. Then $W_j(d, g, x) = -Y(d) - \Gamma d^2/2$, and hence d^+, as defined in Lemma 14, equals $\widehat{d}_j(g, x)$. These yield

$$Y(0) \geq Y(\widehat{d}_j(g, x)) + \Gamma \cdot \left(\widehat{d}_j(g, x) \right)^2 .$$

Since $Y(0) = 0$ and $-Y(\widehat{d}_j(g, x)) = \widehat{W}_j(g, x) + \tfrac{\Gamma}{2} \left(\widehat{d}_j(g, x) \right)^2$, we are done. $\qquad\square$

We are now ready to show Lemmas 2 and 3; they follow directly from Lemma 17 below. We will use the following well-known observation: for any $1 \leq k \leq n$, $x \in \mathbb{R}^n$ and $r \in \mathbb{R}$,

$$f(x + r \cdot e_k) \leq f(x) + \nabla_k f(x) \cdot r + \frac{L_k}{2} \cdot r^2, \tag{18}$$

where e_k is unit vector along coordinate k.

Lemma 17. *Suppose there is an update to coordinate j at time t according to rule (1), and suppose that $\Gamma \geq L_{\max}$. Let $g_j = \nabla_j f(x^t)$ and $\tilde{g}_j = \nabla_j f(\tilde{x})$. Then*

$$F(x^t) - F(x^{t+1}) \geq \frac{\Gamma}{4}(\widehat{d}_j(\tilde{g}_j, x_j^t))^2 - \frac{1}{\Gamma} \cdot (g_j - \tilde{g}_j)^2$$

and
$$F(x^t) - F(x^{t+1}) \geq \widehat{W}_j(g_j, x_j^t) - \frac{1}{\Gamma} \cdot (g_j - \tilde{g}_j)^2 .$$

Proof. To avoid clutter, we use the shorthand $d_j := \widehat{d}_j(g_j, x_j^t)$ and $\tilde{d}_j := \widehat{d}_j(\tilde{g}_j, x_j^t)$. By update rule (1), $\tilde{d}_j = \Delta x_j^t$.

$$F(x^{t+1}) = f(x^{t+1}) + \Psi_j(x_j^{t+1}) + \sum_{k \neq j} \Psi_k(x_k^{t+1})$$

$$\leq f(x^t) + g_j \tilde{d}_j + \frac{\Gamma}{2}(\tilde{d}_j)^2 + \Psi_j(x_j^t + \tilde{d}_j) + \sum_{k \neq j} \Psi_k(x_k^t)$$

$$\text{(By (18), (1), and the assumption } \Gamma \geq L_{\max} \geq L_j)$$

$$= F(x^t) + \tilde{g}_j \tilde{d}_j + \frac{\Gamma}{2}(\tilde{d}_j)^2 - \Psi_j(x_j^t) + \Psi_j(x_j^t + \tilde{d}_j) + (g_j - \tilde{g}_j)\tilde{d}_j$$

$$= F(x^t) - \widehat{W}_j(\tilde{g}_j, x_j^t) + (g_j - \tilde{g}_j)\tilde{d}_j .$$

Hence, $F(x^t) - F(x^{t+1}) \geq \widehat{W}_j(\tilde{g}_j, x_j^t) - (g_j - \tilde{g}_j)\tilde{d}_j .$

31

Then we can apply Lemma 16 to prove the first inequality in Lemma 17:

$$
\begin{aligned}
F(x^t) - F(x^{t+1}) &\geq \widehat{W}_j(\tilde{g}_j, x_j^t) - (g_j - \tilde{g}_j)\tilde{d}_j \geq \frac{\Gamma}{2}(\tilde{d}_j)^2 - |g_j - \tilde{g}_j| \cdot |\tilde{d}_j| \\
&\geq \frac{\Gamma}{2}(\tilde{d}_j)^2 - \frac{1}{2}\left[\frac{2}{\Gamma} \cdot (g_j - \tilde{g}_j)^2 + \frac{\Gamma}{2}(\tilde{d}_j)^2\right] \qquad \text{(by the AM-GM ineq.)} \\
&= \frac{\Gamma}{4}(\tilde{d}_j)^2 - \frac{1}{\Gamma} \cdot (g_j - \tilde{g}_j)^2.
\end{aligned}
$$

We prove the second inequality in Lemma 17 as follows:

$$
\begin{aligned}
F(x^t) - F(x^{t+1}) &\geq \widehat{W}_j(\tilde{g}_j, x_j^t) - (g_j - \tilde{g}_j)\tilde{d}_j \geq W_j(d_j, \tilde{g}_j, x_j^t) - (g_j - \tilde{g}_j)\tilde{d}_j \\
&= W_j(d_j, g_j, x_j^t) + (g_j - \tilde{g}_j)d_j - (g_j - \tilde{g}_j)\tilde{d}_j \\
&= \widehat{W}_j(g_j, x_j^t) + (g_j - \tilde{g}_j)(d_j - \tilde{d}_j) \geq \widehat{W}_j(g_j, x_j^t) - |g_j - \tilde{g}_j| \cdot |d_j - \tilde{d}_j| \\
&\geq \widehat{W}_j(g_j, x_j^t) - \frac{1}{\Gamma}(g_j - \tilde{g}_j)^2. \qquad \text{(By Lemma 15.)}
\end{aligned}
$$

$\square$

A.2   The Expected Progress, Lemma 5

Proof of Lemma 5. Recall that we write $\pi(k,t)$ to denote the path in which coordinate k_t at time t is replaced by coordinate k, and to reduce clutter we abbreviate this as $\pi(k)$. Recall also that we let $\mathrm{prev}(t, k)$ denote the time of the most recent update to coordinate k, if any, in the time range $[t - 2q, t - 1]$; otherwise, we set it to t. From (3),

$$
\begin{aligned}
&\mathbb{E}\left[F(x^t) - F(x^{t+1})\right] \\
&\geq \frac{1}{2n}\mathbb{E}_\pi\left[\sum_{k_t=1}^n \widehat{W}_{k_t}(g_{k_t}^{\pi(k_t),t}, x_{k_t}^{\pi(k_t),t})\right] + \frac{\Gamma}{8}\left(\Delta_t^X\right)^2 - \frac{1}{\Gamma}\mathbb{E}\left[\left(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t}\right)^2\right] \\
&= \frac{1}{2n}\mathbb{E}_\pi\left[\sum_{k=1}^n \widehat{W}_k(g_k^{\pi(k),t}, x_k^{\pi(k),t})\right] + \frac{\Gamma}{8}\left(\Delta_t^X\right)^2 - \frac{1}{\Gamma}\mathbb{E}\left[\left(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t}\right)^2\right] \\
&\geq \frac{1}{2n}\mathbb{E}_\pi\left[\sum_{k=1}^n \frac{1}{n}\sum_{k_t=1}^n \left(\frac{2}{3}\widehat{W}_k(g_k^{\pi(k_t),t}, x_k^{\pi(k),t}) - \frac{4}{3\Gamma}\left(g_k^{\pi(k),t} - g_k^{\pi(k_t),t}\right)^2\right)\right] \\
&\quad + \frac{\Gamma}{8}\left(\Delta_t^X\right)^2 - \frac{1}{\Gamma}\mathbb{E}_\pi\left[\left(g_{k_t}^{\pi,t} - \tilde{g}_{k_t}^{\pi,t}\right)^2\right] \qquad \text{(by Lemma 6).}
\end{aligned}
$$

For the next bound, we will be applying Lemma 7 to shift the $x_k^{\pi(k),t}$ parameter in $\widehat{W}_k$ to $x_k^{\pi,t} = x_k^{\pi(k_t),t}$. Note that applying this lemma introduces additional terms (the case $l > 0$) only if x_k is updated at some time $s \in [t - 2q, t - 1]$; this means that $k = k_s$ where $t - 2q \leq s < t$, and to avoid double counting the effect of updates to the same coordinate, we can further limit s to $s = \mathrm{prev}(t, k)$, or equivalently that $t - 2q \leq s < t$ and $s = \mathrm{prev}(t, k_s)$. This yields the claimed result. $\square$

A.3   Bounding How Much $\widehat{W}$ and $\widehat{d}$ Vary as a Function of Their Arguments

The next five lemmas concern an arbitrary coordinate x_j. To avoid notational clutter, we write W, $\widehat{W}$, $\widehat{d}$, and Ψ in lieu of W_j $\widehat{W}_j$, $\widehat{d}$, and Ψ_j, resp. Also, by x_1 and x_2 we will mean two possible values of x_j, and by g_1 and g_2 two possible values of g_j.

We first present the proofs of Lemma 6 and 8. To prove Lemma 7, we will need two additional lemmas, to be presented below.

Proof of Lemma 6.

$$
\begin{aligned}
\widehat{W}(g_1, x) &= \max_{d \in \mathbb{R}} W(d, g_1, x) \geq W(\widehat{d}(g_2), g_1, x) \\
&= -g_1 \cdot \widehat{d}(g_2) - \Gamma \cdot \widehat{d}(g_2)^2/2 + \Psi(x) - \Psi(x + \widehat{d}(g_2)) \\
&= -g_2 \cdot \widehat{d}(g_2) - \Gamma \cdot \widehat{d}(g_2)^2/2 + \Psi(x) - \Psi(x + \widehat{d}(g_2)) \\
&\qquad\qquad + (g_2 - g_1) \cdot \left[\widehat{d}(g_1) + (\widehat{d}(g_2) - \widehat{d}(g_1))\right] \\
&\geq \widehat{W}(g_2, x) - |g_1 - g_2| \cdot \left|\widehat{d}(g_1)\right| - |g_1 - g_2| \cdot \left|\widehat{d}(g_2) - \widehat{d}(g_1)\right| \\
&\geq \widehat{W}(g_2, x) - |g_1 - g_2| \cdot \left|\widehat{d}(g_1)\right| - \frac{1}{\Gamma}(g_1 - g_2)^2 \qquad \text{(By Lemma 15)} \\
&\geq \widehat{W}(g_2, x) - \frac{1}{\Gamma}(g_1 - g_2)^2 - \frac{\Gamma}{4}(\widehat{d}(g_1))^2 - \frac{1}{\Gamma}(g_1 - g_2)^2 \qquad \text{(AM-GM ineq.)} \\
&\geq \widehat{W}(g_2, x) - \frac{2}{\Gamma}(g_1 - g_2)^2 - \frac{1}{2}\widehat{W}(g_1, x). \qquad \text{(By Lemma 16)}
\end{aligned}
$$

$\square$

Next, we demonstrate Lemma 8; it is a simple corollary of Lemma 15 and the following lemma.

Lemma 18. *For any $g, x_1, x_2 \in \mathbb{R}$, $\left|\widehat{d}(g, x_1) - \widehat{d}(g, x_2)\right| \leq |x_1 - x_2|$.*

Proof. For $i = 1, 2$, let $d_i := \widehat{d}(g, x_i)$. By the definition of $\widehat{d}$, for $i = 1, 2$, there exists a subgradient $\Psi'(x_i + d_i)$ such that
$$
g + \Gamma \cdot d_i + \Psi'(x_i + d_i) = 0.
$$

If $d_1 = d_2$, we are done. If $d_1 > d_2$, then $\Psi'(x_1 + d_1) < \Psi'(x_2 + d_2)$. Since Ψ is convex, $x_1 + d_1 \leq x_2 + d_2$ and hence $0 < d_1 - d_2 \leq x_2 - x_1$.

If $d_2 > d_1$, by the same argument as above we have $0 < d_2 - d_1 \leq x_1 - x_2$. $\square$

Lemma 8 is a simple corollary of Lemmas 15 and 18.

Proof of Lemma 8.

$$
\begin{aligned}
\left(\widehat{d}(g_1, x_1) - \widehat{d}(g_2, x_2)\right)^2 &= \left(\widehat{d}(g_1, x_1) - \widehat{d}(g_1, x_2) + \widehat{d}(g_1, x_2) - \widehat{d}(g_2, x_2)\right)^2 \\
&\leq 2\left(\widehat{d}(g_1, x_1) - \widehat{d}(g_1, x_2)\right)^2 + 2\left(\widehat{d}(g_1, x_2) - \widehat{d}(g_2, x_2)\right)^2 \\
&\leq 2(x_1 - x_2)^2 + \frac{2}{\Gamma^2}(g_1 - g_2)^2.
\end{aligned}
$$

$\square$

The next two lemmas will be needed to prove Lemma 7.

Lemma 19 ($\widehat{W}$ Shifting on x parameter). *Let $\widehat{W}(g, x_1) = W(\check{d}_1, g, x_1)$ and $\widehat{W}(g, x_2) = W(\check{d}_2, g, x_2)$. Then*

$$
\widehat{W}(g, x_1) + \Psi(x_2) - \Psi(x_1) \geq \widehat{W}(g, x_2) - g(x_2 - x_1) - \Gamma\check{d}_2(x_2 - x_1) - \frac{\Gamma}{2} \cdot (x_2 - x_1)^2.
$$

Proof. We use Lemma 14 with $d^- = 0$, $d^+ = \check{d}_1$, and $Y(d) = gd - \Psi(x_1) + \Psi(x_1 + d)$. We note that $Y(d') + \frac{\Gamma}{2} \cdot d'^2 = -W(d', g, x_1)$. Also, we observe that $-W(d', g, x_1)$ is strongly convex with strong convexity parameter Γ. As $W(d', g, x_1)$ is maximized at $\check{d}_1$, we conclude that $-W(d', g, x_1) \geq -W(\check{d}_1, g, x_1) + \frac{\Gamma}{2}(\check{d}_1 - d')^2 = -\widehat{W}(g, x_1) + \frac{\Gamma}{2}(\check{d}_1 - d')^2$. Thus

$$Y(d') + \frac{\Gamma}{2} \cdot (d')^2 \geq -\widehat{W}(g, x_1) + \frac{\Gamma}{2} \cdot (d' - \check{d}_1)^2.$$

The above inequality holds for any d'. In particular, we pick $d' = x_2 - x_1 + \check{d}_2$, yielding

$$\widehat{W}(g, x_1) \geq -g(x_2 - x_1 + \check{d}_2) + \Psi(x_1) - \Psi(x_2 + \check{d}_2) - \frac{\Gamma}{2} \cdot (x_2 - x_1 + \check{d}_2)^2$$
$$+ \frac{\Gamma}{2} \cdot (x_2 - x_1 + \check{d}_2 - \check{d}_1)^2.$$

