Journal of Optimization Theory and Applications (2023) 199:499–540
https://doi.org/10.1007/s10957-023-02297-y

Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization

Ahmed Khaled[1] · **Othmane Sebbouh**[2] · **Nicolas Loizou**[3] · **Robert M. Gower**[4] · **Peter Richtárik**[5]

Received: 16 December 2021 / Accepted: 21 August 2023 / Published online: 27 September 2023

Abstract

We present a unified theorem for the convergence analysis of stochastic gradient algorithms for minimizing a smooth and convex loss plus a convex regularizer. We do this by extending the unified analysis of Gorbunov et al. (in: AISTATS, 2020) and dropping the requirement that the loss function be strongly convex. Instead, we rely only on convexity of the loss function. Our unified analysis applies to a host of existing algorithms such as proximal SGD, variance reduced methods, quantization and some coordinate descent-type methods. For the variance reduced methods, we recover the best known convergence rates as special cases. For proximal SGD, the quantization and coordinate-type methods, we uncover new state-of-the-art convergence rates. Our analysis also includes any form of sampling or minibatching. As such, we are able to determine the minibatch size that optimizes the total complexity of variance

Communicated by Alexander Vladimirovich Gasnikov.

✉ Ahmed Khaled
ahmed.khaled@princeton.edu

Othmane Sebbouh
othmane.sebbouh@gmail.com

Nicolas Loizou
nloizou@jhu.edu

Robert M. Gower
gowerrobert@gmail.com

Peter Richtárik
peter.richtarik@kaust.edu.sa

[1] Princeton University, Princeton, USA

[2] ENS Paris, CREST-ENSAE, Palaiseau, France

[3] Johns Hopkins University, Baltimore, USA

[4] Flatiron Institute, New York, USA

[5] KAUST, Thuwal, Saudi Arabia

reduced methods. We showcase this by obtaining a simple formula for the optimal minibatch size of two variance reduced methods (*L-SVRG* and *SAGA*). This optimal minibatch size not only improves the theoretical total complexity of the methods but also improves their convergence in practice, as we show in several experiments.

Keywords Stochastic optimization · Convex optimization · Variance reduction · Composite optimization

1 Introduction and Background

Consider the following composite convex optimization problem

$$\min_{x\in\mathbb{R}^d} \left\{F(x) \equiv f(x) + R(x)\right\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex and $R : \mathbb{R}^d \to (-\infty, \infty]$ is a proper closed and convex function with an easy-to-compute proximal term. This problem often arises in training machine learning models, where f is a loss function and R is a regularization term, e.g., ℓ_1-regularized logistic regression [33], LASSO regression [41] and elastic net regression [47]. It also includes projected gradient descent, if R is an indicator on a convex set.

A natural algorithm which is well-suited for solving (1) is proximal gradient descent, which requires iteratively taking a proximal step in the direction of the steepest descent. Unfortunately, this method requires computing the gradient ∇f at each iteration, which can be computationally expensive or even impossible in several settings. This has sparked interest in developing cheaper, practical methods that need only a stochastic unbiased estimate $g_k \in \mathbb{R}^d$ of the gradient at each iteration. These methods can be written as

$$x_0 \in \mathbb{R}^d, \quad x_{k+1} = \operatorname{prox}_{\gamma_k R}\left(x_k - \gamma_k g_k\right), \tag{2}$$

where $(\gamma_k)_k$ is a sequence of step sizes. This estimate g_k can take on many different forms depending on the problem of interest. Here we list a few.

1.1 Stochastic Approximation

Most machine learning problems can be cast as minimizing the generalization error of some underlying model, where $f_z(x)$ is the loss over a sample z and

$$f(x) = \mathbb{E}_{z\sim\mathcal{D}}\left[f_z(x)\right]. \tag{3}$$

Since $\mathcal{D}$ is an unknown distribution, computing this expectation is impossible in general. However, by sampling $z \sim \mathcal{D}$, we can compute a *stochastic gradient* $\nabla f_z(x)$. Using Algorithm (2) with $g_k = \nabla f_{z_k}(x_k)$ and $R \equiv 0$ gives the simplest stochastic gradient descent method: stochastic gradient descent (SGD) [29, 35].

1.2 Finite-Sum Minimization

Since the expectation (3) cannot be computed in general, one well-studied solution to approximately solve this problem is to use a Monte Carlo estimator:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{4}$$

where n is the number of samples and $f_i(x)$ is the loss at x on the ith drawn sample. When R is a regularization function, problem (1) with f defined in (4) is often referred to as regularized empirical minimization (R-ERM) [39]. For the approximation (4) to be accurate, we would like n to be as large as possible. This, in turn, makes computing the gradient extremely costly. In this setting, for low-precision problems, SGD scales very favorably compared to gradient descent, since an iteration of SGD requires $\mathcal{O}(d)$ flops compared to $\mathcal{O}(nd)$ for gradient descent. Moreover, several techniques applied to SGD such as importance sampling and minibatching [12, 21, 28, 46] have made SGD the preferred choice for solving Problem (1) + (4). However, one major drawback of SGD is that, using a fixed step size, SGD does not converge and oscillates in the neighborhood of a minimizer. To remedy this problem, *variance reduced methods* [3, 8, 19, 32, 36] were developed. These algorithms get the best of both worlds: the global convergence properties of GD and the small iteration complexity of SGD. In the smooth case, they all share the distinguishing property that the variance of their stochastic gradients g_k converges to 0. This feature allows them to converge to a minimizer with a fixed step size at the cost of some extra storage or computations compared to SGD.

1.3 Distributed Optimization

Another setting where the exact gradient ∇f is impossible to compute is in distributed optimization. The objective function in distributed optimization can be formulated exactly as (4), where each f_i is a loss on the data stored on the ith node. Each node computes the loss on its local data, then the losses are aggregated by the master node. When the number of nodes n is high, the bottleneck of the optimization becomes the cost of communicating the individual gradients. To remedy this issue, various compression techniques were proposed [1, 2, 15, 22, 38, 43, 45], most of which can be modeled as applying a random transformation $Q : \mathbb{R}^d \mapsto \mathbb{R}^d$ to each gradient $\nabla f_i(x_k)$ or to a noisy estimate of the gradient g_i^k. Thus, many proximal quantized stochastic gradient methods fit the form (2) with

$$g_k = \sum_{i=1}^{n} Q(g_i^k).$$

While quantized stochastic gradient methods have been widely used in machine learning applications, it was not until the *DIANA* algorithm [26, 27] that a distributed method was shown to converge to the neighborhood of a minimizer for strongly convex functions. Moreover, in the case where each f_i is itself a finite average of local functions, variance reduced versions of *DIANA*, called *VR-DIANA* [18], were recently developed and proved to converge sublinearly with a fixed step size for convex functions.

1.4 High-Dimensional Function Minimization

Lastly, regardless of the structure of f, if the dimension of the problem d is very high, it is sometimes impossible to compute or to store the gradient at any iteration. Instead, in some cases, one can efficiently compute some coordinates of the gradient, and perform a gradient descent step on the selected coordinates only. These methods are known as (randomized) coordinate descent (RCD) methods [30, 44]. These methods also fit the form (2), for example, with

$$g_k = \nabla f(x_k) e_{i_k},$$

where $(e_i)_i$ is the canonical basis of $\mathbb{R}^d$ and $i_k \in [d]$ is sampled randomly at each iteration. Though RCD methods fit the form (2) their analysis is often very different compared to other stochastic gradient methods. One exception to this observation is *SEGA* [16], the first RCD method known to converge for strongly convex functions with nonseparable regularizers.

While all the methods presented above have been discovered and analyzed independently, most of them rely on the same assumptions and share a similar analysis. It is this observation and the results derived for strongly convex functions by [11] that motivate this work.

2 Contributions

We now summarize the key contributions of this paper.

2.1 Unified Analysis of Stochastic Gradient Algorithms

Under a unified assumption on the gradients g_k, it was shown by [11] that stochastic gradient methods which fit the format (2) converge linearly to a neighborhood of the minimizer for quasi-strongly convex functions when using a fixed step size. We extend this line of work to the convex setting, and further generalize it by allowing for decreasing step sizes. As a result, for all the methods which verify our assumptions, we are able to prove either sublinear convergence to the neighborhood of a minimum with a fixed step size or exact convergence with a decreasing step size.