By adding $\Psi(x_2) - \Psi(x_1)$ to both sides, we obtain

$$\widehat{W}(g, x_1) + \Psi(x_2) - \Psi(x_1)$$
$$\geq -g(x_2 - x_1 + \check{d}_2) + \Psi(x_2) - \Psi(x_2 + \check{d}_2) - \frac{\Gamma}{2} \cdot (x_2 - x_1 + \check{d}_2)^2$$
$$+ \frac{\Gamma}{2} \cdot (x_2 - x_1 + \check{d}_2 - \check{d}_1)^2$$
$$= \widehat{W}(g, x_2) - g(x_2 - x_1) - \Gamma \check{d}_2(x_2 - x_1) - \frac{\Gamma}{2} \cdot (x_2 - x_1)^2 + \frac{\Gamma}{2} \cdot (x_2 - x_1 + \check{d}_2 - \check{d}_1)^2$$
$$\geq \widehat{W}(g, x_2) - g(x_2 - x_1) - \Gamma \check{d}_2(x_2 - x_1) - \frac{\Gamma}{2} \cdot (x_2 - x_1)^2.$$

$\square$

Lemma 20 (Ψ Shifting)**.** *Let* $\widehat{W}(g_1, x_1) = W(\widehat{d}_1, g_1, x_1)$ *and* $\widehat{W}(g_2, x_2) = W(\widehat{d}_2, g_2, x_2)$*. Then*

$$\Psi(x_2 + \widehat{d}_2) - \Psi(x_1 + \widehat{d}_1) \leq g_2(x_1 - x_2 + \widehat{d}_1 - \widehat{d}_2) + \frac{\Gamma}{2} \cdot (x_1 - x_2 + \widehat{d}_1)^2.$$

Proof. By the definition of $\widehat{d}_2$, we have the following inequality, which directly implies the one stated in the lemma.

$$-g_2 \widehat{d}_2 - \frac{\Gamma}{2} \cdot (\widehat{d}_2)^2 - \Psi(x_2 + \widehat{d}_2) \geq -g_2(x_1 - x_2 + \widehat{d}_1) - \frac{\Gamma}{2} \cdot (x_1 - x_2 + \widehat{d}_1)^2 - \Psi(x_1 + \widehat{d}_1).$$

$\square$

Proof. of Lemma 7. Suppose the latest update to coordinate k occurred at time $\check{t}$. Also suppose that

- the changes to x_k from x_k^{t-2q} to $x_k^{\pi(k),t}$ are $d_{11}, d_{12}, \cdots, d_{1\ell}$;

- the changes to x_k from x_k^{t-2q} to $x_k^{\pi,t}$ are $d_{21}, d_{22}, \cdots, d_{2\ell}$.

Furthermore, let

$$g_k^a := \nabla_k f(x^{\pi,t}) \qquad \text{and} \qquad \check{d} := \arg\max_d W(d, g_k^a, x_k^{\pi,t}).$$

In other words, $x_k^{\pi(k),t} = x_k^{t-2q} + \sum_{r=1}^{\ell} d_{1r}$ and $x_k^{\pi,t} = x_k^{t-2q} + \sum_{r=1}^{\ell} d_{2r}$.

By Lemma 19,

$$\widehat{W}(g_k^a, x_k^{\pi(k),t}) + \Psi(x_k^{\pi,t}) - \Psi(x_k^{\pi(k),t}) \geq \widehat{W}(g_k^a, x_k^{\pi,t}) - g_k^a \cdot (x_k^{\pi,t} - x_k^{\pi(k),t})$$
$$- \Gamma \breve{d} \cdot (x_k^{\pi,t} - x_k^{\pi(k),t}) \ - \ \frac{\Gamma}{2} \cdot (x_k^{\pi,t} - x_k^{\pi(k),t})^2.$$

On the other hand, let g_k^b be the gradient used to compute the update $d_{2\ell}$. By Lemma 20, on setting $x_2 = x_k^{\pi,t} - d_{2\ell}$ and $x_1 = x_k^{\pi(k),t} - d_{1\ell}$, and noting that $\widehat{d}_1 = d_{1\ell}$ and $\widehat{d}_2 = d_{2\ell}$, we obtain

$$\Psi(x_k^{\pi,t}) - \Psi(x_k^{\pi(k),t}) \leq g_k^b(x_k^{\pi(k),t} - x_k^{\pi,t}) + \frac{\Gamma}{2}(x_k^{\pi(k),t} - x_k^{\pi,t} + d_{2\ell})^2.$$

Combining the above two inequalities, and letting $\delta := x_k^{\pi,t} - x_k^{\pi(k),t}$, yields

$$\widehat{W}(g_k^a, x_k^{\pi(k),t}) \geq \widehat{W}(g_k^a, x_k^{\pi,t}) + (g_k^b - g_k^a) \cdot \delta - \Gamma \breve{d} \cdot \delta - \frac{\Gamma}{2} \cdot \delta^2 - \frac{\Gamma}{2} \cdot (d_{2\ell} - \delta)^2$$

$$\geq \widehat{W}(g_k^a, x_k^{\pi,t}) - \frac{1}{2\Gamma}(g_k^b - g_k^a)^2 - \frac{\Gamma}{2}\delta^2 - \Gamma(\breve{d} - \delta)\delta - \Gamma\delta^2 - \frac{\Gamma}{2}\delta^2$$

$$- \frac{\Gamma}{2}(d_{2\ell})^2 + \frac{\Gamma}{2} \cdot 2d_{2\ell} \cdot \delta - \frac{\Gamma}{2}\delta^2$$

$$= \widehat{W}(g_k^a, x_k^{\pi,t}) - \frac{1}{2\Gamma}(g_k^b - g_k^a)^2 - \frac{3}{2}\Gamma\delta^2 - \frac{\Gamma}{2}(d_{2\ell})^2 - \Gamma(\breve{d} - \delta)\delta + \Gamma(d_{2\ell} - \delta)\delta$$

$$\geq \widehat{W}(g_k^a, x_k^{\pi,t}) - \frac{1}{2\Gamma}(g_k^b - g_k^a)^2 - \frac{3}{2}\Gamma\delta^2 - \frac{\Gamma}{2}(d_{2\ell})^2 - \Gamma|\breve{d} - d_{2\ell}| \cdot |\delta|$$

$$\geq \widehat{W}(g_k^a, x_k^{\pi,t}) - \frac{1}{2\Gamma}(g_k^b - g_k^a)^2 - 2\Gamma\delta^2 - \frac{\Gamma}{2}(d_{2\ell})^2 - \frac{\Gamma}{2}(\breve{d} - d_{2\ell})^2.$$

By Lemma 8, $\quad \frac{\Gamma}{2} \cdot (\breve{d} - d_{2\ell})^2 \leq \Gamma \cdot (d_{2\ell})^2 + \frac{1}{\Gamma} \cdot (g_k^a - g_k^b)^2.$

Thus, $\quad \widehat{W}(g_k^a, x_k^{\pi(k),t}) \geq \widehat{W}(g_k^a, x_k^{\pi,t}) - \frac{3}{2\Gamma} \cdot (g_k^b - g_k^a)^2 - 2\Gamma\delta^2 - \frac{3\Gamma}{2} \cdot (d_{2\ell})^2.$ (19)

$\square$

A.4   The Recursive Analysis yielding a proof of Lemma 11

Recall that the definition of $\Delta_{\max}^{u,\emptyset} x_{k_s}^{\pi,s}$ (see Section 4.2.1) assumes the first $u - 4q$ updates are already fixed.

To prove Lemma 11, we will make use of a recursive bound on

$$\left(\sum_{l_0 \in [u-q,t+q] \setminus \{u\}} L_{k_{l_0},k_u} \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right) \right)^2,$$

a bound that expresses the effect of performing one level of recursion, along with additional computation — suitable overestimates, and the calculation of some expectations. This bound is presented in the next lemma. It is applied repeatedly in order to prove Lemma 11.

Both lemmas make use of two techniques. It will be helpful to explain them upfront.

Unbalanced Cauchy-Schwartz Inequality To bound an expression of the form $\left(\sum_{1 \leq j \leq q'} z_j \right)^2$, rather than have the bound $q' \sum_{1 \leq j \leq q'} z_j^2$, we would like to multiply a few of the z_j

by just a constant. We achieve this with two applications of the usual Cauchy-Schwartz Inequality, and we illustrate this for the case that "a few" means two of the z_j.

$$\Big(\sum_{1 \le j \le q'} z_j \Big)^2 \le 2(z_k + z_l)^2 + 2\Big(\sum_{\substack{1 \le j \le q' \\ j \ne k,l}} z_j \Big)^2 \le 4(z_k^2 + z_l^2) + 2(q'-2) \sum_{\substack{1 \le j \le q' \\ j \ne k,l}} z_j^2$$

$$\le 4(z_k^2 + z_l^2) + 2q' \sum_{\substack{1 \le j \le q' \\ j \ne k,l}} z_j^2 \quad \text{(this is just a convenient over-estimate)}.$$

Recentering Suppose $a \le b, c \le d$. Then $b^2 \le [c + (d-a)]^2 \le 2c^2 + 2(d-a)^2$. we use this bound when b denotes a value (of the form Δx_k) which varies over the different paths across which we want to average, while the values c and $d - a$ are unvarying. This is a technique we already used in Section 4.2.2.

Lemma 21 is a bound on a sum of Δ_{var} terms. However, we will start by obtaining a recursive bound on an analogous sum of Δ_{span} terms. To this end, we define

$$\mathcal{V}_m \triangleq \mathbb{E}\Bigg[\sum_{l_0 \in [u-q, t+q]\setminus\{u\}} \Bigg(\sum_{\substack{l_1, l_2, \cdots, l_m \in [u-q, t+q], \\ \text{all } l_s \text{ distinct, all } l_s \ne u, l_0; \\ \text{all } l_s \le t_{s-1}+q, \text{ where} \\ t_{s-1} = \min\{t, l_0, l_1, \cdots, l_{s-1}\}; \\ S_m \subseteq \{l_s | k_{l_s} = k_{l_{s-1}}\}.}} \Bigg(\prod_{\substack{l_s \in R_m \setminus S_m \\ \text{where } R_m \\ = \{l_1, l_2, \cdots, l_m\}.}} \frac{L^2_{k_{l_s}, k_{l_{s-1}}}}{\Gamma^2} \Bigg)$$

$$\cdot L^2_{k_{l_0}, k_u} \Big(\Delta_{\mathsf{span}}^{t_m, R_m \setminus \{l_m\}} x^{\pi, l_m}_{k_{l_m}} \Big)^2 \Bigg) \Bigg]. \tag{20}$$

We note that the second summation is over all choices of l_s, $1 \le s \le m$, and of S_m which satisfy the stated conditions. This comment also applies to similar subsequent summations.

Note that $\mathcal{V}_{3q} = 0$ and $\mathcal{V}_0 = \mathbb{E}\Big[\sum_{l_0 \in [u-q, t+q]\setminus\{u\}} L^2_{k_{l_0}, k_u} \Big(\Delta_{\mathsf{span}}^{\min\{t, l_0\}, \emptyset} x^{\pi, l_0}_{k_{l_0}} \Big)^2 \Big]$.

$$\text{Also,} \quad l_{m-1} \le \min\{l_{m-1}, t_{m-2} + q\} \le t_{m-1} + q. \tag{21}$$

Lemma 21 below gives our recursive bound; it is then used to prove Lemmas 22 and 11. Finally, we prove Lemma 21 via Lemma 23, which is stated right before the proof of Lemma 21.

Lemma 21. *For $t - 2q \le u \le t$,*

$$\mathcal{V}_{m-1} \le 40q\mathcal{V}_m + \mathbb{E}\Bigg[\sum_{s \in [t-7q, t+q]\setminus\{u\}} 120q\Gamma^2 \frac{\left(\Lambda^2\right)^{m+1} (4q)^m}{n^{m+1}} \Big(\big(\overline{\Delta}_{\mathsf{span}} x^{\pi, s}_{k_s}\big)^2 + \big(\Delta x^{\pi, s}_{k_s}\big)^2 \Big) \Bigg] \tag{22}$$

$$+ \mathbb{E}\Bigg[24 L^2_{\mathsf{max}} \frac{\left(\Lambda^2\right)^m (4q)^m}{n^m} \Big(\big(\overline{\Delta}_{\mathsf{span}} x^{\pi, u}_{k_u}\big)^2 + \big(\Delta x^{\pi, u}_{k_u}\big)^2 \Big) \Bigg]. \tag{23}$$

Lemma 21 readily yields a bound on $\mathcal{V}_0$. Recall that $r = \frac{160q^2\Lambda^2}{n}$.

Lemma 22.

$$\mathcal{V}_0 \le \frac{3r^2\Gamma^2}{160(1-r)q^2} \sum_{\substack{s \in [t-7q, t+q] \\ \setminus\{u\}}} \Big[(\mathcal{D}_s)^2 + (\Delta^X_s)^2 \Big] + \frac{3r\Gamma^2}{5q(1-r)} \cdot \Big[(\mathcal{D}_u)^2 + (\Delta^X_u)^2 \Big].$$

Proof. We will apply Lemma 21 recursively to $\mathcal{V}_0$. Note that as m increases by 1, each sum is multiplied by $\frac{\Lambda^2 \cdot (4q)}{n}$, and the first term has a multiplier of $40q$ (which is why we chose $r = 40q \cdot \frac{\Lambda^2(4q)}{n}$). We obtain

$$
\mathcal{V}_0 \leq (1 + r + r^2 + \cdots) \cdot \mathbb{E}\Bigg[\Bigg(\sum_{s \in [t-7q, t+q] \setminus \{u\}} \frac{480 q^2 \Gamma^2 (\Lambda^2)^2}{n^2} \left(\left(\overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi,s}\right)^2 + \left(\Delta x_{k_s}^{\pi,s}\right)^2 \right)
$$
$$
+ \frac{96 q (\Lambda^2) L_{\max}^2}{n} \left(\left(\overline{\Delta}_{\mathsf{span}} x_{k_u}^{\pi,u}\right)^2 + \left(\Delta x_{k_u}^{\pi,u}\right)^2 \right) \Bigg) \Bigg]
$$
$$
\leq \frac{3 r^2 \Gamma^2}{160(1-r) q^2} \sum_{s \in [t-7q, t+q] \setminus \{u\}} \left[(\mathcal{D}_s)^2 + (\Delta_s^X)^2 \right] + \frac{3 r \Gamma^2}{5 q (1-r)} \left[(\mathcal{D}_u)^2 + (\Delta_u^X)^2 \right],
$$

as $L_{\max} \leq \Gamma$, $r < 1$, and replacing $160 q^2 \Lambda^2 / n$ by r. $\qquad \square$

Recall that $R_{m-1} = \{l_0, l_1, \cdots, l_{m-1}\}$ and $t_{m-1} = \min\{t, l_0, l_1, \cdots, l_{m-1}\}$. In the proofs of Lemmas 11 and 21, we define $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s}$ as follows. If $s \notin A^{\pi, t_{m-1}}$, then $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s}$ is the value of $\Delta x_{k_s}^{\pi,s}$ when $\mathcal{U}_s$ reads all its inputs and makes its update immediately after first $t_{m-1} - 4q - 1$ updates committed; otherwise, $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s} = \Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s}$. We make the following observation.

Observation 1. *For $m \geq 1$,*
i. If $s \in [t_{m-1} - 4q, t_{m-1} + q] \setminus \{u\}$, then $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s} \in \left[\Delta_{\min}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s}, \Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s} \right]$;
ii. If $s \in [t_{m-1} - 4q, t_{m-1} + q] \setminus \{u\}$, then $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s}$ is independent of $\mathcal{U}_u$ and $\mathcal{U}_v$ for $v \in R_{m-1}$;
iii. If $s \in [t - 4q, t + q] \setminus \{u\}$, then $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s} \in \left[\Delta_{\min}^{t_{m-1}, \emptyset} x_{k_s}^{\pi,s}, \Delta_{\max}^{t_{m-1}, \emptyset} x_{k_s}^{\pi,s} \right]$;
iv. If $s \in [t - 4q, t + q]$, then $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s}$ is independent of $\mathcal{U}_u$.

Proof. i. When $s \in A^{\pi, t_{m-1}}$, $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s} = \Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s}$m and so the result is immediate. We also note, for use in (ii), that in this case $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s}$ is independent of R_{m-1}.

While if $s \notin A^{\pi, t_{m-1}}$, we argue as follows. Note that every element in R_{m-1} is at least t_{m-1}. By Lemma 1, the updates in R_{m-1} have commit time at least $t_{m-1} + 1$, while each of the first $t_{m-1} - 4q - 1$ updates have commit time at most $t_{m-1} - 4q - 1 + q + 1 = t_{m-1} - 3q$. Thus, in this case too, the updates in R_{m-1} are excluded from the computation of update $\mathcal{U}_s$.

Also, by the definition of $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s}$, $\mathcal{U}_s$ reads all its inputs and makes its update immediately after first $t_{m-1} - 4q - 1$ updates committed, all updates in $A^{\pi, t_{m-1}}$ have been fixed before $\mathcal{U}_s$ starts its reads. Consequently (i) holds in this case too.

ii. We have already shown the independence for $v \in R_{m-1}$ in the proof of (i).

For update $\mathcal{U}_u$, first recall that $t_{m-1} - 2q \leq u$. If $s \in A^{\pi, t_{m-1}}$, we want to show that $\Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s}$ is independent of $\mathcal{U}_u$. By the definition of $A^{\pi, t_{m-1}}$, $\mathcal{U}_s$ commits earlier than some update $\mathcal{U}_p$ where $p < t_{m-1} - 4q$. By Lemma 1, the commit time of $\mathcal{U}_p$ is at most $p + q + 1 \leq t_{m-1} - 3q \leq u - q$, hence $\mathcal{U}_s$ must commit before time $u - q$, and hence it is independent of $\mathcal{U}_u$ whose start time is at least $u - q + 1$ by Lemma 1. If $s \notin A^{\pi, t}$, $\widetilde{\Delta}^{t} x_{k_s}^{\pi,s}$ depends only on the first $t_{m-1} - 4q - 1 < u - 2q$ updates, and these updates must be independent of $\mathcal{U}_u$ by Lemma 1.

iii. The argument is similar to (i).

iv. The argument is similar to (ii). $\qquad \square$

Proof of Lemma 11. For brevity, we write $\mathcal{V} = \left(\sum_{l_0 \in [t-4q,t+q] \setminus \{u\}} L_{k_{l_0},k_u} \Delta_{\mathsf{var}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right)^2$. Recall that

$$
\mathcal{V} = \left(\sum_{l_0 \in [t-4q,t+q] \setminus \{u\}} L_{k_{l_0},k_u} \Delta_{\mathsf{var}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right)^2
$$

$$
= \left(\sum_{l_0 \in [t-4q,t+q] \setminus \{u\}} L_{k_{l_0},k_u} \max \left\{ \Delta_{\mathsf{span}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0}, \left| \Delta_{\mathsf{max}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right|, \left| \Delta_{\mathsf{min}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right| \right\} \right)^2.
$$

By Observation 1(iii), recentering and applying the Cauchy-Schwarz inequality,

$$
\mathcal{V} \leq 10q \sum_{l_0 \in [t-4q,t+q] \setminus \{u\}} L_{k_{l_0},k_u}^2 \left[\left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right)^2 + \left(\widetilde{\Delta}^t x_{k_{l_0}}^{\pi,l_0} \right)^2 \right]
$$

$$
\leq 10q \sum_{l_0 \in [u-q,t+q] \setminus \{u\}} L_{k_{l_0},k_u}^2 \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \right)^2 + 10q \sum_{s \in [t-4q,u-q)} L_{k_s,k_u}^2 \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_s}^{\pi,s} \right)^2
$$

$$
+ 10q \sum_{s \in [t-4q,t+q] \setminus \{u\}} L_{k_s,k_u}^2 \left(\widetilde{\Delta}^t x_{k_s}^{\pi,s} \right)^2.
$$

Note that by Lemma 9(ii), $\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_{l_0}}^{\pi,l_0} \leq \Delta_{\mathsf{span}}^{\min\{t,l_0\},\emptyset} x_{k_{l_0}}^{\pi,l_0}$. For $s < u - q$, $\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_s}^{\pi,s}$ is independent of update $\mathcal{U}_u$, as $\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_s}^{\pi,s}$ is independent of updates $\mathcal{U}_v$ for $v > s + q$. Also, as already noted, $\widetilde{\Delta}^t x_{k_s}^{\pi,s}$ is independent of k_u by Observation 1(iv). Thus, taking the expectation of the previous bound as k_u varies, yields

$$
\mathbb{E}_{k_u}[\mathcal{V}] \leq \mathbb{E}_{k_u}\left[10q \sum_{l_0 \in [u-q,t+q] \setminus \{u\}} L_{k_{l_0},k_u}^2 \left(\Delta_{\mathsf{span}}^{\min\{t,l_0\},\emptyset} x_{k_{l_0}}^{\pi,l_0} \right)^2 \right]
$$

$$
+ \mathbb{E}_{k_u}\left[10q \sum_{s \in [t-4q,u-q)} \frac{L_{\overline{\mathsf{res}}}^2}{n} \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_s}^{\pi,s} \right)^2 \right]
$$

$$
+ \mathbb{E}_{k_u}\left[10q \sum_{s \in [t-4q,t+q] \setminus \{u\}} \frac{L_{\overline{\mathsf{res}}}^2}{n} \left(\widetilde{\Delta}^t x_{k_s}^{\pi,s} \right)^2 \right].
$$

Note that for $s \in [t-4q, t+q] \setminus \{u\}$, $\widetilde{\Delta}^t x_{k_s}^{\pi,s}, \Delta x_{k_s}^{\pi,s} \in [\overline{\Delta}_{\min} x_{k_s}^{\pi,s}, \overline{\Delta}_{\max} x_{k_s}^{\pi,s}]$. Thus, by recentering,

$\left(\widetilde{\Delta}^t x_{k_s}^{\pi,s}\right)^2 \le 2\left(\Delta x_{k_s}^{\pi,s}\right)^2 + 2\left(\overline{\Delta}_{\text{span}} x_{k_{l_0}}^{\pi,s}\right)^2$. Then we average over all paths π:

$$\mathbb{E}[\mathcal{V}] \le \mathbb{E}\left[10q \sum_{l_0 \in [u-q,t+q]\setminus\{u\}} L_{k_{l_0},k_u}^2 \left(\Delta_{\text{span}}^{\min\{t,l_0\},\emptyset} x_{k_{l_0}}^{\pi,l_0}\right)^2\right]$$

$$+ \mathbb{E}\left[30q \sum_{s\in[t-4q,t+q]\setminus\{u\}} \frac{L_{\overline{\text{res}}}^2}{n}\left(\left(\overline{\Delta}_{\text{span}} x_{k_s}^{\pi,s}\right)^2 + \left(\Delta x_{k_s}^{\pi,s}\right)^2\right)\right]$$

$$\le \frac{3r^2\Gamma^2}{16(1-r)q} \sum_{\substack{s\in[t-7q,t+q]\\ \setminus\{u\}}} \left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right] + \frac{6r\Gamma^2}{1-r}\left[\left(\mathcal{D}_u\right)^2 + \left(\Delta_u^X\right)^2\right]$$

$$+ 30q \sum_{s\in[t-4q,t+q]\setminus\{u\}} \frac{L_{\overline{\text{res}}}^2}{n}\left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right] \quad \text{(by Lemma 22)}$$

$$\le \frac{3r^2\Gamma^2}{16(1-r)q} \sum_{\substack{s\in[t-7q,t+q]\\ \setminus\{u\}}} \left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right] + \frac{6r\Gamma^2}{1-r}\cdot\left[\left(\mathcal{D}_u\right)^2 + \left(\Delta_u^X\right)^2\right]$$

$$+ \frac{3r\Gamma^2}{16q} \sum_{s\in[t-4q,t+q]\setminus\{u\}} \left[\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right] \quad \left(\text{as } r = \frac{160\Lambda^2 q^2}{n} \text{ and } \Lambda = \frac{L_{\overline{\text{res}}}^2}{\Gamma^2} + 1.\right)$$

$\square$

In the next lemma we bound $\left(\Delta_{\text{span}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}} x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2$.

Lemma 23. *Suppose* $l_0, l_1, \cdots \in [u-q,t+q]\setminus\{u\}$, $t_s = \min\{t, l_0, l_1, \cdots, l_s\}$, $R_{m-1} = \{l_0, l_1, \cdots, l_{m-1}\}$ *and* $l_s \le t_s + q$ *for* $s < m$. *Then*

$$\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}}x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2$$

$$\leq 40q\sum_{\substack{l_m\in[t_{m-1}-4q,\min\{l_{m-1}-1,u-q-1\}]\setminus(\{u\}\cup R_{m-1})\\ \text{and } k_{l_m}=k_{l_{m-1}}}}$$

$$\left[\underbrace{\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{F}+\underbrace{\left(\widetilde{\Delta}^{t_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{G}\right]$$

$$+40q\sum_{\substack{l_m\in[\min\{l_{m-1}-1,u-q-1\}+1,l_{m-1}-1]\setminus(\{u\}\cup R_{m-1})\\ \text{and } k_{l_m}=k_{l_{m-1}}}}$$

$$\left[\underbrace{\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{H}+\underbrace{\left(\widetilde{\Delta}^{t_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{I}\right]$$

$$+8\cdot\mathbb{1}_{(k_{l_{m-1}}=k_u \text{ and } u\leq l_{m-1})}\cdot\underbrace{\left[\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2+\left(\Delta_{\mathsf{max}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2\right]}_{J}$$

$$+\frac{40q}{\Gamma^2}\left(\sum_{l_m\in[t_{m-1}-4q,\min\{t_{m-1}+q,u-q-1\}]\setminus(\{l_{m-1},u\}\cup R_{m-1})}\right.$$

$$\left.\left(\underbrace{L_{k_{l_m}k_{l_{m-1}}}^2\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{K}+\underbrace{L_{k_{l_m}k_{l_{m-1}}}^2\left(\widetilde{\Delta}^{t_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{L}\right)\right)$$

$$+\frac{40q}{\Gamma^2}\left(\sum_{l_m\in[\min\{t_{m-1}+q,u-q-1\}+1,t_{m-1}+q]\setminus(\{l_{m-1},u\}\cup R_{m-1})}\right.$$

$$\left.\left(\underbrace{L_{k_{l_m}k_{l_{m-1}}}^2\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{M}+\underbrace{L_{k_{l_m}k_{l_{m-1}}}^2\left(\widetilde{\Delta}^{t_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2}_{N}\right)\right)$$

$$+\frac{8}{\Gamma^2}\cdot\mathbb{1}_{(u\leq t_{m-1}+q)}\cdot L_{k_uk_{l_{m-1}}}^2\underbrace{\left[\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2+\left(\Delta_{\mathsf{max}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2\right]}_{O}. \qquad (24)$$

Proof. Recall that $t_{m-1}=\min\{t,l_0,l_1,\cdots,l_{m-1}\}$, and R_{m-1} contains l_{m-1}. Also recall that

$$\Delta_{\mathsf{var}}^{t_{m-1},R}x_{k_s}^{\pi,s}=\max\left\{\Delta_{\mathsf{max}}^{t_{m-1},R}x_{k_s}^{\pi,s}-\Delta_{\mathsf{min}}^{t_{m-1},R}x_{k_s}^{\pi,s},\left|\Delta_{\mathsf{max}}^{t_{m-1},R}x_{k_s}^{\pi,s}\right|,\left|\Delta_{\mathsf{min}}^{t_{m-1},R}x_{k_s}^{\pi,s}\right|\right\}.$$

First, let's expand the term

$$\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}}x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2$$

$$=\left(\Delta_{\mathsf{max}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}}x_{k_{l_{m-1}}}^{\pi,l_{m-1}}-\Delta_{\mathsf{min}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}}x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2,$$

By Lemma 8, and recalling that the definition of $\Delta_{\mathsf{var}}^{t_{m-1}}x_{k_{l_m}}^{\pi,l_m}$ allows the choice of whether to read the results of updates $\mathcal{U}_s$ with $s\in[t_{m-1}-4q,t_{m-1}+q]$, we obtain

$$\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}}x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2$$

$$\leq 2\left(\sum_{\substack{l_m\in[t_{m-1}-4q,\\ l_{m-1}-1]\setminus R_{m-1}\\ \text{and } k_{l_m}=k_{l_{m-1}}}}\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$+\frac{2}{\Gamma^2}\left(\widetilde{g}_{\max,k_{l_{m-1}}}^{t_{m-1},R_{m-1},\pi,l_{m-1}}-\widetilde{g}_{\min,k_{l_{m-1}}}^{t_{m-1},R_{m-1},\pi,l_{m-1}}\right)^2$$

$$\leq 2\left(\sum_{\substack{l_m\in[t_{m-1}-4q,\\ l_{m-1}-1]\setminus R_{m-1}\\ \text{and } k_{l_m}=k_{l_{m-1}}}}\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$+\frac{2}{\Gamma^2}\left(\sum_{\substack{l_m\in[t_{m-1}-4q,\\ t_{m-1}+q]\setminus R_{m-1}}}L_{k_{l_m}k_{l_{m-1}}}\cdot\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2. \tag{25}$$