2.2 Analysis of SGD Without the Bounded Gradients Assumption

Most of the existing analysis on SGD assume a uniform bound on the second moments of the stochastic gradients or on their variance. Indeed, for the analysis of stochastic (sub)gradient descent, this is often necessary to apply the classical convergence proofs. However, for large classes of convex functions, it has been shown that these assumptions do not to hold [20, 31]. As a result, there has been a recent surge in trying to avoid these assumptions on the stochastic gradients for several classes of smooth functions: strongly convex [14, 25, 31], convex [14, 25, 40, 42], or even nonconvex functions [20, 24, 25]. Surprisingly, a general analysis for SGD with proximal iterations for convex functions without these bounded gradient assumptions is still lacking. As a special case of our unified analysis, assuming only convexity and smoothness, we provide a general analysis of proximal SGD in the convex setting. Moreover, using the *arbitrary sampling* framework of [12], we are able to prove convergence rates for SGD under minibatching, importance sampling, or virtually any form of sampling.

2.3 Extension of the Analysis of Existing Algorithms to the Convex Case

As another special case of our analysis, we also provide the first convergence rates for the (variance reduced) stochastic coordinate descent method *SEGA* [16] and the distributed (variance reduced) compressed SGD method *DIANA* [26] in the convex setting. Our results can also be applied to all the recent methods developed in [11].

2.4 Optimal Minibatches for L-SVRG and SAGA in the Convex Setting

With a unifying convergence theory in hand, we can now ask sweeping questions across families of algorithms. We demonstrate this by answering the question

What is the optimal minibatch size for variance reduced methods?

Recently, precise estimates of the minibach sizes which minimize the total complexity for *SAGA* [8] and *SVRG* [4, 19, 34] applied to strongly convex functions were derived by [9] and [37]. We showcase the flexibility of our unifying framework by deriving new optimal minibatch sizes for *SAGA* [8] and *L-SVRG* [17, 23] in the general convex setting. Unlike prior work in the strongly convex setting [9, 37], our resulting optimal minibatch sizes can be computed using only the smoothness constants. To verify the validity of our claims, we show through extensive experiments that our theoretically derived optimal minibatch sizes are competitive against a grid search.

3 Unified Analysis for Proximal Stochastic Gradient Methods

3.1 Notation

The Bregman divergence associated with f is the mapping

$$D_f(x,y) \overset{\text{def}}{=} f(x) - f(y) - \langle \nabla f(y), x - y \rangle\,, \quad x, y \in \mathbb{R}^d,$$

and the proximal operator of γR is the function

$$\operatorname{prox}_{\gamma R}(x) \overset{\text{def}}{=} \operatorname{argmin}_{u\in\mathbb{R}^d}\left\{\gamma R(x) + \frac{1}{2}\|x-u\|^2\right\}.$$

Let $[n] \overset{\text{def}}{=} \{1,\dots,n\}$. We denote the expectation of a random variable X by $\mathbb{E}[X]$ and the conditional expectation of X given a random variable Y by $\mathbb{E}[X \mid Y]$.

[11] analyze stochastic gradient methods that fit the form (2) for smooth quasi-strongly convex functions. In this work, we extend these results to the general convex setting. We formalize our assumptions on f and R in the following.

Assumption 1 The function f is L–smooth and convex:

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2, \quad \text{for all } x, y \in \mathbb{R}^d, \tag{5}$$

$$f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle, \quad \text{for all } x, y \in \mathbb{R}^d. \tag{6}$$

The function R is convex:

$$R(\alpha x + (1-\alpha)y) \ge \alpha R(x) + (1-\alpha)R(y), \quad \text{for all } x, y \in \mathbb{R}^d, \alpha \in [0,\ 1].$$

When f has the form (4), we assume that for all $i \in [n]$, f_i is L_i-smooth and convex, and we denote $L_{\max} \overset{\text{def}}{=} \max_{i\in[n]} L_i$.

The innovation introduced by [11] is the following unifying assumption on the stochastic gradients g_k used in (2) which allows to simultaneously analyze classical SGD, variance reduced methods, quantized stochastic gradient methods, and some randomized coordinate descent methods.

Assumption 2 (*Assumption 4.1 in* [11]) Consider the iterates $(x_k)_k$ and gradients $(g_k)_k$ in (2).

1. The gradient estimates are conditionally unbiased:

$$\mathbb{E}[g_k \mid x_k] = \nabla f(x_k). \tag{7}$$

2. There exist constants $A, B, C, D_1, D_2, \rho \ge 0$, and a sequence of random variables $\sigma_k^2 \ge 0$ such that for all possible minimizers x_* of F:

$$\mathbb{E}\left[\|g_k - \nabla f(x_*)\|^2 \mid x_k\right] \le 2AD_f(x_k, x_*) + B\sigma_k^2 + D_1, \tag{8}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid x_k\right] \le (1-\rho)\,\sigma_k^2 + 2CD_f(x_k, x_*) + D_2. \tag{9}$$

Though we chose to present Eqs. (7), (8) and (9) as an assumption, we show throughout the main paper and in the appendix that for all the algorithms we consider (excluding *DIANA*), these equations all hold with known constants when Assumption 1 holds. An extensive yet nonexhaustive list of algorithms satisfying Assumption 2 and the corresponding constants can be found in [11, Table 2]. We report in Section B of the appendix these constants for five algorithms: *SGD*, two variance reduced methods *L-SVRG* and *SAGA*, a distributed method *DIANA* and a coordinate descent-type method *SEGA*.

We now state our main theorem.

Theorem 3.1 *Suppose that Assumptions 1 and 2 hold. Let* $M \overset{def}{=} B/\rho$ *and let* $(\gamma_k)_{k\geq 0}$ *be a decreasing, strictly positive sequence of step sizes chosen such that*

$$0 < \gamma_0 < \min\left\{\frac{1}{2(A+MC)}, \frac{1}{L}\right\}.$$

The iterates given by (2) satisfy

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma_0\left(\delta_0 + \gamma_0 M\sigma_0^2\right) + 2\left(D_1 + 2MD_2\right)\sum\limits_{k=0}^{t-1}\gamma_k^2}{2\sum_{i=0}^{t-1}\left(1 - 2\gamma_i\left(A + MC\right)\right)\gamma_i}, \tag{10}$$

where $\bar{x}_t \overset{def}{=} \sum_{k=0}^{t-1}\frac{(1-2\gamma_k(A+MC))\gamma_k}{\sum_{i=0}^{t-1}(1-2\gamma_i(A+MC))\gamma_i}x_k$ *and* $\delta_0 \overset{def}{=} F(x_0) - F(x_*)$.

The proof of Theorem 3.1 is deferred to the appendix (Section C).

4 The Main Corollaries

In contrast to [11], our analysis allows both constant and decreasing step sizes. In this section, we will present two corollaries corresponding to these two choices of step sizes and discuss the resulting convergence rates depending on the constants obtained from Assumption 2. Then, we specialize our theorem to SGD, which allows us to recover the first analysis of proximal SGD without the bounded gradients or bounded gradient variance assumptions in the general convex setting. We apply the same analysis to *DIANA* and present convergence results for this algorithm in the convex setting.

First, we show that by using a constant step size the average of iterates of any stochastic gradient method of the form (2) satisfying Assumptions 1 and 2 converges sublinearly to the neighborhood of the minimum.

Corollary 4.1 *Consider the setting of Theorem 3.1. Let* $M = B/\rho$. *Choose stepsizes* $\gamma_k = \gamma > 0$ *for all* k, *where* $\gamma \leq \min\left\{\frac{1}{4(A+MC)}, \frac{1}{2L}\right\}$; *then substituting in the rate in (10), we have,*

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{2\gamma\left(\delta_0 + \gamma M\sigma_0^2\right) + \|x_0 - x_*\|^2}{\gamma t} + 2\gamma\left(D_1 + MD_2\right).$$

One can already see that to ensure convergence with a fixed step size, we need to have $D_1 = D_2 = 0$. The only known stochastic gradient methods which satisfy this property are variance reduced methods, as we show in Sect. 5. When $D_1 \neq 0$ or $D_2 \neq 0$, which is the case for *SGD* and *DIANA* (See Section B), the solution to ensure anytime convergence is to use decreasing step sizes.

Corollary 4.2 *Consider the setting of Theorem 3.1. Let $M = B/\rho$. Choose stepsizes $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$ for all $k \geq 0$, where $\gamma \leq \min\left\{\frac{1}{4(A+MC)}, \frac{1}{2L}\right\}$. Then substituting in the rate in (10), we have*

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\gamma\left(\delta_0 + \gamma M \sigma_0^2\right) + \|x_0 - x_*\|^2 + \left(\frac{D_1}{2} + M D_2\right)(\log(t) + 1)}{\gamma\left(\sqrt{t} - 1\right)}$$
$$\sim \mathcal{O}\left(\frac{\log(t)}{\sqrt{t}}\right).$$

4.1 SGD Without the Bounded Gradients Assumption

To better illustrate the significance of the convergence rates derived in Corollaries 4.1 and 4.2, consider the SGD method for the finite sum setting (4):

$$x_0 \in \mathbb{R}^d, \quad x_{k+1} = \operatorname{prox}_{\gamma_k R}\left(x_k - \gamma_k \nabla f_{i_k}(x_k)\right), \tag{11}$$

where i_k is sampled uniformly at random from $[n]$. Note that we consider the finite sum setting just for illustration, and our results continue to hold under the more general Monte Carlo setting.