By applying the unbalanced Cauchy-Schwarz inequality, the first term on the RHS of (25) can be bounded as follows (the $5q$ in the first inequality occurs because by (21), $l_{m-1}\leq t_{m-1}+q$, and hence the range for l_m is of size at most $5q$):

$$2\left(\sum_{\substack{l_m\in[t_{m-1}-4q,l_{m-1}-1]\setminus R_{m-1}\\ \text{and } k_{l_m}=k_{l_{m-1}}}}\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$\leq 4\cdot(5q)\sum_{\substack{l_m\in[t_{m-1}-4q,\\ l_{m-1}-1]\setminus(\{u\}\cup R_{m-1})\\ \text{and } k_{l_m}=k_{l_{m-1}}}}\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$+4\cdot\mathbb{1}_{(k_{l_{m-1}}=k_u\text{ and } u\leq l_{m-1}-1)}\cdot\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2. \tag{26}$$

Again, by using an unbalanced Cauchy-Schwarz inequality, the second term on the RHS of (25) can be bounded as follows:

$$\frac{2}{\Gamma^2}\left(\sum_{\substack{l_m\in[t_{m-1}-4q,t_{m-1}+q]\setminus R_{m-1}}}L_{k_{l_m}k_{l_{m-1}}}\cdot\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$\leq\frac{4(5q)}{\Gamma^2}\sum_{\substack{l_m\in[t_{m-1}-4q,\\ t_{m-1}+q]\\ \setminus(\{u\}\cup R_{m-1})}}L_{k_{l_m}k_{l_{m-1}}}^2\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_{l_m}}^{\pi,l_m}\right)^2$$

$$+\frac{4}{\Gamma^2}\cdot\mathbb{1}_{(u\leq t_{m-1}+q)}\cdot L_{k_uk_{l_{m-1}}}^2\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2. \tag{27}$$

Observation 1(i) implies that for any $s=l_m\in[t_{m-1}-4q,t_{m-1}+q]\setminus\{u\}$, $\left(\Delta_{\min}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2$, $\left(\Delta_{\max}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2\leq\left(\left|\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right|+\left|\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right|\right)^2$, and consequently by recentering,

$$\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2\leq 2\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2+2\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2. \tag{28}$$

Similarly, since $\Delta_{\max}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u} \in \left[\Delta_{\min}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u}, \Delta_{\max}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u}\right],$

$$\left(\Delta_{\mathsf{var}}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u}\right)^2 \leq 2\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u}\right)^2 + 2\left(\Delta_{\max}^{t_{m-1},R_{m-1}} x_{k_u}^{\pi,u}\right)^2. \tag{29}$$

In summary, (25) is bounded by (26) and (27); we split the range of the first term in each of these bounds, and bound the Δ_{var} terms as in (28) and (29). This yields the result. □

We can now prove Lemma 21.

Proof of Lemma 21. Next, we bound the LHS of the inequality in Lemma 21. To obtain our bound, we start with (24) and multiply both sides by

$$\left(\prod_{\substack{l_s \in R_{m-1}\setminus S_{m-1} \\ \text{where } R_{m-1} \\ =\{l_1,l_2,\cdots,l_{m-1}\}.}} \frac{L_{k_{l_s}k_{l_{s-1}}}^2}{\Gamma^2}\right) L_{k_{l_0}k_u}^2,$$

and sum over all choices of $l_0, l_1, \cdots, l_{m-1}$, and over all choices of a set S_{m-1} defined shortly, where these parameters satisfy the following constraints.
i. $l_0, l_1, \cdots, l_{m-1} \in [u-q, t+q]$.
ii. All the l_s are distinct and $l_s \neq u$ for all s.
iii. Let $t_{s-1} = \min\{t, l_0, l_1, \cdots, l_{s-1}\}$; $l_s \leq t_{s-1} + q$ for all s.
iv. $S_{m-1} \subseteq \{l_s | k_{l_s} = k_{l_{s-1}}\}$.
v. Let $R_{m-1} = \{l_1, l_2, \ldots, l_m\}$; for all $s \geq 1$, $l_s \in R_{m-1} \setminus S_{m-1}$.
This multiplication of the LHS term of (23) yields

$$\mathbb{E}\left[\sum_{l_0 \in [u-q,t+q]\setminus\{u\}} \left(\sum_{\substack{l_1,l_2,\cdots,l_{m-1}\in[u-q,t+q], \\ \text{all } l_s \text{ distinct, all } l_s\neq u,l_0; \\ \text{all } l_s\leq t_{s-1}+q, \text{ where} \\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\} \\ \text{all } S_{m-1}\subseteq\{l_s|k_{l_s}=k_{l_{s-1}}\}.}} \left(\prod_{\substack{l_s \in R_{m-1}\setminus S_{m-1} \\ \text{where } R_{m-1} \\ =\{l_1,l_2,\cdots,l_{m-1}\}.}} \frac{L_{k_{l_s}k_{l_{s-1}}}^2}{\Gamma^2}\right) \right.$$
$$\left. L_{k_{l_0}k_u}^2 \left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}\setminus\{l_{m-1}\}} x_{k_{l_{m-1}}}^{\pi,l_{m-1}}\right)^2\right)\right],$$

which is $\mathcal{V}_{m-1}$.

On the RHS, we start by looking at terms H and M to obtain the recursive term $\mathcal{V}_m$ on the RHS of the Lemma 21. We use the inequality $\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}} x_{k_{l_m}}^{\pi,l_m}\right)^2 \leq \left(\Delta_{\mathsf{span}}^{t_m,R_{m-1}} x_{k_{l_m}}^{\pi,l_m}\right)^2$, which holds by Lemma 9(ii). The multiplication applied to these two terms yields

$$\leq \mathbb{E}\left[40q \sum_{\substack{l_0 \in [u-q,t+q] \\ \backslash \{u\}}} \left(\sum_{\substack{l_1,l_2,\cdots,l_m \in [u-q,t+q], \\ \text{all } l_s \text{ distinct, all } l_s \neq u,l_0; \\ \text{all } l_s \leq t_{s-1}+q, \text{where} \\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\}; \\ \text{all } S_{m-1} \subseteq \{l_s | k_{l_s}=k_{l_{s-1}} \\ \text{and } s \leq m-1\}; \\ k_{l_m}=k_{l_{m-1}}.}} \left(\prod_{\substack{l_s \in R_{m-1} \backslash S_{m-1} \\ \text{where } R_{m-1} \\ =\{l_1,l_2,\cdots,l_{m-1}\}.}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \right) L^2_{k_{l_0}k_u} \right.\right.$$

$$\left.\left. \cdot \left(\Delta^{t_m,R_{m-1}}_{\mathsf{span}} x^{\pi,l_m}_{k_{l_m}} \right)^2 \right) \right]$$

$$+ \mathbb{E}\left[40q \sum_{l_0 \in [u-q,t+q] \backslash \{u\}} \left(\sum_{\substack{l_1,l_2,\cdots,l_m \in [u-q,t+q], \\ \text{all } l_s \text{ distinct, all } l_s \neq u,l_0; \\ \text{all } l_s \leq t_{s-1}+q, \text{ where} \\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\}; \\ S_{m-1} \subseteq \{l_s | k_{l_s}=k_{l_{s-1}} \\ \text{and } s \leq m-1\}.}} \left(\prod_{\substack{l_s \in R_{m-1} \backslash S_{m-1} \\ \text{where } R_{m-1} \\ =\{l_1,l_2,\cdots,l_{m-1}\}.}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \right) \right.\right.$$

$$\left.\left. \cdot L^2_{k_{l_0}k_u} \cdot \frac{L^2_{k_{l_m}k_{l_{m-1}}}}{\Gamma^2} \left(\Delta^{t_m,R_{m-1}}_{\mathsf{span}} x^{\pi,l_m}_{k_{l_m}} \right)^2 \right) \right]$$

These two terms correspond to including l_m in S_m and not including it in the constraint for $\mathcal{V}_m$; in other words, summed together these two terms equal $\mathcal{V}_m$.

We now analyze the remaining non-recursive terms.

Recall that $\Lambda^2 = \frac{L^2_{\overline{res}}}{\Gamma^2} + 1$. Term J is bounded as follows, as we explain below:

$$\mathbb{E}\left[\sum_{\substack{l_0 \in [u-q,t+q] \\ \backslash \{u\}}} \left(\sum_{\substack{l_1,l_2,\cdots,l_{m-1} \in [u-q,t+q] \\ \text{all } l_s \text{ distinct, all } l_s \neq u,l_0; \\ \text{all } l_s \leq t_{s-1}+q, \text{ where} \\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\}; \\ \text{all } S_{m-1} \subseteq \{l_s | k_{l_s}=k_{l_{s-1}}\}.}} \left(\prod_{\substack{l_s \in R_{m-1} \backslash S_{m-1} \\ \text{where } R_{m-1} \\ =\{l_1,l_2,\cdots,l_{m-1}\}.}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \right) L^2_{k_{l_0}k_u} \right.\right.$$

$$\left.\left. \cdot 8 \cdot \mathbb{1}_{\substack{k_{l_{m-1}}=k_u \\ \text{and } u \leq l_{m-1}}} \left[\left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,u}_{k_u} \right)^2 + \left(\Delta^{t_{m-1},R_{m-1}}_{\max} x^{\pi,u}_{k_u} \right)^2 \right] \right) \right]$$

$$\leq \mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1} \\ \in [u-q,t+q] \backslash \{u\} \\ u \leq t_{m-1}+q}} 8 L^2_{\max} \frac{(\Lambda^2)^{m-1}}{n^m} \left[\left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,u}_{k_u} \right)^2 + \left(\Delta^{t_{m-1},R_{m-1}}_{\max} x^{\pi,u}_{k_u} \right)^2 \right] \right]. \tag{30}$$

Recall that by (21), $l_{m-1} \leq t_{m-1} + q$, justifying the bound on the sum. We first bound $L^2_{k_{l_0}k_u}$ by $L^2_{\max}$. Then we can average, in turn, over the random coordinate choices of k_{l_0}, k_{l_1}, ..., $k_{l_{m-1}}$. As we will explain below, depending on whether each l_{s-1} is in S_{m-1} or not, the averaging of $\frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2}$ over $k_{l_{s-1}}$ will provide either an $\frac{1}{n}$ or an $\frac{1}{n} \cdot \frac{L^2_{\overline{res}}}{\Gamma^2}$ factor. To elaborate, the factor $\frac{(\Lambda^2)^{m-1}}{n^m}$ on the RHS is due to a combination of the following observations:

i. As in Section 4.2.2, the $m-1$ factors of $\frac{\left(\frac{L^2_{\overline{res}}}{\Gamma^2} + 1 \right)}{n} = \frac{\Lambda^2}{n}$ are due to the expectation over the bundle of n paths obtained by varying $k_{l_{s-1}}$, which does not affect the terms $\Delta^{t,R_{m-1}}_{\mathsf{span}} x^{\pi,u}_{k_u}$, $\Delta^{t,R_{m-1}}_{\max} x^{\pi,u}_{k_u}$, as

43

$l_{s-1} \in R_{m-1}$ and $l_{s-1} \neq u$. In more detail, there is always a term due to the case $k_{l_{s-1}} \notin S_{m-1}$, which, on averaging, yields a term $\frac{1}{n} \cdot \frac{L_{\text{res}}^2}{\Gamma^2}$; when $k_{l_s} = k_{l_{s-1}}$, there is also a term for the case that $k_{l_{s-1}} \in S_{m-1}$, which, on averaging, yields a term of the form $\frac{1}{n}$; in combination, they yield a term $\frac{\Lambda^2}{n}$. (if $l_s \in S_{m-1}$, then $k_{l_s} = k_{l_{s-1}}$, and in this case taking the expectation yields an $1/n$ factor).

ii. The extra factor of n in the denominator is due to the expectation of $\mathbb{1}_{k_{l_{m-1}} = k_u}$.

For the remaining terms, first note that, on the RHS of (24), for any $s \in [t_{m-1} - 4q, t_{m-1} + q] \setminus \{u\}$, $\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s}$ is independent of any updates in R_{m-1} or update $\mathcal{U}_u$ by Observation 1(ii). In addition, for $s < u - q$, $\Delta_{\text{span}}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s}$ also doesn't depend on any updates in R_{m-1} or update $\mathcal{U}_u$.