Lemma 4.1 *Assume that f has a finite sum structure (4) and that Assumption 1 holds. The iterates defined by (11) verify Assumption 1 with*

$$A = 2L_{\max}, \;\; B = 0, \;\; \rho = 1, \;\; C = 0, \;\; D_1 = 2\sigma^2, D_2 = 0, \tag{12}$$

where $\sigma^2 = \frac{1}{n} \sup\limits_{x_* \in X^*} \sum_{i=1}^n \|\nabla f_i(x_*)\|^2$, where X_* is the set of minimizers of F, and $L_{\max} = \max\limits_{i \in [n]} L_i$.

Proof See Lemma A.1 in [11]. □

This analysis can be easily extended to include minibatching, importance sampling, and virtually all forms of sampling by using the constants given in (12), with the exception of $L_{\max}$ which should be replaced by the *expected smoothness* constant [12]. Due to lack of space, we defer this general analysis of SGD to the appendix (Sections A and B). Using Theorem 3.1 and Lemma 4.1, we arrive at the following result.

Corollary 4.3 *Let $(\gamma_k)_k$ be a sequence of decreasing step sizes such that $0 < \gamma_0 \leq 1/4L_{\max}$ for all $k \in \mathbb{N}$. Let Assumption 1 hold. The iterates of (11) verify*

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma_0\left(F(x_0) - F(x_*)\right)}{\sum_{i=0}^{t-1} \gamma_i} + \frac{2\sigma^2 \sum_{k=0}^{t-1} \gamma_k^2}{\sum_{i=0}^{t-1} \gamma_i}.$$

Moreover, as we did in Corollaries 4.1 and 4.2, we can show sublinear convergence to a neighborhood of the minimum if we use a fixed step size, or $\mathcal{O}(\log(k)/\sqrt{k})$ convergence to the minimum using a step size $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$. Moreover, if we know the stopping time of the algorithm, we can derive a $\mathcal{O}(1/\sqrt{k})$ upper bound as done in [29].

Corollary 4.3 fills a gap in the theory of SGD. Indeed, to the best of our knowledge, this is the first analysis of proximal SGD in the convex setting which does not assume neither bounded gradients nor bounded variance (as done in, e.g., [10, 29]). Instead, it relies only on convexity and smoothness. The closest results to ours here are Theorem 1.6 in [14] and Theorem 5 in [40], both of which are in the same setting as Lemma 4.1 but study more restrictive variants of proximal SGD. [14] studies SGD with projection onto closed convex sets and [40] studies vanilla SGD, without proximal or projection operators. Unfortunately, neither result extends easily to include using proximal operators, and hence our results necessitate a different approach. When specialized to the setting where $g = 0$ (i.e., no prox) and L-smoothness ($A = L$, $B = 0$), Corollary 4.1 gives us the following guarantee after T steps of SGD:

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leq \frac{2\gamma\delta_0 + \|x_0 - x_*\|^2}{\gamma t} + 2\gamma D_1.$$

Observe that by the smoothness of f we have $\delta_0 \leq \frac{L}{2}\|x_0 - x_*\|^2$, therefore

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leq \frac{(1+\gamma L)\|x_0 - x_*\|^2}{\gamma t} + 2\gamma D_1.$$

Optimizing over γ, we get that for $\gamma = \min\left\{\frac{1}{2L}, \frac{\|x_0 - x_*\|}{\sqrt{2D_1 T}}\right\}$ we have the convergence rate

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leq \frac{3L\|x_0 - x_*\|^2}{T} + \frac{3\sqrt{2}R\sqrt{D_1}}{\sqrt{T}}.$$

This matches the guarantees given in [14, 40] up to constants. Therefore, our analysis is more general while sacrificing no degradation in the convergence rate.

4.2 Convergence of *DIANA* in the Convex Setting

DIANA was the first distributed quantized stochastic gradient method proven to converge to the minimizer in the strongly convex case and to a critical point in the nonconvex case [26]. See Section B.2 in the appendix for the definition of *DIANA* and its parameters.

Lemma 4.2 *Assume that f has a finite sum structure and that Assumption 1 holds. The iterates of DIANA (Algorithm 4) satisfy Assumption 2 with constants:*

$$A = \left(1 + \frac{2w}{n}\right) L_{\max}, \ B = \frac{2w}{n}, \ \rho = \alpha,$$
$$C = L_{\max}\alpha, \ D_1 = \frac{(1+w)\sigma^2}{n}, \ D_2 = \alpha\sigma^2,$$

where $w > 0$ and $\alpha \leq \frac{1}{1+w}$ are parameters of Algorithm 4 and σ^2 is such that

$$\forall k \in \mathbb{N}, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\| g_i^k - \nabla f(x_k)\right\|^2\right] \leq \sigma^2.$$

Proof See Lemma A.12 in [11]. □

As yet another corollary of Theorem 3.1, we can extend the results of [26] to the convex case and show that *DIANA* converges sublinearly to the neighborhood of the minimum using a fixed step size, or to the minimum exactly using a decreasing step size.

Corollary 4.4 *Assume that f has a finite sum structure (4) and that Assumption 1 holds. Let $(\gamma_k)_{k\geq 0}$ be a decreasing, strictly positive sequence of step sizes chosen such that*

$$0 < \gamma_0 < \frac{1}{4(1+\frac{4w}{n})L_{\max}}.$$

By Theorem 3.1 and Lemma 4.2, we have that the iterates given by Algorithm 4 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma_0\left(F(x_0) - F(x_*) + \frac{2w\gamma_0}{\alpha n}\sigma_0^2\right) + \frac{2(1+5w)\sigma^2}{n}\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{i=0}^{t-1}\gamma_i}.$$

5 Optimal Minibatch Sizes for Variance Reduced Methods

Variance reduced methods are of particular interest because they do not require a decreasing step size in order to ensure convergence. This is because for variance reduced methods we have $D_1 = D_2 = 0$, and thus, these methods converge sublinearly with a fixed step size.

Variance reduced methods were designed for solving (1) in the special case where f has a finite sum structure. In this case, in order to further improve the convergence properties of variance reduced methods, several techniques can be applied such as adding momentum [3] or using importance sampling [13], but the most popular of such techniques is by far minibatching. Minibatching has been used in conjunction with variance reduced methods since their inception [21], but it was not until [9, 37] that a theoretical justification for the effectiveness of minibatching was proved for SAGA [8] and SVRG [19] in the strongly convex setting. In this section, we show how our theory allows us to determine the optimal minibatch sizes which minimize the total complexity of any variance reduced method. This allows us to compute the first estimates of these minibatch sizes in the nonstrongly convex setting. For simplicity, in the remainder of this section, we will consider the special case where $R = 0$. Hence, in this section

$$F(x) = f(x) \equiv \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

To derive a meaningful optimal minibatch size from our theory, we need to use the tightest possible upper bounds on the total complexity. When $R = 0$, we can derive a slightly tighter upper bound than the one we obtained in Theorem 3.1 as follows.

Proposition 5.1 *Let $R = 0$ and $M = B/2\rho$. Suppose that Assumption 2 holds with $D_1 = D_2 = 0$. Let the step sizes $\gamma_k = \gamma$ for all $k \in \mathbb{N}$, with $\gamma_k = \gamma \leq 1/(4(A+MC))$ for all $k \in \mathbb{N}$. Then,*

$$\mathbb{E}\left[f(\bar{x}_k) - f(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2M\gamma^2\sigma_0^2}{\gamma k}. \tag{13}$$

We can translate this upper bound into a convenient complexity result as follows.