For term F, note that on the RHS, l_m is being renamed s. Here we can start by averaging over k_u, as $l_m \leq u - q$, which yields a factor of $\frac{L_{\text{res}}^2}{n} \leq \frac{\Gamma^2 \Lambda^2}{n}$. The remaining $m-1$ factors of $\frac{\Lambda^2}{n}$ are as before. The final factor of $\frac{1}{n}$ is due to the condition $k_{l_m} = k_{l_{m-1}}$.

$$
\mathbb{E}\Bigg[\sum_{\substack{l_0 \in [u-q,t+q] \\ \setminus \{u\}}} \Bigg(\sum_{\substack{l_1,l_2,\cdots,l_{m-1} \in [u-q,t+q] \\ \text{all } l_s \text{ distinct, all } l_s \neq u, l_0; \\ \text{all } l_s \leq t_{s-1}+q, \text{ where} \\ t_{s-1} = \min\{t,l_0,l_1,\cdots,l_{s-1}\}; \\ \text{all } S_{m-1} \subseteq \{l_s | k_{l_s} = k_{l_{s-1}}\}}} \Bigg(\prod_{\substack{l_s \in R_{m-1} \setminus S_{m-1} \\ \text{where } R_{m-1} \\ = \{l_1,l_2,\cdots,l_{m-1}\};}} \frac{L_{k_{l_s} k_{l_{s-1}}}^2}{\Gamma^2} \Bigg) L_{k_{l_0} k_u}^2
$$

$$
\cdot 40q \cdot \sum_{\substack{l_m \in [t_{m-1}-4q,u-q-1] \\ \setminus (\{u\} \cup R_{m-1}) \\ k_{l_m} = k_{l_{m-1}}}} \Big(\Delta_{\text{span}}^{t_{m-1}, R_{m-1}} x_{k_{l_m}}^{\pi,l_m} \Big)^2
$$

$$
\leq \mathbb{E}\Bigg[\sum_{\substack{l_0,l_1,\cdots,l_{m-1} \\ \in [u-q,t+q] \setminus \{u\}}} 40q\Gamma^2 \frac{(\Lambda^2)^m}{n^{m+1}} \sum_{\substack{s \in [t_{m-1} \\ -4q,u-q)}} \Big(\Delta_{\text{span}}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi,s} \Big)^2 \Bigg]. \tag{31}
$$

For the next expression, we sum terms G and I. Again, we can start by averaging over k_u, as $\widetilde{\Delta}^{t_{m-1}} x_{k_{l_m}}^{\pi,l_m}$ is independent of $\mathcal{U}_u$ by Observation 1(ii), as $l_m \in [t_{m-1} - 4q, l_{m-1} - 1] \subset [t_{m-1} - 4q, t_{m-1} + q]$ (as can be seen by applying (21)).

$$
\mathbb{E}\Bigg[\sum_{\substack{l_0 \in [u-q,t+q] \\ \setminus \{u\}}} \Bigg(\sum_{\substack{l_1,l_2,\cdots,l_{m-1} \in [u-q,t+q] \\ \text{all } l_s \text{ distinct, all } l_s \neq u, l_0; \\ \text{all } l_s \leq t_{s-1}+q, \text{ where} \\ t_{s-1} = \min\{t,l_0,l_1,\cdots,l_{s-1}\}; \\ \text{all } S_{m-1} \subseteq \{l_s | k_{l_s} = k_{l_{s-1}}\}}} \Bigg(\prod_{\substack{l_s \in R_{m-1} \setminus S_{m-1} \\ \text{where } R_{m-1} \\ = \{l_1,l_2,\cdots,l_{m-1}\};}} \frac{L_{k_{l_s} k_{l_{s-1}}}^2}{\Gamma^2} \Bigg) L_{k_{l_0} k_u}^2
$$

$$
\cdot 40q \cdot \sum_{\substack{l_m \in [t_{m-1}-4q,l_{m-1}-1] \\ \setminus (\{u\} \cup R_{m-1}) \\ k_{l_m} = k_{l_{m-1}}}} \Big(\widetilde{\Delta}^{t_{m-1}} x_{k_{l_m}}^{\pi,l_m} \Big)^2
$$

$$
\leq \mathbb{E}\Bigg[\sum_{\substack{l_0,l_1,\cdots,l_{m-1} \\ \in [u-q,t+q] \setminus \{u\}}} 40q\Gamma^2 \frac{(\Lambda^2)^m}{n^{m+1}} \sum_{s \in [t_{m-1}-4q,l_{m-1}-1] \setminus \{u\}} \Big(\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s} \Big)^2 \Bigg]. \tag{32}
$$

For term K, there is an additional term $L_{k_{l_m} k_{l_{m-1}}}^2$ to average over; we bound it by $\frac{L_{\text{res}}^2}{n}$.

$$\mathbb{E}\Bigg[\sum_{\substack{l_0\in[u-q,t+q]\\ \setminus\{u\}}} \Bigg(\sum_{\substack{l_1,l_2,\cdots,l_{m-1}\in[u-q,t+q]\\ \text{all } l_s \text{ distinct, all } l_s\neq u,l_0;\\ \text{all } l_s\leq t_{s-1}+q,\ \text{where}\\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\};\\ \text{all } S_{m-1}\subseteq\{l_s|k_{l_s}=k_{l_{s-1}}\}.}} \Bigg(\prod_{\substack{l_s\in R_{m-1}\setminus S_{m-1}\\ \text{where } R_{m-1}\\ =\{l_1,l_2,\cdots,l_{m-1}\};}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \Bigg) L^2_{k_{l_0}k_u}$$

$$\cdot 40q \sum_{\substack{l_m\in[t_{m-1}-4q,u-q-1]\\ \setminus(\{u\}\cup R_{m-1})}} \frac{L^2_{k_{l_m}k_{l_{m-1}}}}{\Gamma^2} \left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,l_m}_{k_{l_m}} \right)^2$$

$$\leq \mathbb{E}\Bigg[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\setminus\{u\}}} 40q\Gamma^2 \frac{L^2_{\mathrm{res}}(\Lambda^2)^m}{\Gamma^2 n^{m+1}} \sum_{s\in[t_{m-1}-4q,u-q-1]} \left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,s}_{k_s} \right)^2 \Bigg]. \tag{33}$$

For the next expression, the sum of terms L and N, we obtain

$$\mathbb{E}\Bigg[\sum_{\substack{l_0\in[u-q,t+q]\\ \setminus\{u\}}} \Bigg(\sum_{\substack{l_1,l_2,\cdots,l_{m-1}\in[u-q,t+q]\\ \text{all } l_s \text{ distinct, all } l_s\neq u,l_0;\\ \text{all } l_s\leq t_{s-1}+q,\ \text{where}\\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\};\\ \text{all } S_{m-1}\subseteq\{l_s|k_{l_s}=k_{l_{s-1}}\}.}} \Bigg(\prod_{\substack{l_s\in R_{m-1}\setminus S_{m-1}\\ \text{where } R_{m-1}\\ =\{l_1,l_2,\cdots,l_{m-1}\};}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \Bigg) L^2_{k_{l_0}k_u}$$

$$\cdot 40q \cdot \sum_{\substack{l_m\in[t_{m-1}-4q,t_{m-1}+q]\\ \setminus(\{u\}\cup R_{m-1})}} \frac{L^2_{k_{l_m}k_{l_{m-1}}}}{\Gamma^2} \left(\widetilde{\Delta}^{t_{m-1}} x^{\pi,l_m}_{k_{l_m}} \right)^2$$

$$\leq \mathbb{E}\Bigg[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\setminus\{u\}}} 40q\Gamma^2 \frac{L^2_{\mathrm{res}}(\Lambda^2)^m}{\Gamma^2 n^{m+1}} \sum_{s\in[t_{m-1}-4q,t_{m-1}+q]\setminus\{u\}} \left(\widetilde{\Delta}^{t_{m-1}} x^{\pi,s}_{k_s} \right)^2 \Bigg]. \tag{34}$$

Finally, for term O, the bound is

$$\mathbb{E}\Bigg[\sum_{\substack{l_0\in[u-q,t+q]\\ \setminus\{u\}}} \Bigg(\sum_{\substack{l_1,l_2,\cdots,l_{m-1}\in[u-q,t+q]\\ \text{all } l_s \text{ distinct, all } l_s\neq u,l_0;\\ \text{all } l_s\leq t_{s-1}+q,\ \text{where}\\ t_{s-1}=\min\{t,l_0,l_1,\cdots,l_{s-1}\};\\ \text{all } S_{m-1}\subseteq\{l_s|k_{l_s}=k_{l_{s-1}}\}.}} \Bigg(\prod_{\substack{l_s\in R_{m-1}\setminus S_{m-1}\\ \text{where } R_{m-1}\\ =\{l_1,l_2,\cdots,l_{m-1}\};}} \frac{L^2_{k_{l_s}k_{l_{s-1}}}}{\Gamma^2} \Bigg) L^2_{k_{l_0}k_u}$$

$$\cdot 8 \cdot \frac{L^2_{k_u k_{l_{m-1}}}}{\Gamma^2} \mathbb{1}_{u\leq t_{m-1}+q} \left[\left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,u}_{k_u} \right)^2 + \left(\Delta^{t_{m-1},R_{m-1}}_{\max} x^{\pi,u}_{k_u} \right)^2 \right] \Bigg) \Bigg]$$

$$\leq \mathbb{E}\Bigg[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\\ \setminus\{u\}\\ u\leq t_{m-1}+q}} 8L^2_{\max} \frac{L^2_{\mathrm{res}}(\Lambda^2)^{m-1}}{\Gamma^2 n^m} \left[\left(\Delta^{t_{m-1},R_{m-1}}_{\mathsf{span}} x^{\pi,u}_{k_u} \right)^2 + \left(\Delta^{t_{m-1},R_{m-1}}_{\max} x^{\pi,u}_{k_u} \right)^2 \right] \Bigg]. \tag{35}$$

Therefore, the sum of the bounds on the non-recursive terms, reordered as 32, 34, 31, 33, 30, 35, equals

45

$$\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,l_{m-1}-1]\\ \backslash\{u\}}}40q\Gamma^2\frac{(\Lambda^2)^m}{n^{m+1}}\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2\right] \tag{36}$$

$$+\,\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,t_{m-1}+q]\\ \backslash\{u\}}}40q\Gamma^2\frac{L_{\overline{\text{res}}}^2(\Lambda^2)^m}{\Gamma^2 n^{m+1}}\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2\right] \tag{37}$$

$$+\,\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,u-q)\\ \backslash\{u\}}}40q\Gamma^2\frac{(\Lambda^2)^m}{n^{m+1}}\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2\right] \tag{38}$$

$$+\,\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,u-q)\\ \backslash\{u\}}}40q\Gamma^2\frac{L_{\overline{\text{res}}}^2(\Lambda^2)^m}{\Gamma^2 n^{m+1}}\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_s}^{\pi,s}\right)^2\right] \tag{39}$$

$$+\,\mathbb{E}\left[8\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}\\ u\le t_{m-1}+q}}L_{\max}^2\frac{(\Lambda^2)^{m-1}}{n^m}\left(\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2+\left(\Delta_{\max}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2\right)\right] \tag{40}$$

$$+\,\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}\\ u\le t_{m-1}+q}}8L_{\max}^2\frac{L_{\overline{\text{res}}}^2(\Lambda^2)^{m-1}}{\Gamma^2 n^m}\right.$$

$$\left.\cdot\left(\left(\Delta_{\mathsf{span}}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2+\left(\Delta_{\max}^{t_{m-1},R_{m-1}}x_{k_u}^{\pi,u}\right)^2\right)\right]. \tag{41}$$

To obtain the final result, we combine these bounds. We start by bounding the sum of (36) and (37) as follows. Noting that $\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}, \Delta x_{k_u}^{\pi,u}\in\left[\overline{\Delta}_{\min}x_{k_u}^{\pi,u},\overline{\Delta}_{\max}x_{k_u}^{\pi,u}\right]$ for $s\le t_{m-1}+q$, by recentering, $\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2\le 2\left(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\right)^2+2\left(\Delta x_{k_s}^{\pi,s}\right)^2$, we obtain

$$\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,l_{m-1}-1]\\ \backslash\{u\}}}40q\Gamma^2\frac{(\Lambda^2)^m}{n^{m+1}}\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2\right]$$

$$+\,\mathbb{E}\left[\sum_{\substack{l_0,l_1,\cdots,l_{m-1}\\ \in[u-q,t+q]\backslash\{u\}}}\sum_{\substack{s\in[t_{m-1}-4q,t_{m-1}+q]\\ \backslash\{u\}}}40q\Gamma^2\frac{L_{\overline{\text{res}}}^2(\Lambda^2)^m}{\Gamma^2 n^{m+1}}\left(\widetilde{\Delta}^{t_{m-1}}x_{k_s}^{\pi,s}\right)^2\right]$$

$$\le\mathbb{E}\left[\sum_{\substack{s\in[t-7q,t+q]\\ \backslash\{u\}}}80q\Gamma^2\frac{(\Lambda^2)^{m+1}(4q)^m}{n^{m+1}}\left(\left(\overline{\Delta}_{\mathsf{span}}x_{k_s}^{\pi,s}\right)^2+\left(\Delta x_{k_s}^{\pi,s}\right)^2\right)\right]. \tag{42}$$

The additional Λ^2 is from $1+\frac{L_{\overline{\text{res}}}^2}{\Gamma^2}$ on the LHS of (42), and the $(4q)^m$ is due to the $t+q-(u-q)\le (t+q)-(t-3q)=4q$ choices of $l_0,l_1,\cdots,l_{m-1}$. We also relax the range $[t_{m-1}-4q,t_{m-1}+q]\setminus\{u\}$ and the range $[t_{m-1}-4q,l_{m-1}-1]\setminus\{u\}$ to $[t-7q,t+q]\setminus\{u\}$ as $l_{m-1}\le t+q$, $t_{m-1}+q\le t+q$, and $t_{m-1}-4q\ge u-5q\ge t-7q$.

Next, we sum (38) and (39). As $t_{m-1} = \min\{t, l_0, \cdots, l_{m-1}\} \geq s - q$, we obtain

$$\mathbb{E}\Bigg[\sum_{\substack{l_0, l_1, \cdots, l_{m-1} \\ \in [u-q, t+q] \setminus \{u\}}} \sum_{\substack{s \in [t_{m-1}-4q, u-q) \\ \setminus \{u\}}} 40q\Gamma^2 \frac{(\Lambda^2)^m}{n^{m+1}} \left(\Delta_{\mathsf{span}}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi, s}\right)^2 \Bigg]$$

$$+ \mathbb{E}\Bigg[\sum_{\substack{l_0, l_1, \cdots, l_{m-1} \\ \in [u-q, t+q] \setminus \{u\}}} \sum_{\substack{s \in [t_{m-1}-4q, u-q) \\ \setminus \{u\}}} 40q\Gamma^2 \frac{L_{\mathsf{res}}^2 (\Lambda^2)^m}{\Gamma^2 n^{m+1}} \left(\Delta_{\mathsf{span}}^{t_{m-1}, R_{m-1}} x_{k_s}^{\pi, s}\right)^2 \Bigg]$$

$$\leq \mathbb{E}\Bigg[\sum_{s \in [t-7q, t+q] \setminus \{u\}} 40q\Gamma^2 \frac{(\Lambda^2)^{m+1}(4q)^m}{n^{m+1}} \left(\overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi, s}\right)^2 \Bigg]. \tag{43}$$

Adding (42) and (43) yields the term (22) on the RHS of Lemma 21.

Finally, we sum (40) and (41). For $u \leq t_{m-1}+q$, as $\Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}, \Delta x_{k_u}^{\pi, u} \in \left[\overline{\Delta}_{\min} x_{k_u}^{\pi, u}, \overline{\Delta}_{\max} x_{k_u}^{\pi, u}\right]$, by recentering $\left(\Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 \leq 2\left(\overline{\Delta}_{\mathsf{span}} x_{k_u}^{\pi, u}\right)^2 + 2\left(\Delta x_{k_u}^{\pi, u}\right)^2$, and $\left(\Delta_{\mathsf{span}}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 \leq \left(\overline{\Delta}_{\mathsf{span}} x_{k_u}^{\pi, u}\right)^2$.

$$\mathbb{E}\Bigg[8 \sum_{\substack{l_0, l_1, \cdots, l_{m-1} \\ \in [u-q, t+q] \setminus \{u\} \\ u \leq t_{m-1}+q}} L_{\max}^2 \frac{(\Lambda^2)^{m-1}}{n^m} \left(\left(\Delta_{\mathsf{span}}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 + \left(\Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 \right) \Bigg]$$

$$+ \mathbb{E}\Bigg[\sum_{\substack{l_0, l_1, \cdots, l_{m-1} \\ \in [u-q, t+q] \setminus \{u\} \\ u \leq t_{m-1}+q}} 8L_{\max}^2 \frac{L_{\mathsf{res}}^2 (\Lambda^2)^{m-1}}{\Gamma^2 n^m}$$

$$\cdot \left(\left(\Delta_{\mathsf{span}}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 + \left(\Delta_{\max}^{t_{m-1}, R_{m-1}} x_{k_u}^{\pi, u}\right)^2 \right) \Bigg]$$

$$\leq \mathbb{E}\left[24 L_{\max}^2 \frac{(\Lambda^2)^m (4q)^m}{n^m} \left(\left(\overline{\Delta}_{\mathsf{span}} x_{k_u}^{\pi, u}\right)^2 + \left(\Delta x_{k_u}^{\pi, u}\right)^2 \right) \right].$$

This yields the term (23) on the RHS of Lemma 21, which concludes the proof. $\qquad \square$

A.5 The Claims from Section 4.2.4

Recall that $\pi(k) \equiv \pi(k, t)$. The following observation will be useful.

Observation 2. *For $u < t$, $\Delta_{\max}^{t, \emptyset} x_{k_u}^{\pi(k_t), u} \geq \Delta_{\max}^{t, \emptyset} x_{k_u}^{\pi(k_u), u}$, $\Delta_{\min}^{t, \emptyset} x_{k_u}^{\pi(k_t), u} \leq \Delta_{\min}^{t, \emptyset} x_{k_u}^{\pi(k_u), u}$, and thus $\Delta_{\mathsf{span}}^{t, \emptyset} x_{k_u}^{\pi(k_u), u} \leq \Delta_{\mathsf{span}}^{t, \emptyset} x_{k_u}^{\pi(k_t), u}$.*

Proof. This is immediate from the fact that replacing k_u by k_t on path π allows more choices of input in calculating the update to x_{k_u}. $\qquad \square$

Proof of Claim 5. [Bounding Term B] We begin by noting that

$$x_{k_s}^{\pi(k_s), t} = x_{k_s}^{\pi, t-2q} + \sum_{\substack{t-2q \leq u < t \\ \& k_u = k_s}} \Delta x_{k_u}^{\pi(k_s), u} = x_{k_s}^{\pi, t-2q} + \sum_{\substack{t-2q \leq u < t \\ \& k_u = k_s}} \Delta x_{k_u}^{\pi(k_u), u}, \text{ and}$$

$$x_{k_s}^{\pi(k_t), t} = x_{k_s}^{\pi, t-2q} + \sum_{\substack{t-2q \leq u < t \\ \& k_u = k_s}} \Delta x_{k_u}^{\pi(k_t), u}.$$

Note that as $t - 2q \leq u$, and as the updates $\mathcal{U}_r$ are fixed for $r < t - 4q$ on the RHS of the following expression, we have $\Delta x_{k_u}^{\pi(k_u),u} \in \left[\Delta_{\min}^{t,\emptyset} x_{k_u}^{\pi(k_u),u}, \Delta_{\max}^{t,\emptyset} x_{k_u}^{\pi(k_u),u} \right]$. Also, as $u < t$, and by Observation 2, the range $\left[\Delta_{\min}^{t,\emptyset} x_{k_u}^{\pi(k_u),u}, \Delta_{\max}^{t,\emptyset} x_{k_u}^{\pi(k_u),u} \right] \subseteq \left[\Delta_{\min}^{t,\emptyset} x_{k_u}^{\pi(k_t),u}, \Delta_{\max}^{t,\emptyset} x_{k_u}^{\pi(k_t),u} \right]$. Further note that $\Delta x_{k_u}^{\pi(k_t),u}$ lies in the same range, $\left[\Delta_{\min}^{t,\emptyset} x_{k_u}^{\pi(k_t),u}, \Delta_{\max}^{t,\emptyset} x_{k_u}^{\pi(k_t),u} \right]$.

Therefore,

$$
\mathbb{E}\left[\frac{2}{3n^2} \sum_{\substack{t-2q \leq s < t \\ \& s = \mathrm{prev}(t,k_s)}} \sum_{k_t=1}^{n} \Gamma \cdot \left(x_{k_s}^{\pi(k_s),t} - x_{k_s}^{\pi(k_t),t} \right)^2 \right]
$$

$$
\leq \frac{2\Gamma}{3n^2} \cdot \mathbb{E}\left[\sum_{\substack{t-2q \leq s < t \\ \& s = \mathrm{prev}(t,k_s)}} \sum_{k_t=1}^{n} \left[\sum_{\substack{t-2q \leq u < t \\ \& k_u = k_s}} \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_u}^{\pi(k_t),u} \right) \right]^2 \right]
$$

$$
\leq \frac{2\Gamma \cdot 2q}{3n^2} \cdot \mathbb{E}\left[\sum_{k_t=1}^{n} \sum_{t-2q \leq u < t} \left(\Delta_{\mathsf{span}}^{t,\emptyset} x_{k_u}^{\pi(k_t),u} \right)^2 \right] \qquad \text{(by the Cauchy-Schwarz inequality)}
$$

$$
\leq \frac{4q}{3n} \sum_{u \in [t-2q,t-1]} \Gamma \cdot (\mathcal{D}_u)^2 \;=\; \frac{\nu_1}{15q} \sum_{s \in [t-2q,t-1]} \Gamma \cdot (\mathcal{D}_s)^2 \qquad \text{(recall that } \nu_1 = \frac{20q^2}{n}\text{)}
$$

$\square$

Proof of Claim 3. [Bounding Term A] We begin by observing:

$$
\mathbb{E}\left[\sum_{\substack{t-2q \leq s < t \\ \& \ s = \mathrm{prev}(t,k_s)}} \sum_{k_t=1}^{n} \frac{1}{2n^2\Gamma} \cdot \left(\tilde{g}_{k_s}^{\pi(k_t),s} - g_{k_s}^{\pi(k_t),t} \right)^2 \right]
$$

$$
\leq \mathbb{E}\left[\sum_{t-2q \leq s < t} \sum_{k_t=1}^{n} \frac{1}{2n^2\Gamma} \cdot \left(\tilde{g}_{k_s}^{\pi(k_t),s} - g_{k_s}^{\pi(k_t),t} \right)^2 \right]. \tag{44}
$$

Note that $\tilde{g}_{k_s}^{\pi(k_t),s}$ and $g_{k_s}^{\pi(k_t),t}$ are on the same path $\pi(k_t)$. The difference of these two values is bounded by a sum of terms of the form $L_{k_{l_0} k_s} \cdot \left| \Delta x_{k_{l_0}}^{\pi(k_t),l_0} \right|$, where l_0 needs to satisfy two conditions: 1. $l_0 \in [s - 2q, \max\{s + q, t - 1\}]$; 2. If $l_0 \geq t$, $\mathcal{U}_{l_0}$ commits before $\mathcal{U}_s$ reads coordinate $x_{k_{l_0}}$. As $t - 2q \leq s < t$, if l_0 satisfies these conditions, $\left| \Delta x_{k_{l_0}}^{\pi(k_t),l_0} \right| \leq \max \left\{ \left| \Delta_{\min}^{t,\emptyset} x_{k_{l_0}}^{\pi(k_t),l_0} \right|, \left| \Delta_{\max}^{t,\emptyset} x_{k_{l_0}}^{\pi(k_t),l_0} \right| \right\} \leq$

$\Delta_{\mathsf{var}}^{t,\emptyset} x_{kl_0}^{\pi(k_t),l_0}$.   Thus

$$\mathbb{E}\left[\sum_{t-2q\leq s<t}\sum_{k_t=1}^{n}\frac{1}{2n^2\Gamma}\cdot\left(\tilde{g}_{k_s}^{\pi(k_t),s}-g_{k_s}^{\pi(k_t),t}\right)^2\right]$$

$$\leq\mathbb{E}\left[\sum_{t-2q\leq s<t}\sum_{k_t=1}^{n}\frac{1}{n^2\Gamma}\left(\left(\sum_{l_0\in[s-2q,\max\{s+q,t-1\}]\setminus\{s\}}L_{k_{l_0}k_s}\Delta_{\mathsf{var}}^{t,\emptyset}x_{kl_0}^{\pi(k_t),l_0}\right)^2\right.\right.$$

$$\left.\left.+\left(L_{k_sk_s}\Delta x_{k_s}^{\pi(k_t),s}\right)^2\right)\right]$$

$$\leq\mathbb{E}\left[\sum_{t-2q\leq s<t}\frac{1}{n\Gamma}\left(\left(\sum_{l_0\in[t-4q,t+q]\setminus\{s\}}L_{k_{l_0}k_s}\Delta_{\mathsf{var}}^{t,\emptyset}x_{kl_0}^{\pi,l_0}\right)^2+\left(L_{k_sk_s}\Delta x_{k_s}^{\pi,s}\right)^2\right)\right]$$

$$\text{(recall that }\pi=\pi(k_t).)$$

Then, using Lemma 11 for the first inequality, $\Gamma\geq L_{\max}$ for the second inequality, and Lemma 10 to bound $\Gamma\left(\mathcal{D}_t\right)^2$ for the third inequality, yields

$$\mathbb{E}\left[\sum_{\substack{t-2q\leq s<t\\ \&\ s=\mathsf{prev}(t,k_s)}}\sum_{k_t=1}^{n}\frac{1}{2n^2\Gamma}\cdot\left(\tilde{g}_{k_s}^{\pi(k_t),\mathsf{prev}(t,k_s)}-g_{k_s}^{\pi(k_t),t}\right)^2\right]$$

$$\leq\frac{\nu_3\Gamma}{qn}\sum_{t-2q\leq s<t}\sum_{r\in[t-7q,t+q]\setminus\{s\}}\left[\left(\mathcal{D}_r\right)^2+\left(\Delta_r^X\right)^2\right]+\sum_{t-2q\leq s<t}\frac{\nu_4\Gamma}{n}\left[\left(\mathcal{D}_s\right)^2+\left(\Delta_s^X\right)^2\right]$$

$$+\sum_{t-2q<s<t}\frac{L_{\max}^2}{n\Gamma^2}\cdot\Gamma\left(\Delta_s^X\right)^2\quad\text{(as }L_{k_sk_s}\leq L_{\max})$$

$$\leq\frac{2(\nu_3+\nu_4)}{n}\cdot\Gamma\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[\left(\mathcal{D}_s\right)^2+\left(\Delta_s^X\right)^2\right]+\frac{2\nu_3}{n}\cdot\Gamma\left[\left(\mathcal{D}_t\right)^2+\left(\Delta_t^X\right)^2\right]$$

$$+\frac{\Gamma}{n}\sum_{s\in[t-2q,t-1]}\left(\Delta_s^X\right)^2$$

$$\leq\frac{2(\nu_3+\nu_4)}{n}\cdot\Gamma\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[\left(\mathcal{D}_s\right)^2+\left(\Delta_s^X\right)^2\right]+\frac{\Gamma}{n}\sum_{s\in[t-2q,t-1]}\left(\Delta_s^X\right)^2$$

$$+\frac{2\nu_3}{n}\cdot\Gamma\left(\frac{\nu_1}{q}+\frac{\nu_2}{q}\right)\sum_{s\in[t-5q,t+q]\setminus\{t\}}\left[\left(\mathcal{D}_s\right)^2+\left(\Delta_s^X\right)^2\right]+\frac{2\nu_3}{n}\cdot\Gamma\left(\Delta_t^X\right)^2.$$

$\square$

Proof of Claim 6. [Bounding Term C]

$$\mathbb{E}\left[\frac{\Gamma}{2n^2}\sum_{\substack{t-2q\leq s<t\\ \&s=\mathsf{prev}(t,k_s)}}\sum_{k_t=1}^{n}\left(\Delta x_{k_s}^{\pi(k_t),s}\right)^2\right]\leq\mathbb{E}\left[\frac{\Gamma}{2n}\sum_{t-2q\leq s<t}\left(\Delta x_{k_s}^{\pi,s}\right)^2\right]$$

$$\leq\frac{\Gamma}{2n}\sum_{t-2q\leq s\leq t-1}\left(\Delta_s^X\right)^2.$$

$\square$

Proof of Claim 4. [Bounding Term D] We bound the term

$\mathbb{E}\left[\frac{2}{3\Gamma n^2}\sum_{k=1}^{n}\sum_{k_t=1}^{n}\cdot\left(g_k^{\pi(k),t}-g_k^{\pi(k_t),t}\right)^2\right]$. We will use the term $g_k^{\pi(k),t-2q}$ as an intermediary to allow us to compare values on two different paths, as follows.

$$\frac{2}{3}\left(g_k^{\pi(k),t}-g_k^{\pi(k_t),t}\right)^2 \le \frac{4}{3}\left(g_k^{\pi(k),t}-g_k^{\pi(k),t-2q}\right)^2 + \frac{4}{3}\left(g_k^{\pi(k),t-2q}-g_k^{\pi(k_t),t}\right)^2$$
$$\le \frac{4}{3}\left(g_k^{\pi(k),t}-g_k^{\pi(k),t-2q}\right)^2 + \frac{4}{3}\left(g_k^{\pi(k_t),t-2q}-g_k^{\pi(k_t),t}\right)^2, \qquad (45)$$

as $g_k^{\pi(k),t-2q}=g_k^{\pi(k_t),t-2q}$ since the gradients at $x^{\pi,t-2q}$ do not depend on update $\mathcal{U}_t$.

For the first term on the RHS of (45), we apply Lemma 11 for the third inequality, and then apply Lemma 10 to bound the $(\mathcal{D}_t)^2$ term which was generated by Lemma 11, as follows.