Corollary 5.1 *Assume that there exists a constant $G \geq 0$ such that*

$$\sigma_0^2 \leq G\,\|x_0 - x_*\|^2. \tag{14}$$

Let $\epsilon > 0$ and $\gamma = \frac{1}{4(A+\frac{BC}{2\rho})}$. It follows that

$$k \geq \left(4(A + \frac{BC}{2\rho}) + \frac{BG}{2(2\rho A + BC)}\right)\frac{\|x_0 - x_*\|^2}{\epsilon} \tag{15}$$

$$\implies \mathbb{E}\left[f(\bar{x}_k) - f(x_*)\right] \leq \epsilon. \tag{16}$$

Proof The result follows from taking $\gamma = \frac{1}{4(A+\frac{BC}{2\rho})}$ and upper bounding σ_0^2 by $G\|x_0 - x_*\|^2$ in (13). □

In the same way we specialized the general convergence rate given in Theorem 3.1 to the cases of *SGD* and *DIANA* in Sect. 4, we can specialize the iteration complexity result (15) to any method which verifies $D_1 = D_2 = 0$. Due to their popularity, we chose to analyze minibatch variants of *SAGA* [8] and *L-SVRG* [17, 23], a single-loop variant of the original SVRG algorithm [19]. The pseudocode for these algorithms is presented in Algorithms 1 and 2. We define for any subset $B \subseteq [n]$ the minibatch average of f over B as $f_B(x) = \frac{1}{b}\sum_{i\in B} f_i(x)$.

Algorithm 1 b-SAGA

```
**Parameters** minibatch size $b$, step size $\gamma$
**Initialization** $x_0 \in \mathbb{R}^d$ and $J_0^i = \nabla f_i(x_0)$ for $i = 1, \ldots, n$.
for $k = 0, 1, \ldots$ do
    Sample a batch $B \subseteq [n]$ with $|B| = b$
    $g_k = \frac{1}{n}\sum_{i=1}^n J_k^i + \nabla f_B(x_k) - \frac{1}{b}\sum_{i \in B} J_k^i$
    $x_{k+1} = x_k - \gamma g_k$
    $J_{k+1}^i = \begin{cases} J_k^i & \text{if } i \notin B \\ \nabla f_i(x_k) & \text{if } i \in B \end{cases}$
end for
```

Algorithm 2 b-L-SVRG

```
**Parameters** minibatch size $b$, step size $\gamma$, $p \in (0, 1]$
**Initialization** $w_0 = x_0 \in \mathbb{R}^d$
for $k = 0, 1, \ldots$ do
    Sample a batch $B \subseteq [n]$ with $|B| = b$
    $g_k = \nabla f_B(x_k) - \nabla f_B(w_k) + \nabla f(w_k)$
    $x_{k+1} = x_k - \gamma g_k$
    $w_{k+1} = \begin{cases} x_k & \text{w. prob. } p \\ w_k & \text{w. prob. } 1 - p \end{cases}$
end for
```

As we will show next, the iterates of Algorithms 1 and 2 satisfy Assumption 2 with constants which depend on the minibatch size b. These constants will depend on the following *expected smoothness* and *expected residual* constants $\mathcal{L}(b)$ and $\zeta(b)$ used in the analysis of *SAGA* and *SVRG* in [9, 37]:

$$\mathcal{L}(b) \overset{\text{def}}{=} \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \frac{n}{b}\frac{b-1}{n-1}L, \quad \text{and} \quad \zeta(b) \overset{\text{def}}{=} \frac{1}{b}\frac{n-b}{n-1}L_{\max}. \tag{17}$$

5.1 Optimal Minibatch Size for *b*-SAGA

Consider the b-SAGA method in Algorithm 1. Define

$$H(x) \overset{\text{def}}{=} [f_1(x), \ldots, f_n(x)] \in \mathbb{R}^d$$

and let $\nabla H(x) \in \mathbb{R}^{d \times n}$ denote the Jacobian of H. Let $J_k = [J_k^1, \ldots, J_k^n]$ be the current *stochastic Jacobian*.

Lemma 5.1 *The iterates of Algorithm 1 satisfy Assumption 2 and Eq. (14) with*

$$\sigma_k^2 = \frac{1}{nb}\frac{n-b}{n-1}\left\| J_k - \nabla H(x_*)\right\|_{\text{Tr}}^2, \tag{18}$$

where for all $Z \in \mathbb{R}^{d \times n}$, $\|Z\|_{\text{Tr}}^2 = \text{tr}\left(ZZ^\top\right)$, and constants

$$A = 2\mathcal{L}(b),\ B = 2,\ \rho = \frac{b}{n},\ C = \frac{b\zeta(b)}{n},\ D_1 = D_2 = 0,\ G = \zeta(b)L. \tag{19}$$

Using Corollary 5.1, we can determine the iteration complexity of Algorithm 1.

Corollary 5.2 (Iteration complexity of $b-$SAGA) *Consider the iterates of Algorithm 1. Let the stepsize used be $\gamma = \frac{1}{4(2\mathcal{L}(b)+\zeta(b))}$. Given the constants obtained for Algorithm 1 in (19), by Corollary 5.1 we have that*

$$k \geq \left(4(2\mathcal{L}(b)+\zeta(b)) + \frac{n\zeta(b)L}{2b\left(2\mathcal{L}(b)+\zeta(b)\right)}\right)\frac{\|x_0 - x_*\|^2}{\epsilon}$$
$$\implies \mathbb{E}\left[F(\bar{x}_k) - F(x_*)\right] \leq \epsilon.$$

We define the total complexity as the number of gradients computed per iteration (b) times the iteration complexity required to reach an ϵ-approximate solution. Thus, multiplying by b the iteration complexity in Corollary 5.2 and plugging in (17), the total complexity for Algorithm 1 is upper bounded by

$$K_{saga}(b) \overset{\text{def}}{=} \left(\frac{4\left(3(n-b)L_{\max} + 2n(b-1)L\right)}{n-1} + \frac{n(n-b)L_{\max}L}{2\left(3(n-b)L_{\max} + 2n(b-1)L\right)}\right)\frac{\|x_0 - x_*\|^2}{\epsilon}. \tag{20}$$

Minimizing this upper bound in the minibatch size b gives us an estimate of the optimal empirical minibatch size, which we verify in our experiments.

Proposition 5.2 *Let $b^*_{saga} = \underset{b\in[n]}{\operatorname{argmin}}\, K_{saga}(b)$, where $K_{saga}(b)$ is defined in (20).*

- *If $L_{\max} \leq \frac{2nL}{3}$ then*

$$b^*_{saga} = \begin{cases} 1 & \text{if} \quad \bar{b} < 2 \\ \lfloor b_1 \rfloor & \text{if} \quad 2 \leq \bar{b} < n \\ n & \text{if} \quad \bar{b} \geq n, \end{cases} \tag{21}$$

where

$$b_1 \overset{def}{=} \frac{n\left((n-1)L\sqrt{L_{\max}} - 2\sqrt{2nL - 3L_{\max}}(3L_{\max} - 2\,L)\right)}{2(2nL - 3L_{\max})^{\frac{3}{2}}}.$$

- *Otherwise, if $L_{\max} > \frac{2nL}{3}$ then $b^* = n$.*

5.2 Optimal Minibatch Size for *b*-L-SVRG

Since the analysis for Algorithm 2 is similar to that of Algorithm 1, we defer its details to the appendix and only present the total complexity and the optimal minibatch size. Indeed, as shown in Section E.3, an upper bound on the total complexity to find an

ϵ-approximate solution for Algorithm 2 is given by

$$K_{svrg}(b) \overset{\text{def}}{=} (1+2b)\left(\frac{12\left((n-b)L_{\max}+n(b-1)L\right)}{b(n-1)} + \frac{nL}{6}\right)\frac{\|x_0 - x_*\|^2}{\epsilon}.$$

Proposition 5.3 *Let* $b^*_{svrg} = \underset{b\in[n]}{\operatorname{argmin}}\, K_{svrg}(b)$*, where* $K_{svrg}(b)$ *is defined in* (44)*. Then,*

$$b^*_{svrg} = 6\sqrt{\frac{n\left(L_{\max}-L\right)}{72\left(nL - L_{\max}\right) + n(n-1)L}}.$$

6 Experiments

Here we test our new formula for optimal minibatch size of SAGA given by (21) against the best minibatch size found over a grid search. We used logistic regression with no regularization ($\lambda = 0$) to emphasize that our results hold for nonstrongly convex functions with data sets taken from the LIBSVM collection [7]. For each data set, we ran minibatch SAGA with the stepsize given in Corollary 5.2 and until a solution with

$$F(x_t) - F(x^*) < 10^{-4}(F(x_0) - F(x^*))$$

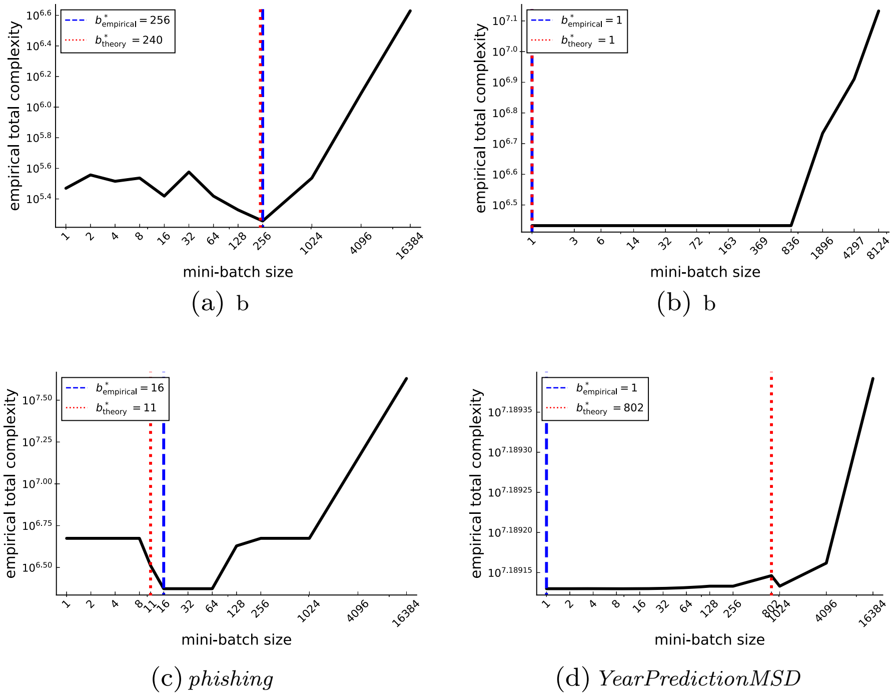
was reached.

In Fig. 1 we plot the total complexity (number of iterations times the minibatch size) to reach this tolerance for each minibatch size on the grid. We can see in Fig. 1 that for *ijcnn* and *phishing* the optimal minibatch size $b^*_{\text{theory}} = b^*_{saga}$ (21) is remarkably close to the best minibatch size over the grid $b^*_{\text{empirical}}$. Even when b^*_{theory} is not close to $b^*_{\text{empirical}}$, such as on the *YearPredictionMSD* problem, the resulting total complexity is still very close to the total complexity of $b^*_{\text{empirical}}$.

Outline of the Appendix

The appendix is organized as follows:

- **Section A**: we present the arbitrary sampling framework for stochastic gradient methods introduced by [13], which will be used for the analysis of *SGD* and *L-SVRG*.
- **Section B**: we present specializations of Theorem 3.1 to the algorithms we discuss: *SGD*, *DIANA*, *L-SVRG*, *SAGA* and *SEGA*.
- **Section C**: we present the proof of our main Theorem 3.1.
- **Section D**: we present the proof of Corollary 4.2.
- **Section E**, we present the proof of Proposition 5.1, and the detailed analysis of the optimal minibatch results for *b-SAGA* and *b-L-SVRG*, in addition to an analysis for the optimal miniblock size for *b-SEGA*.
- **Section F**: we present some technical lemmas which we use in our analysis.



(a) b

(b) b

(c) *phishing*

(d) *YearPredictionMSD*

Fig. 1 Comparing the theoretical optimal batchsize (21) with the best over a grid

Appendix A: Arbitrary Sampling

In this section, we recall the arbitrary sampling framework [12] which allows us to analyze our algorithms for minibatching, importance sampling and virtually all possible forms of sampling.

Appendix A.1: Stochastic reformulation

To see importance sampling and minibatch variants of stochastic gradient methods all through the same lens, we introduce a *sampling vector* which we will use to rewrite (1).

Definition A.1 We say that a random element-wise positive vector $v \in \mathbb{R}^n_+$ drawn from some distribution $\mathcal{D}$ is a sampling vector if its expectation is the vector of all ones:

$$\mathbb{E}_{\mathcal{D}}\left[v_i\right] = 1, \ \text{ for all } i \in [n]. \tag{22}$$

For a given distribution $\mathcal{D}$, we introduce a *stochastic reformulation* of (1) as follows

$$\min_{x\in\mathbb{R}^d} \left\{ \mathbb{E}_{\mathcal{D}}\left[f_v(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n v_i f_i(x) \right] + R(x) \right\}. \tag{23}$$

By definition of the sampling vector, $f_v(x)$ and $\nabla f_v(x)$ are unbiased estimators of $f(x)$ and $\nabla f(x)$, respectively, and hence problem (23) is indeed equivalent (i.e., a reformulation) of the original problem (1). In the case of the gradient, for instance, we get

$$\mathbb{E}_{\mathcal{D}}\left[\nabla f_v(x)\right] \overset{(23)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}\left[v_i\right]\nabla f_i(x) \overset{(22)}{=} \nabla f(x).$$

Reformulation (23) can be solved using proximal stochastic gradient descent via

$$x_{k+1} = \operatorname{prox}_{\gamma_k R}\left(x_k - \gamma \nabla f_{v_k}(x_k)\right), \tag{24}$$

where $v_k \sim \mathcal{D}$ is sampled i.i.d. at each iteration and $\gamma > 0$ is a stepsize. By substituting specific choices of $\mathcal{D}$, we obtain specific variants of SGD for solving (1). We further show that (24) is a special case of (2) with a sequence of vectors $g_k = \nabla f_{v_k}(x_k)$ and use the unified analysis in Theorem 3.1 to obtain convergence rates for (24).

Appendix A.2: Expected Smoothness and Gradient Noise

In order to analyze (24), we will make use of the following result, which characterizes the smoothness of the subsampled functions f_v.

Lemma A.1 (Expected Smoothness) *If for all $i \in [n]$, f_i is convex and L_i-smooth, then there exists a constant $\mathcal{L} \geq 0$ such that*

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x_*)\|^2\right] \leq 2\mathcal{L}\, D_f(x, x_*), \tag{25}$$

for all $x \in \mathbb{R}^d$ and where x_* is any minimizer of (1).

The proof of this result follows closely that of Lemma 1 in [9].

Proof Since for all $i \in [n]$, f_i is L_i-smooth and convex, we have that each realization f_v (defined in (23)) is L_v-smooth and convex. Thus, from Lemma F.1, we have that for all $x \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla f_v(x) - f_v(x_*)\|^2 &\leq 2L_v\left(f_v(x) - f_v(x_*) - \langle \nabla f_v(x_*), x - x_*\rangle\right)\\ &= \frac{2}{n}\sum_{i=1}^{n} L_v v_i\left(f_i(x) - f_i(x_*) - \langle \nabla f_i(x_*), x - x_*\rangle\right).\end{aligned}$$

Taking expectation over the samplings,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - f_v(x_*)\|^2\right] &\leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}\left[v_i L_v\right]\left(f_i(x) - f_i(x_*) - \langle \nabla f_i(x_*), x - x_*\rangle\right)\\ &\leq 2\max_{j=1,\ldots,n}\mathbb{E}_{\mathcal{D}}\left[L_v v_j\right]\left(f(x) - f(x_*) - \langle \nabla f(x_*), x - x_*\rangle\right)\\ &= 2\max_{j=1,\ldots,n}\mathbb{E}_{\mathcal{D}}\left[L_v v_j\right] D_f(x, x_*).\end{aligned}$$

□

Next, we define the gradient noise.

Definition A.2 (*Gradient Noise*) The gradient noise $\sigma^2 = \sigma^2(f, \mathcal{D})$ is defined by

$$\sigma^2 \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x_*) - \nabla f(x_*)\|^2\right]. \tag{26}$$

Appendix A.3: Minibatching Elements Without Replacement

Since analyzing minibatching for variance reduced methods is one of the main focuses of our work, we present minibatching without replacement as an example of the use of arbitrary sampling.

First, we define samplings.

Definition A.3 (*Sampling*) A sampling $S \subseteq [n]$ is any random set-valued map which is uniquely defined by the probabilities $\sum_{B \subseteq [n]} p_B = 1$ where $p_B \overset{\text{def}}{=} \mathbb{P}(S = B), \quad \forall B \subseteq [n]$. A sampling S is called proper if for every $i \in [n]$, we have that $p_i \overset{\text{def}}{=} \mathbb{P}(i \in S) = \sum\limits_{C : i \in C} p_C > 0$.

We can build a sampling vector using a sampling as follows.

Lemma A.2 (Sampling vector, Lemma 3.3 in [12]) *Let S be a proper sampling. Let $p_i \overset{def}{=} \mathbb{P}(i \in S)$ and $\mathbf{P} \overset{def}{=} \operatorname{diag}(p_1, \ldots, p_n)$. Let $v = v(S)$ be a random vector defined by*

$$v(S) \;=\; \mathbf{P}^{-1} \sum_{i \in S} e_i \;\overset{def}{=}\; \mathbf{P}^{-1} e_S. \tag{27}$$

It follows that v is a sampling vector.