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\frac{4}{3}\left(g_k^{\pi(k),t}-g_k^{\pi(k),t-2q}\right)^2\right] \le \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\frac{4}{3}\left(\sum_{t-2q\le l_0\le t-1}L_{k_{l_0}k}\Delta x_{k_{l_0}}^{\pi(k),l_0}\right)^2\right]$$

$$\le \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\frac{4}{3}\left(\sum_{t-2q\le l_0\le t-1}L_{k_{l_0}k}\Delta_{\text{var}}x_{k_{l_0}}^{\pi(k),l_0}\right)^2\right]$$

$$\le \frac{4\nu_3\Gamma^2}{3q}\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[(\mathcal{D}_s)^2+(\Delta_s^X)^2\right] + \frac{4\nu_4\Gamma^2}{3}\left[(\mathcal{D}_t)^2+(\Delta_t^X)^2\right]$$

$$\le \frac{4\nu_3\Gamma^2}{3q}\sum_{s\in[t-7q,t+q]\setminus\{t\}}\left[(\mathcal{D}_s)^2+(\Delta_s^X)^2\right]$$

$$+ \frac{4\nu_4\Gamma^2}{3}\cdot\left(\frac{\nu_1}{q}+\frac{\nu_2}{q}\right)\sum_{s\in[t-5q,t+q]\setminus\{t\}}\left[(\mathcal{D}_s)^2+(\Delta_s^X)^2\right] + \frac{4\nu_4\Gamma^2}{3}\cdot(\Delta_t^X)^2. \qquad (46)$$

For the second term on the RHS of (45), for $t-2q \le s \le t-1$, as $\Delta x_{k_s}^{\pi(k_t),s}$ and $\Delta_{\max}^{t,\{t\}}x_{k_s}^{\pi(k_t),s}$ are in the range $\left[\Delta_{\min}^{t,\emptyset}x_{k_s}^{\pi(k_t),s},\Delta_{\max}^{t,\emptyset}x_{k_s}^{\pi(k_t),s}\right]$, by recentering and by the Cauchy-Schwarz inequality,

$$\frac{4}{3}\left(g_k^{\pi(k_t),t-2q}-g_k^{\pi(k_t),t}\right)^2 = \frac{4}{3}\left(\sum_{s\in[t-2q,t-1]}L_{k_s,k}\Delta x_{k_s}^{\pi(k_t),s}\right)^2$$

$$= \frac{16}{3}q\sum_{s\in[t-2q,t-1]}L_{k_s,k}^2\left(\left(\Delta_{\text{span}}^{t,\emptyset}x_{k_s}^{\pi(k_t),s}\right)^2+\left(\Delta_{\max}^{t,\{t\}}x_{k_s}^{\pi(k_t),s}\right)^2\right).$$

This yields

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{k=1}^{n}\sum_{k_t=1}^{n}\frac{4}{3}\left(g_k^{\pi(k_t),t-2q}-g_k^{\pi(k_t),t}\right)^2\right]$$

$$\le \mathbb{E}\left[\frac{1}{n^2}\sum_{k_t=1}^{n}\sum_{k=1}^{n}\frac{16}{3}q\sum_{s\in[t-2q,t-1]}L_{k_s,k}^2\left(\left(\Delta_{\text{span}}^{t,\emptyset}x_{k_s}^{\pi(k_t),s}\right)^2+\left(\Delta_{\max}^{t,\{t\}}x_{k_s}^{\pi(k_t),s}\right)^2\right)\right]$$

$$\le \mathbb{E}\left[\frac{1}{n^2}\sum_{k_t=1}^{n}\frac{16}{3}q\sum_{s\in[t-2q,t-1]}L_{\overline{\text{res}}}^2\left(\left(\Delta_{\text{span}}^{t,\emptyset}x_{k_s}^{\pi(k_t),s}\right)^2+\left(\Delta_{\max}^{t,\{t\}}x_{k_s}^{\pi(k_t),s}\right)^2\right)\right].$$

50

By Lemma 9, $\left(\Delta_{\text{span}}^{t,\emptyset} x_{k_s}^{\pi(k_t),s}\right)^2 \leq \left(\overline{\Delta}_{\text{span}} x_{k_s}^{\pi(k_t),s}\right)^2$ and $\left(\Delta_{\max}^{t,\{t\}} x_{k_s}^{\pi(k_t)}\right)^2 \leq 2\left(\overline{\Delta}_{\text{span}} x_{k_s}^{\pi(k_t),s}\right)^2 +$
$2\left(\Delta x_{k_s}^{\pi(k_t),s}\right)^2$. Thus,

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{k=1}^{n}\sum_{k_t=1}^{n}\frac{4}{3}\left(g_k^{\pi(k_t),t-2q} - g_k^{\pi(k_t),t}\right)^2\right] \leq \frac{16qL_{\overline{\text{res}}}^2}{n}\sum_{s\in[t-2q,t-1]}\left(\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right)$$

$$\leq \frac{2\nu_2\Gamma^2}{3q}\sum_{s\in[t-2q,t-1]}\left(\left(\mathcal{D}_s\right)^2 + \left(\Delta_s^X\right)^2\right). \qquad (47)$$

Combining (45), (46) and (47) yields the result. □

Claim 7. *The term G in* (13) *is bounded by* $\frac{1}{q}\left[\frac{2r}{3} + \frac{3r^2}{1280} + \frac{9r^3}{25600} + \frac{3r^2}{1-r} + \frac{r^3}{426(1-r)} + \frac{r^4}{2844(1-r)}\right]$.

Proof. Recall that $r = \frac{160q^2}{n}\cdot\left(\frac{L_{\overline{res}}^2}{\Gamma^2}+1\right)$ (see the second paragraph of Section 4.2.3). As stated before Lemma 10, $\nu_1 = \frac{20q^2}{n}$, $\nu_2 = \frac{24q^2L_{\overline{res}}^2}{n\Gamma^2}$, and as stated before Lemma 11, $\nu_3 = \frac{3}{16}\left(\frac{r^2}{1-r}+r\right)$, $\nu_4 = \frac{6r}{1-r}$. Also, note that $\frac{1}{n} \leq \frac{q}{n}$ as $q \geq 1$, $\nu_1 + \nu_2 \leq \frac{3r}{20}$, and $\nu_2 + \frac{24q^2}{n} = \frac{3r}{20}$.

$$\max\left\{\frac{\nu_1}{15q}, \frac{3}{2n}\right\} + \frac{8\nu_2}{11q} + \frac{2(\nu_3+\nu_4)}{n} + \frac{7\nu_3}{3q} + \left(\frac{2\nu_3}{nq} + \frac{7\nu_4}{3q}\right)(\nu_1+\nu_2)$$

$$\leq \frac{1}{q}\left[\max\left\{\frac{\nu_1}{15}, \frac{3q^2}{2n}\right\} + \frac{8\nu_2}{11} + \left(\frac{3}{4}\left(\frac{r^2}{1-r}+r\right) + \frac{24r}{1-r}\right)\frac{r}{320} + \frac{7}{16}\left(\frac{r^2}{1-r}+r\right)\right.$$

$$\left. + \left(\frac{3}{4}\left(\frac{r^2}{1-r}+r\right)\frac{r}{320} + \frac{14r}{1-r}\right)\frac{3r}{20}\right]$$

$$\leq \frac{1}{q}\left[r\left(\frac{1}{9}+\frac{7}{16}\right) + r^2\frac{3}{1280} + r^3\frac{9}{25600} + \frac{r^2}{1-r}\left(\frac{3}{40}+\frac{7}{16}+\frac{21}{10}\right)\right.$$

$$\left. + \frac{r^3}{1-r}\left(\frac{3}{1280}\right) + \frac{r^4}{1-r}\left(\frac{9}{25600}\right)\right]$$

$$\leq \frac{1}{q}\left[\frac{2r}{3} + \frac{3r^2}{1280} + \frac{9r^3}{25600} + \frac{3r^2}{1-r} + \frac{r^3}{426(1-r)} + \frac{r^4}{2844(1-r)}\right].$$

□

A.6 Proofs for the Amortized Analysis, Section 4.3: Theorem 3 and Lemma 13

The following lemma is key to the demonstration of progress in both the strongly convex and convex cases.

For any $t \geq 1$, we define:
$$\text{PRG}(t) \triangleq \sum_{k=1}^{n}\widehat{W}_k(\nabla_k f(x^t), x_k^t).$$

We will use the following lemma from [25, Lemmas 4,6]. The version we present here is slightly different from the one in [25], but the proofs are essentially the same.

Lemma 24 ([25, Lemmas 4,6]).
(a) Suppose that f, F are strongly convex with parameters $\mu_f, \mu_F > 0$ respectively, and suppose that $\Gamma \geq \mu_f$. Then
$$\text{PRG}(t) \geq \frac{\mu_F}{\mu_F + \Gamma - \mu_f}\cdot F(x^t).$$

(b) Suppose that f, F are convex functions. Suppose that $\mathcal{R} := \min_{x^* \in X^*} \|x^t - x^*\| < \infty$. Then

$$\mathsf{PRG}(t) \geq \min\left\{\frac{1}{2}, \frac{F(x^t)}{2\Gamma\mathcal{R}^2}\right\} \cdot F(x^t).$$

Proof of Theorem 3. We begin by showing (i). By assumption (c) and Lemma 24(a),

$$
\begin{aligned}
H(t) - H(t+1) &\geq \left[\frac{\alpha}{n} \cdot \mathbb{E}\big[\mathsf{PRG}(t)\big] + \frac{\beta}{n} \cdot A^+(t)\right] \\
&\geq \left[\frac{\alpha}{n} \cdot \frac{\mu_F}{\mu_F + \Gamma - \mu_f} \cdot \mathbb{E}\big[F(x^t)\big] + \frac{\beta}{n} \cdot A^+(t)\right] \geq \delta \cdot H(t),
\end{aligned}
$$

where $\delta \triangleq \min\left\{\frac{\alpha}{n} \cdot \frac{\mu_F}{\mu_F + \Gamma - \mu_f}, \frac{\beta}{n}\right\}$.

Thus $H(t+1) \leq (1-\delta)\, H(t)$ for all $t \geq 1$. Iterating the above inequality T times yields $H(T+1) \leq (1-\delta)^T H(1)$.

To finish the proof note that since $A^+(1) = 0$ and $A^-(1) \geq 0$, $H(1) \leq F(x^1)$.

Now we show (ii). By the second assumption, Lemma 24 and the fact that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$,

$$
\begin{aligned}
&H(t) - H(t+1) \\
&\geq \left[\frac{\alpha}{n} \cdot \mathbb{E}\big[\mathsf{PRG}(t)\big] + \frac{\beta}{n} \cdot A^+(t)\right] \geq \left[\frac{\alpha}{n} \cdot \min\left\{\frac{1}{2}, \frac{\mathbb{E}[F(x^t)]}{2\Gamma\mathcal{R}^2}\right\} \cdot \mathbb{E}[F(x^t)] + \frac{\beta}{n} \cdot A^+(t)\right].
\end{aligned}
$$

We consider two cases:

- If $\mathbb{E}[F(x^t)] \leq A^+(t)$, then $A^+(t) \geq \frac{H(t)}{2}$, thus

$$\frac{\alpha}{n} \cdot \min\left\{\frac{1}{2}, \frac{\mathbb{E}[F(x^t)]}{2\Gamma\mathcal{R}^2}\right\} \cdot \mathbb{E}[F(x^t)] + \frac{\beta}{n} \cdot A^+(t) \geq \frac{\beta}{2n} \cdot H(t).$$

- If $\mathbb{E}[F(x^t)] > A^+(t)$, then $\mathbb{E}[F(x^t)] > \frac{H(t)}{2}$, thus

$$\frac{\alpha}{n} \cdot \min\left\{\frac{1}{2}, \frac{\mathbb{E}[F(x^t)]}{2\Gamma\mathcal{R}^2}\right\} \cdot \mathbb{E}[F(x^t)] + \frac{\beta}{n} \cdot A^+(t) > \frac{\alpha}{2n} \cdot \min\left\{\frac{1}{2}, \frac{H(t)}{4\Gamma\,\mathcal{R}^2}\right\} \cdot H(t).$$

Since H is a decreasing function, $H(t) \leq H(1) \leq F(x^1)$. Thus, unconditionally,

$$
\begin{aligned}
\frac{\alpha}{n} \cdot \min\left\{\frac{1}{2}, \frac{\mathbb{E}[F(x^t)]}{2\Gamma\mathcal{R}^2}\right\} \cdot \mathbb{E}[F(x^t)] + \frac{\beta}{n} \cdot A^+(t) &\geq \min\left\{\frac{\beta}{2n}, \frac{\alpha}{4n}, \frac{\alpha \cdot H(t)}{8n\Gamma\mathcal{R}^2}\right\} \cdot H(t) \\
&\geq \min\left\{\frac{\beta}{2n \cdot F(x^1)}, \frac{\alpha}{4n \cdot F(x^1)}, \frac{\alpha}{8n\Gamma\mathcal{R}^2}\right\} \cdot H(t)^2.
\end{aligned}
$$

Note that the term $\min\left\{\frac{\beta}{2n\ F(x^1)}, \frac{\alpha}{4nF(x^1)}, \frac{\alpha}{8n\Gamma\mathcal{R}^2}\right\}$ is independent of t. We denote it by ε. Thus, $H(t) - H(t+1) \geq \varepsilon\, H(t)^2$. Dividing both sides by $H(t) \cdot H(t+1)$ yields

$$\frac{1}{H(t+1)} - \frac{1}{H(t)} \geq \varepsilon \frac{H(t)}{H(t+1)} \geq \varepsilon.$$

Iterating the above inequality T times yields $\quad \dfrac{1}{H(T+1)} - \dfrac{1}{H(1)} \geq \varepsilon T$,

and hence $\quad \dfrac{1}{H(T+1)} \geq \varepsilon T + \dfrac{1}{H(1)} \geq \varepsilon T + \dfrac{1}{F(x^1)}. \quad$ (since $H(1) \leq F(x^1)$)

(ii) follows by taking the reciprocal on both sides of the above inequality. $\qquad \square$

Proof of Lemma 13. By calculation,

$$
A^+(t) - A^+(t+1) = \sum_{s=t-7q}^{t-1} \frac{1}{\left(1 - \frac{1}{3n}\right)} \left[c(\mathcal{D}_s)^2 + c(\Delta_s^X)^2 \right]
$$

$$
+ \sum_{s=t-7q+1}^{t-1} \sum_{v=t+1}^{s+7q} \frac{1}{3n} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t+1}} \left[c(\mathcal{D}_s)^2 + c(\Delta_s^X)^2 \right]
$$

$$
- \sum_{v=t+1}^{t+7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t}} \left[c(\mathcal{D}_t)^2 + c(\Delta_t^X)^2 \right]
$$

$$
= \frac{1}{3n} A^+(t) + \sum_{s=t-7q}^{t-1} \left[c(\mathcal{D}_s)^2 + c(\Delta_s^X)^2 \right]
$$

$$
- \sum_{v=t+1}^{t+7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t}} \left[c(\mathcal{D}_t)^2 + c(\Delta_t^X)^2 \right]
$$

$$
A^-(t+1) - A^-(t) = \sum_{v=t+1}^{t+q} \left[c(\mathcal{D}_v)^2 + c\left(\Delta_v^X\right)^2 \right] - \sum_{s=t-q}^{t-1} \left[c(\mathcal{D}_t)^2 + c\left(\Delta_t^X\right)^2 \right].
$$