Proof The ith coordinate of $v(S)$ is $v_i(S) = \mathbb{1}(i \in S)/p_i$ and thus

$$\mathbb{E}\left[v_i(S)\right] \;=\; \frac{\mathbb{E}\left[\mathbb{1}(i \in S)\right]}{p_i} \;=\; \frac{\mathbb{P}(i \in S)}{p_i} \;=\; 1.$$

□

Next, we define b-nice sampling, also known as minibatching without replacement.

Definition A.4 (*b-nice sampling*) S is a b-nice sampling if it is a sampling such that

$$\mathbb{P}(S = B) = \frac{1}{\binom{n}{b}}, \quad \forall B \subseteq [n], \quad \text{with } |B| = b.$$

To construct such a sampling vector based on the b–nice sampling, note that $p_i = \frac{b}{n}$ for all $i \in [n]$ and thus we have that $v(S) = \frac{n}{b} \sum_{i \in S} e_i$ according to Lemma A.2. The resulting subsampled function is then $f_v(x) = \frac{1}{|S|}\sum_{i\in S} f_i(x)$, which is simply the minibatch average over S.

A remarkable result for b-nice sampling is that when all the functions f_i, $i \in [n]$ are L_i-smooth and convex, then the expected smoothness constant (25) nicely interpolates between L, the smoothness constant of f, and $L_{\max} = \max_{i\in[n]} L_i$.

Lemma A.3 ($\mathcal{L}$ for $b-$nice sampling, Proposition 3.8 in [12]) *Let v be a sampling vector based on the b-nice sampling defined in A.4. If for all $i \in [n]$, f_i is convex and L_i-smooth, then (25) holds with*

$$\mathcal{L}(b) = \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \frac{n}{b}\frac{b-1}{n-1}L,$$

where L is the smoothness constant of f and $L_{\max} = \max_{i\in[n]} L_i$.

Appendix B: Notable Corollaries of Theorem 3.1

In this section, we present corollaries of Theorem 3.1 for five algorithms:

- *SGD* with arbitrary sampling (Algorithm 3).
- *DIANA* (Algorithm 4).
- *L-SVRG* with arbitrary sampling (Algorithm 5), and minibatch *L-SVRG* as a special case (Algorithm 2).
- Minibatch *SAGA* (Algorithm 1).
- Miniblock *SEGA* (Algorithm 6).

This means that for each method, we will present the constants which satisfy Assumption 2 and specialize Theorem 3.1 using these constants.

Appendix B.1: SGD with Arbitrary Sampling

Algorithm 3 SGD-AS

```
Parameters step sizes $(\gamma_k)_k$, a sampling vector $v \sim \mathcal{D}$
Initialization $x_0 \in \mathbb{R}^d$
for $k = 1, 2, \ldots$ do
    Sample $v_k \sim \mathcal{D}$
    $g_k = \nabla f_{v_k}(x_k)$
    $x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k - \gamma_k g_k)$
end for
```

Lemma B.1 *The iterates of Algorithm 3 satisfy Assumption 2 with*

$$\sigma_k^2 = 0$$

and constants:

$$A = 2\mathcal{L},\ \ B = 0,\ \ \rho = 1,\ \ C = 0,\ \ D_1 = 2\sigma^2,\ \ D_2 = 0,$$

where $\mathcal{L}$ is defined in (25) *and σ^2 in* (26).

Proof See Lemma A.2 in [11]. □

Using the constants given in the above lemma, we have the following immediate corollary of Theorem 3.1.

Corollary B.1 *Assume that f has a finite sum structure* (4) *and that Assumption 1 holds. Let $(\gamma_k)_{k\geq 0}$ be a decreasing, strictly positive sequence of step sizes chosen such that*

$$0 < \gamma_0 < \min\left\{\frac{1}{4\mathcal{L}}, \frac{1}{L}\right\}.$$

Then, from Theorem 3.1 and Lemma B.1, we have that the iterates given by Algorithm 3 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma_0\left(F(x_0) - F(x_*)\right) + 4\sigma^2\sum_{k=0}^{t-1}\gamma_k^2}{2\sum_{i=0}^{t-1}\left(1 - 4\gamma_i\mathcal{L}\right)\gamma_i},$$

where $\bar{x}_t \overset{def}{=} \sum_{k=0}^{t-1}\frac{(1-4\gamma_k\mathcal{L})\gamma_k}{\sum_{i=0}^{t-1}(1-4\gamma_i\mathcal{L})\gamma_i}x_k$.

Appendix B.2: DIANA

A complete description of the *DIANA* algorithm can be found in [26].

To analyze the *DIANA* algorithm (Algorithm 4), we introduce quantization operators.

Definition B.1 (*w-quantization operator, Definition 4 in*[26]) Let $w > 0$. A random operator $Q : \mathbb{R}^d \to \mathbb{R}$ with the properties:

$$\mathbb{E}\left[Q(x)\right] = x, \quad \mathbb{E}\left[\|Q(x)\|^2\right] \leq (1 + w)\|x\|^2, \tag{28}$$

for all $x \in \mathbb{R}^d$ is called a w-quantization operator.

Several examples of quantization operators can be found in [26].

For convenience, we repeat the statement of Lemma 4.2 below.

Lemma B.2 *Assume that f has a finite sum structure and that Assumption 1 holds. The iterates of DIANA (Algorithm 4) satisfy Assumption 2 with constants:*

$$A = \left(1 + \frac{2w}{n}\right)L_{\max},\ \ B = \frac{2w}{n},\ \ \rho = \alpha,\ \ C = L_{\max}\alpha,\ \ D_1 = \frac{(1+w)\sigma^2}{n},\ \ D_2 = \alpha\sigma^2,$$

Algorithm 4 DIANA

```
Parameters $w$-quantization operator $Q$, Learning rates $\alpha > 0$ and $\gamma > 0$, initial vectors $x^0, h_1^0, \dots, h_n^0 \in \mathbb{R}^d$ and $h^0 = \frac{1}{n}\sum_{i=1}^{n} h_i^0$
Initialization $x^0, h_1^0, \dots, h_n^0 \in \mathbb{R}^d$
Set $h^0 = \frac{1}{n}\sum_{i=1}^{n} h_i^0$
for $k = 1, 2, \dots$ do
    Broadcast $x_k$ to all workers.
    for $k = 1, 2, \dots$ do
        Sample $g_i^k$ such that $\mathbb{E}_k\left[g_i^k\right] = \nabla f_i(x_k)$
        $\Delta_i^k = g_i^k - h_i^k$
        Sample $\hat{\Delta}_i^k \sim Q(\Delta_i^k)$
        $h_i^{k+1} = h_i^k + \alpha\Delta_i^k$
        $\hat{g}_i^k = h_i^k + \hat{\Delta}_i^k$
    end for
    $\hat{\Delta}^k = \frac{1}{n}\sum_{i=1}^{n} \Delta_i^k$
    $g_k = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_i^k = h^k + \hat{\Delta}^k$
    $x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k - \gamma_k g_k)$
    $h^{k+1} = \frac{1}{n}\sum_{i=1}^{n} h_i^{k+1} = h^k + \alpha\hat{\Delta}^k$
end for
```

where $w > 0$ and $\alpha \leq \frac{1}{1+w}$ are parameters of Algorithm 4 and σ^2 is such that

$$\forall k \in \mathbb{N}, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\|g_i^k - \nabla f(x_k)\right\|^2\right] \leq \sigma^2.$$

Proof See Lemma A.12 in [11]. □

Now using the constants given in the above lemma in Theorem 3.1 gives the following corollary.

Corollary B.2 *Assume that f has a finite sum structure (4) and that Assumption 1 holds. Let $(\gamma_k)_{k\geq 0}$ be a decreasing, strictly positive sequence of step sizes chosen such that*

$$0 < \gamma_0 < \frac{1}{2(1 + \frac{4w}{n})L_{\max}}.$$

Then, from Theorem 3.1 and Lemma B.2, we have that the iterates given by Algorithm 4 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \\ \leq \frac{\|x_0 - x_*\|^2 + 2\gamma_0\left(F(x_0) - F(x_*) + \frac{2w\gamma_0}{\alpha n}\sigma_0^2\right) + \frac{2(1+5w)\sigma^2}{n}\sum_{k=0}^{t-1}\gamma_k^2}{2\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i},$$

where $\eta \overset{def}{=} 2(1+\frac{4w}{n})L_{\max}$, $\bar{x}_t \overset{def}{=} \sum_{k=0}^{t-1}\frac{(1-\gamma_k\eta)\gamma_k}{\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i}x_k$ *and* $\delta_0 \overset{def}{=} F(x_0) - F(x_*)$.

Appendix B.3: L-SVRG with Arbitrary Sampling

Algorithm 5 L-SVRG-AS

```
Parameters step size $\gamma$, sampling vector $v \sim \mathcal{D}$
Initialization $w_0 = x_0 \in \mathbb{R}^d$
for $k = 1, 2, \ldots$ do
    Sample $v_k \sim \mathcal{D}$
    $g_k = \nabla f_{v_k}(x_k) - \nabla f_{v_k}(w_k) + \nabla f(w_k)$
    $x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma g_k)$
    $w_{k+1} = \begin{cases} x_k & \text{with probability } p \\ w_k & \text{with probability } 1-p \end{cases}$
end for
```

Lemma B.3 *If Assumption 1 holds then the iterates of Algorithm 5 satisfy*

$$\mathbb{E}_k\left[\|g_k - \nabla f(x_*)\|^2\right] \leq 4\mathcal{L}D_f(x_k, x_*) + 2\sigma_k^2$$
$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-p)\sigma_k^2 + 2p\mathcal{L}D_f(x_k, x_*),$$

where

$$\sigma_k^2 = \mathbb{E}_{\mathcal{D}}\left[\left\|\nabla f_{v_k}(w_k) - \nabla f_{v_k}(x_*) - (\nabla f(w_k) - \nabla f(x_*))\right\|^2\right]$$

and $\mathcal{L}$ *is defined in* (25).

Proof By Lemma A.1 we have that (25) holds with $\mathcal{L} > 0$. Furthermore

$$\begin{aligned}\mathbb{E}_k\left[\|g_k\|^2\right] &= \mathbb{E}_k\left[\left\|\nabla f_{v_k}(x_k) - \nabla f_{v_k}(w_k) + \nabla f(w_k) - \nabla f(x_*)\right\|^2\right] \\ &\leq 2\mathbb{E}_k\left[\left\|\nabla f_{v_k}(x_k) - \nabla f_{v_k}(x_*)\right\|^2\right] \\ &\quad + 2\mathbb{E}_k\left[\left\|\nabla f_{v_k}(w_k) - \nabla f_{v_k}(x_*) - (\nabla f(w_k) - \nabla f(x_*))\right\|^2\right],\end{aligned}$$

where we used in the inequality that for all $a, b \in \mathbb{R}^d$, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. Thus,

$$\mathbb{E}_k\left[\|g_k\|^2\right] \overset{(25)}{\le} 4\mathcal{L}D_f(x_k, x_*) + 2\sigma_k^2.$$

Moreover,

$$\begin{aligned}\mathbb{E}_k\left[\sigma_{k+1}\right] &= (1-p)\sigma_k^2 + p\mathbb{E}_k\left[\left\|\nabla f_{v_k}(x_k) - \nabla f_{v_k}(x_*) - (\nabla f(x_k) - \nabla f(x_*))\right\|^2\right] \\ &\overset{(25)}{\le} (1-p)\sigma_k^2 + 2p\mathcal{L}D_f(x_k, x_*),\end{aligned}$$

where we also used in the last inequality that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] = \mathbb{E}\left[\|X\|^2\right] - \|\mathbb{E}[X]\|^2 \le \mathbb{E}\left[\|X\|^2\right]$. □

We have the following immediate consequence of the previous lemma.

Lemma B.4 *If Assumption 1 holds then the iterates of Algorithm 5 satisfy Assumption 2 with*

$$\sigma_k^2 = \mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x_k) - \nabla f_v(w_k) + \nabla f(w_k)\|^2\right]$$

and constants

$$A = 2\mathcal{L},\ B = 2,\ \rho = p,\ C = p\mathcal{L},\ D_1 = D_2 = 0,$$

where $\mathcal{L}$ is defined in (25).

Using the constant derived in Lemma B.4 in Theorem 3.1 gives the following corollary.

Corollary B.3 *Assume that f has a finite sum structure (4) and that Assumption 1 holds. Let $\gamma_k = \gamma$ for all $k \in \mathbb{N}$, where*

$$0 < \gamma < \min\left\{\frac{1}{8\mathcal{L}}, \frac{1}{L}\right\}.$$

Then, from Theorem 3.1 and Lemma B.4, we have that the iterates given by Algorithm 5 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \le \frac{\|x_0 - x_*\|^2 + 2\gamma\left(F(x_0) - F(x_*) + \frac{2\gamma}{p}\sigma_0^2\right)}{2\gamma\left(1 - 8\gamma\mathcal{L}\right)t},$$

where $\bar{x}_t \overset{def}{=} \frac{1}{t}\sum_{k=0}^{t-1} x_k$ and where $\mathcal{L}$ is defined in (25).

Appendix B.3.1: *b*-L-SVRG

As we demonstrated in Section A.3, we can specialize the results derived for arbitrary sampling to minibatching without replacement by using a $b-$nice sampling defined in Definition A.4 and the corresponding sampling vector (27).

Indeed, using Algorithm 5 with b-nice sampling is equivalent to using Algorithm 2. Thus, we have the following lemma.

Corollary B.4 *From Lemma B.4, we have that the iterates of Algorithm 2 satisfy Assumption 2 with constants:*

$$A = 2\mathcal{L}(b),\ B = 2,\ \rho = p,\ C = p\mathcal{L}(b),\ D_1 = D_2 = 0,$$

where $\mathcal{L}(b)$ is defined in (17).

A convergence result for Algorithm 2 can be easily concluded from Corollary B.3, with $\mathcal{L}(b)$ in place of $\mathcal{L}$.

Appendix B.4: *b*-SAGA

Lemma 5.1 in the main text is a consequence of the following lemma.

Lemma B.5 *Consider the iterates of Algorithm 1. We have:*

$$\mathbb{E}_k\left[\|g_k\|^2\right] \le 4\mathcal{L}(b)\left(f(x_k) - f(x_*)\right) + 2\sigma_k^2 \tag{29}$$

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \le \left(1 - \frac{b}{n}\right)\sigma_k^2 + 2\frac{b\zeta(b)}{n}\left(f(x_k) - f(x_*)\right), \tag{30}$$

where:

$$\sigma_k^2 = \frac{1}{nb}\frac{n-b}{n-1}\|J_k - \nabla H(x_*)\|_{\mathrm{Tr}}^2 \quad \textit{and} \quad \zeta(b) \overset{def}{=} \frac{1}{b}\frac{n-b}{n-1}L_{\max},$$

with $\|Z\|_{\mathrm{Tr}}^2 = \mathrm{tr}(Z^\top Z)$ for any $Z \in \mathbb{R}^{d\times n}$.

Proof The inequality (29) corresponds to Lemma 3.10 and (30) to Lemma 3.9 in [13]. □

The previous Lemma gives us the constants for Assumption 2 for Algorithm 1.

Lemma B.6 *The iterates of Algorithm 1 satisfy Assumption 2 with*

$$\sigma_k^2 = \frac{1}{nb}\frac{n-b}{n-1}\|J_k - \nabla H(x_*)\|_{\mathrm{Tr}}^2$$

and constants

$$A = 2\mathcal{L}(b),\ B = 2,\ \rho = \frac{b}{n},\ C = \frac{b\zeta(b)}{n},\ D_1 = D_2 = 0.$$

Using the constant derived in Lemma B.6 in Theorem 3.1 gives the following corollary.

Corollary B.5 *Assume that f has a finite sum structure (4) and that Assumption 1 holds. Choose for all $k \in \mathbb{N}$ $\gamma_k = \gamma$, where*

$$0 < \gamma < \frac{1}{2(2\mathcal{L}(b) + \zeta(b))}.$$

Then, from Theorem 3.1 and Lemma B.6, we have that the iterates given by Algorithm 1 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma\left(F(x_0) - F(x_*) + \frac{2n\gamma}{b}\sigma_0^2\right)}{2\gamma\left(1 - 2\gamma\left(2\mathcal{L}(b) + 2\zeta(b)\right)\right)t},$$

where $\bar{x}_t \overset{def}{=} \frac{1}{t}\sum\limits_{k=0}^{t-1} x_k$.