Therefore

$$
\left[\left(1 - \frac{1}{3n}\right) A^+(t) - A^-(t) \right] - \left[A^+(t+1) - A^-(t+1) \right]
$$

$$
= \sum_{s=t-7q}^{t-1} \left[c(\mathcal{D}_s)^2 + c(\Delta_s^X)^2 \right] - \sum_{v=t+1}^{t+7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t}} \left[c(\mathcal{D}_t)^2 + c(\Delta_t^X)^2 \right]
$$

$$
+ \sum_{v=t+1}^{t+q} \left[c(\mathcal{D}_v)^2 + c(\Delta_v^X)^2 \right] - \sum_{s=t-q}^{t-1} \left[c(\mathcal{D}_t)^2 + c(\Delta_t^X)^2 \right]
$$

$$
= \sum_{s \in [t-7q, t+q]\setminus\{t\}} \left[c(\mathcal{D}_s)^2 + c(\Delta_s^X)^2 \right]
$$

$$
- \left(\sum_{v=t+1}^{t+7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^{v-t}} + q \right) \left[c(\mathcal{D}_t)^2 + c(\Delta_t^X)^2 \right]. \tag{48}
$$

In order to achieve (17), we compare the coefficients of each of the terms $c(\mathcal{D}_t)^2$, $c\left(\Delta_t^X\right)^2$, $c(\mathcal{D}_s)^2$, $c\left(\Delta_s^X\right)^2$ in (17) and (48). Since $c = \varpi + (\gamma + \varpi)\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)\Gamma$ the coefficient of $(\mathcal{D}_s)^2$ and $(\Delta_s^X)^2$ in (48) is at least as big as in (17). Therefore, it suffices to have the coefficients of $(\mathcal{D}_t)^2$ and $(\Delta_t^X)^2$ satisfy the following inequalities.

$$
\gamma \geq \frac{c}{\Gamma} \cdot \left[\sum_{i=1}^{7q} \frac{1}{\left(1 - \frac{1}{3n}\right)^i} + q \right]
$$

$$
= \frac{1}{q} \left[q\varpi + (\gamma + \varpi)(\nu_1 + \nu_2) \right] \cdot \left[3n \left(\frac{1}{\left(1 - \frac{1}{3n}\right)^{7q+1}} - \frac{1}{1 - \frac{1}{3n}} \right) + q \right]
$$

and
$$\varrho - \varpi \geq \frac{1}{q}\left[q\varpi + (\gamma + \varpi)(\nu_1 + \nu_2)\right] \cdot \left[3n\left(\frac{1}{\left(1 - \frac{1}{3n}\right)^{7q+1}} - \frac{1}{1 - \frac{1}{3n}}\right) + q\right].$$

If $7q < 3n - 2$, then by the fact that $(1 + x)^s \leq 1 + \frac{sx}{1-(s-1)x}$ for any $x < \frac{1}{s-1}$ and $s \geq 1$,

$$3n\left(\frac{1}{\left(1 - \frac{1}{3n}\right)^{7q+1}} - \frac{1}{1 - \frac{1}{3n}}\right) + q \leq 3n\left[\left(1 + \frac{1}{3n - 1}\right)^{7q+1} - \frac{3n}{3n - 1}\right] + q$$

$$\leq 3n\left[1 + \frac{(7q + 1)\left(\frac{1}{3n-1}\right)}{1 - \frac{7q}{3n-1}} - 1\right] + q$$

$$\leq 3n\left(\frac{7q + 1}{3n - 1 - 7q}\right) + q \leq 14q + 2 + q,$$

if $\frac{3n}{3n-1-7q} \leq 2$, i.e., if $7q \leq n - 1$.

Then it suffices that
$$\gamma \geq \frac{1}{q}\left[q\varpi + (\gamma + \varpi)(\nu_1 + \nu_2)\right](15q + 2)$$

and
$$\varrho - \varpi \geq \frac{1}{q}\left[q\varpi + (\gamma + \varpi)(\nu_1 + \nu_2)\right](15q + 2).$$

Recall that $\nu_1 = \frac{20q^2}{n} \leq \frac{r}{8}$ and $\nu_2 = \frac{24q^2 L_{\text{res}}^2}{n\Gamma^2} \leq \frac{r}{6}$, $\varrho = \frac{1}{8} - \frac{15r}{1-r}$ (see the first line of Section 4.3), and $\varpi = \frac{1}{q}\left[\frac{2r}{3} + \frac{3r^2}{1280} + \frac{9r^3}{25600} + \frac{3r^2}{1-r} + \frac{r^3}{426(1-r)} + \frac{r^4}{2844(1-r)}\right]$. One choice of values that suffices is $\gamma = \varrho - \varpi$ and $r \leq \frac{1}{225}$. $\qquad\square$

A Table of Definitions and Parameters

Notation / Parameter	Definition / Description	First Appearance
$F : \mathbb{R}^n \to \mathbb{R}$ $f : \mathbb{R}^n \to \mathbb{R}$ $\Psi_k : \mathbb{R} \to \mathbb{R}$	$F(x) = f(x) + \sum_{k=1}^{n} \Psi_k(x_k)$ F is the convex function we want to minimize.	Abstract
e_j	the unit vector along coordinate j	Definition 1
L_{jk}	$\|\nabla_k f(x + r \cdot e_j) - \nabla_k f(x)\| \leq L_{jk} \cdot \|r\|$	
L_{res}	$\|\nabla f(x + r \cdot e_j) - \nabla f(x)\| \leq L_{\text{res}} \cdot \|r\|$	
$L_{\overline{\text{res}}}$	$L_{\overline{\text{res}}} \triangleq \max_k \left(\sum_{j=1}^{n} (L_{kj})^2 \right)^{1/2}$	
$L_{\max}$	$L_{\max} \triangleq \max_{j,k} L_{jk}$ if f is twice differentiable, then $L_{\max} = \max_j L_{jj}$	
μ_f, μ_F	strong convexity parameters of f, F $f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{1}{2}\mu_f \|y - x\|^2$ $F(y) - F(x) \geq \langle \nabla F(x), y - x \rangle + \frac{1}{2}\mu_F \|y - x\|^2$	Definition 2
$\mathcal{U}_t$	the t-th update (in the SCC order)	Near Eqn. (1)
Γ	parameter used in the update rule	
$W_j(d, g, x)$	$W_j(d, g, x) \triangleq gd + \Gamma d^2/2 + \Psi_j(x + d) - \Psi_j(x)$	
$\widehat{W}_j(g, x)$	$\widehat{W}_j(g, x) \triangleq \max_d W(d, g, x)$	
$\widehat{d}_j(g, x)$	$\widehat{d}_j(g, x) \triangleq \arg\max_d W(d, g, x)$	
asynchronous update rule	$x_j^{t+1} \leftarrow x_j^t + \widehat{d}_j(\tilde{g}_j, x_j^t)$, where $\tilde{g}_j$ is the (inaccurate) measured gradient.	
q	the updates that can interfere with the update at time t in the ST order are those that commit at times $t + 1, t + 2, \cdots, t + q$	Assumption 4
$\Delta x_{k_t}^{\pi, t}$	the increment computed by the t-th update on path π	Near Lemma 1
k_t	the coordinate that is updated at time t	Beginning of Section 3
g_k^t	$g_k^t \triangleq \nabla_k f(x^t)$ accurate gradient along coordinate k at time t	
π	a root-to-leaf path in the branching tree	Beginning of Section 3.2
$\pi(k, t)$	the root-to-leaf path with time t coordinate on path π replaced by coordinate k	
$\pi(k_t, t)$	$\pi(k_t, t) = \pi$	
$\bullet^{\pi, t}$	for any variable $\bullet$, $\bullet^{\pi, t}$ denotes its value at time t along the path π	

Notation / Parameter	Definition / Description	First Appearance
$A^{\pi,u}$	$A^{\pi,u} = \{r \mid u - 4q \le r < u - 2q$ and $\mathcal{U}_r$ has committed before some $\mathcal{U}_p$ for $p \le u - 4q - 1\}$	Section 4.2.1
$\Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s}$	see the table in the next page	
$\Delta_{\max}^{u,R} x_{k_s}^{\pi,s}$	$\Delta_{\max}^{u,R} x_{k_s}^{\pi,s} \triangleq \max_S \Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s}$	
$\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s}$	$\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s} \triangleq \Delta_{\max}^{u,R} x_{k_s}^{\pi,s} - \Delta_{\min}^{u,R} x_{k_s}^{\pi,s}$	
$\Delta_{\max}^{u} x_{k_s}^{\pi,s}$	$\Delta_{\max}^{u} x_{k_s}^{\pi,s} \triangleq \Delta_{\max}^{u,\emptyset} x_{k_s}^{\pi,s}$	
$\Delta_{\mathsf{var}}^{u,R} x_{k_s}^{\pi,s}$	$\Delta_{\mathsf{var}}^{u,R} x_{k_s}^{\pi,s} \triangleq \max\left\{\Delta_{\mathsf{span}}^{u,R} x_{k_s}^{\pi,s}, \left|\Delta_{\max}^{u,R} x_{k_s}^{\pi,s}\right|, \left|\Delta_{\min}^{u,R} x_{k_s}^{\pi,s}\right|\right\}$	Section 4.2.1
$\overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s} \triangleq \max_{s-q \le t \le s} \Delta_{\max}^{t,R} x_{k_s}^{\pi,s}$	
$\overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s} - \overline{\Delta}_{\min}^{R} x_{k_s}^{\pi,s}$	
$\overline{\Delta}_{\mathsf{var}}^{R} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\mathsf{var}}^{R} x_{k_s}^{\pi,s} \triangleq \max\left\{\overline{\Delta}_{\mathsf{span}}^{R} x_{k_s}^{\pi,s}, \left|\overline{\Delta}_{\max}^{R} x_{k_s}^{\pi,s}\right|, \left|\overline{\Delta}_{\min}^{R} x_{k_s}^{\pi,s}\right|\right\}$	
$\overline{\Delta}_{\max} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\max} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\max}^{\emptyset} x_{k_s}^{\pi,s}$	
$\overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\mathsf{span}} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\mathsf{span}}^{\emptyset} x_{k_s}^{\pi,s}$	Section 4.2.1
$\overline{\Delta}_{\mathsf{var}} x_{k_s}^{\pi,s}$	$\overline{\Delta}_{\mathsf{var}} x_{k_s}^{\pi,s} \triangleq \overline{\Delta}_{\mathsf{var}}^{\emptyset} x_{k_s}^{\pi,s}$	
$\widetilde{g}_{\max,k_s}^{u,R,S,\pi,s}$	see the next table	
$\widetilde{g}_{\max,k_s}^{u,R,\pi,s}$	$\widetilde{g}_{\max,k_s}^{u,R,\pi,s} = \max_{S \subseteq [u-4q,u+q] \setminus (A^{\pi,u} \cup \{s\})} \widetilde{g}_{\max,k_s}^{u,R,S,\pi,s}$	
$\overline{g}_{\max,k_s}^{R,\pi,s}$	$\overline{g}_{\max,k_s}^{R,\pi,s} = \max_{s-q \le u \le s} \widetilde{g}_{\max,k_s}^{u,R,\pi,s}$	
$\overline{g}_{\max,k_s}^{\pi,s}$	$\overline{g}_{\max,k_s}^{\pi,s} = \overline{g}_{\max,k_s}^{\emptyset,\pi,s}$	
$\overline{g}_{\mathsf{span},k_s}^{\pi,s}$	$\overline{g}_{\mathsf{span},k_s}^{\pi,s} \triangleq \overline{g}_{\max,k_s}^{\pi,s} - \overline{g}_{\min,k_s}^{\pi,s}$	
$(\mathcal{D}_t)^2$	$(\mathcal{D}_t)^2 \triangleq \mathbb{E}\left[\left(\overline{\Delta}_{\max} x_{k_t}^{\pi,t} - \overline{\Delta}_{\min} x_{k_t}^{\pi,t}\right)^2\right]$	Section 4.2.1
$\left(\Delta_t^X\right)^2$	$\left(\Delta_t^X\right)^2 \triangleq \mathbb{E}\left[\left(\Delta x_{k_t}^{\pi,t}\right)^2\right]$	
$\mathrm{prev}(t,k)$	The time of the most recent update to coordinate k, if any, in the time range $[t-2q, t-1]$; otherwise, we set it to t.	Lemma 7
ν_1, ν_2	$\nu_1 \triangleq \frac{20q^2}{n}$ and $\nu_2 \triangleq \frac{24q^2 L_{\text{res}}^2}{n\Gamma^2}$	Lemma 10

Notation / Parameter	Definition / Description	First Appearance
Λ, r ν_3, ν_4	$\Lambda \triangleq \frac{L_{res}^2}{\Gamma^2} + 1$ and $r \triangleq \frac{160q^2}{n} \cdot \Lambda^2$ $\nu_3 \triangleq \frac{3}{16}\left(\frac{r^2 r}{1-r} + r\right)$ and $\nu_4 \triangleq \frac{6r}{1-r}$	Lemma 11
$\mathcal{V}_m$		Eqn. (20)
$\widetilde{\Delta}^{t_{m-1}} x_{k_s}^{\pi,s}$	Before Observation 1	
ϖ	$\varpi \triangleq \frac{1}{q}\left[\frac{2r}{3} + \frac{3r^2}{1280} + \frac{9r^3}{25600} + \frac{3r^2}{1-r} + \frac{r^3}{426(1-r)} + \frac{r^4}{2844(1-r)}\right]$	Lemma 12
ϱ	$\varrho \triangleq \frac{1}{8} - \frac{15r}{1-r}$	Section 4.3
γ	a parameter introduced for amortization	Eqn. (16)
c	$c \triangleq \varpi + (\gamma + \varpi)\left(\frac{\nu_1}{q} + \frac{\nu_2}{q}\right)$	Lemma 13

For any set $R, S \subset [u - 4q, u + q] \setminus A^{\pi,u} \cup \{s\}$, when the first $(u - 4q - 1)$ updates on path π have been fixed, and update $\mathcal{U}_v$ is excluded from the computation of $\mathcal{U}_s$ for $v \in R \cup S$ and for $v > u + q$; for all $r \in A^{\pi,u}$, the value of the update $\mathcal{U}_r$ is already fixed, then:

$\Delta_{\max}^{u,R,S} x_{k_s}^{\pi,s}$	$\triangleq$	the maximum value that $\Delta x_{k_s}^{\pi,s}$ can assume
$\widetilde{g}_{\max,k_s}^{u,R,S,\pi,s}$	$\triangleq$	the maximum value of $\widetilde{g}_{k_s}^{\pi,s}$ can assume