Appendix B.5: *b*-SEGA

Lemma B.7 *Consider the iterates of Algorithm 6. We have:*

$$\mathbb{E}_k\left[\|g_k\|^2\right] \leq \frac{4dL}{b} D_f\left(x_k, x_*\right) + 2\left(\frac{d}{b} - 1\right)\sigma_k^2$$
$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq \left(1 - \frac{b}{d}\right)\sigma_k^2 + \frac{2bL}{d} D_f\left(x_k, x_*\right),$$

where:

$$\sigma_k^2 = \|h_k - \nabla f(x_*)\|^2.$$

Proof Let S be a random miniblock s.t. $\mathbb{P}(S = B) = \frac{1}{\binom{n}{b}}$ for any $B \subseteq [n]$ s.t. $|B| = b$. Then, for any vector $a = [a_1, \ldots, a_n] \in \mathbb{R}^d$, we have:

$$\mathbb{E}\left[\|I_S a\|^2\right] = \frac{b}{d}\|a\|^2 \quad \text{and} \quad \mathbb{E}\left[\left\|(I - \frac{d}{b} I_S)a\right\|^2\right] = \left(\frac{d}{b} - 1\right)\|a\|^2. \tag{31}$$

Indeed,

$$\mathbb{E}\left[\|I_S a\|^2\right] = \mathbb{E}\left[\sum_{i \in S} a_i^2\right] = \sum_{B \subseteq [d], |B| = b} \mathbb{P}(S = B)\sum_{i \in B} a_i^2 = \frac{1}{\binom{d}{b}} \sum_{B \subseteq [d], |B| = b} \sum_{i=1}^{d} a_i^2 \mathbb{1}_B(i)$$
$$= \frac{1}{\binom{d}{b}} \sum_{i=1}^{d} a_i^2 \sum_{B \subseteq [d], |B| = b} \mathbb{1}_B(i) = \frac{\binom{d-1}{b-1}}{\binom{d}{b}} \sum_{i=1}^{d} a_i^2 = \frac{b}{d}\|a\|^2,$$

where we used that $|B \in [d] : |B| = b \wedge i \in B| = \binom{d-1}{b-1}$. And

$$\left\|(I - \frac{d}{b} I_S)a\right\|^2 = \sum_{i \in S} \left(1 - \frac{d}{b}\right)^2 a_i^2 + \sum_{i \notin S} a_i^2 = \frac{d^2 - 2bd}{b^2} \sum_{i \in S} a_i^2 + \|a\|^2$$
$$= \frac{d^2 - 2bd}{b^2} \|I_S a\|^2 + \|a\|^2.$$

Thus,

$$\mathbb{E}\left[\left\|(I - \frac{d}{b} I_S)a\right\|^2\right] = \left(\frac{d^2 - 2bd}{b^2} \frac{b}{d} + 1\right) \|a\|^2 = \left(\frac{d}{b} - 1\right) \|a\|^2.$$

We have

$$\begin{aligned}
&\mathbb{E}_k\left[\|g_k - \nabla f(x_*)\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{d}{b} I_{B_k}(\nabla f(x_k) - \nabla f(x_*)) + \left(I - \frac{d}{b} I_{B_k}\right)(h_k - \nabla f(x_*))\right\|^2\right] \\
&\leq \frac{2d^2}{b^2} \mathbb{E}_k\left[\left\|I_{B_k}(\nabla f(x_k) - \nabla f(x_*))\right\|^2\right] \\
&\quad + 2\, \mathbb{E}_k\left[\left\|\left(I - \frac{d}{b} I_{B_k}\right)(h_k - \nabla f(x_*))\right\|^2\right] \\
&\overset{(31)}{=} \frac{2d}{b} \|\nabla f(x_k) - \nabla f(x_*)\|^2 + 2\left(\frac{d}{b} - 1\right) \|h_k - \nabla f(x_*)\|^2.
\end{aligned}$$

where we used in the first inequality that for all $a, b \in \mathbb{R}^d$, $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Thus, using the fact that f is L-smooth, we have

$$\mathbb{E}_k\left[\|g_k\|^2\right] \leq \frac{4dL}{b} D_f\,(x_k, x_*) + 2\left(\frac{d}{b} - 1\right) \sigma_k^2.$$

Moreover,

$$\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \mathbb{E}_k\left[\|h_{k+1} - \nabla f(x_*)\|^2\right] \\
&= \mathbb{E}_k\left[\left\|I_{B_k^c}(h_k - \nabla f(x_*)) + I_{B_k}(\nabla f(x_k) - \nabla f(x_*))\right\|^2\right] \\
&\overset{(31)}{=} \left(1 - \frac{b}{d}\right) \|h_k - \nabla f(x_*)\|^2 + \frac{b}{d} \|\nabla f(x_k) - \nabla f(x_*)\|^2 \\
&\quad + 2\left\langle I_{B_k^c}(h_k - \nabla f(x_*)), I_{B_k}(\nabla f(x_k) - \nabla f(x_*))\right\rangle \\
&= \left(1 - \frac{b}{d}\right) \|h_k - \nabla f(x_*)\|^2 + \frac{b}{d} \|\nabla f(x_k) - \nabla f(x_*)\|^2
\end{aligned}$$

$$
\begin{aligned}
&+2\left\langle \underbrace{I_{B_k} I_{B_k^c}}_{=0}(h_k - \nabla f(x_*)), \nabla f(x_k) - \nabla f(x_*) \right\rangle \\
&\leq \left(1 - \frac{b}{d}\right) \|h_k - \nabla f(x_*)\|^2 + \frac{2bL}{d} D_f\left(x_k, x_*\right),
\end{aligned}
$$

where we used in the last inequality the $L-$smoothness of f. □

Lemma B.8 *From Lemma B.7, we have that the iterates of Algorithm 6 satisfy Assumption 2 and Eq. (14) with*

$$\sigma_k^2 = \|h_k - \nabla f(x_*)\|^2$$

and constants:

$$A = \frac{2dL}{b},\; B = 2\left(\frac{d}{b} - 1\right),\; \rho = \frac{b}{d},\; C = \frac{bL}{d},\; D_1 = D_2 = 0,\; G = 0.$$

Using the constant derived in Lemma B.8 in Theorem 3.1 gives the following corollary.

Corollary B.6 *Assume that f satisfies Assumption 1. Choose for all $k \in \mathbb{N}$, $\gamma_k = \gamma$, where*

$$0 < \gamma < \frac{1}{4(\frac{2d}{b} - 1)L}.$$

Then, from Theorem 3.1 and Lemma B.8, we have that the iterates given by Algorithm 6 verify

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{\|x_0 - x_*\|^2 + 2\gamma\left(F(x_0) - F(x_*) + \frac{2d}{b}\left(\frac{d}{b} - 1\right)\gamma\sigma^2\right)}{2\gamma\left(1 - 4\gamma\left(\frac{2d}{b} - 1\right)\right)t},$$

where $\bar{x}_t \overset{def}{=} \frac{1}{t}\sum_{k=0}^{t-1} x_k$.

Appendix C: Proofs for Sect. 3

Appendix C.1: Proof of Theorem 3.1

Before proving Theorem 3.1, we present several useful lemmas.

Lemma C.1 (Bounding the gradient variance) *Assuming that the g_k are unbiased and that Assumption 2 holds, we have*

$$\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \leq 2AD_f(x_k, x_*) + B\sigma_k^2 + D_1. \tag{32}$$

Proof Starting from the left-hand side of (32), we have

$$\begin{aligned}\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] &= \mathbb{E}\left[\|g_k - \nabla f(x_*) - (\nabla f(x_k) - \nabla f(x_*))\|^2\right] \\ &= \mathbb{E}\left[\|g_k - \nabla f(x_*) - \mathbb{E}\left[g_k - \nabla f(x_*)\right]\|^2\right] \\ &\leq \mathbb{E}\left[\|g_k - \nabla f(x_*)\|^2\right] \leq 2AD_f(x_k, x_*) + B\sigma_k^2 + D_1,\end{aligned}$$

where we used that $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] = \mathbb{E}\left[\|X\|^2\right] - \|\mathbb{E}[X]\|^2 \leq \mathbb{E}\left[\|X\|^2\right]$ for any random variable X. □

Lemma C.2 (Lemma 8 in [5]) *Suppose that Assumption 1 holds and let* $\gamma \in \left(0, \frac{1}{L}\right]$, *then for all* $x, y \in \mathbb{R}^d$ *and* $p = \mathrm{prox}_{\gamma g}(y)$ *we have,*

$$-2\gamma\left(F(p) - F(x_*)\right) \geq \|p - z\|^2 + 2\left\langle p - x_*, x - \gamma\nabla f(x) - y\right\rangle - \|x_* - x\|^2. \tag{33}$$

Proof We leave the proof to Section F.3. □

Lemma C.3 *For any* $x \in \mathbb{R}^d$ *and minimizer* x_* *of* F, *we have,*

$$D_f(x, x_*) \leq F(x) - F(x_*). \tag{34}$$

Proof Because x_* is a minimizer of F, we have that $-\nabla f(x_*) \in \partial R(x_*)$. By the definition of subgradients, we have

$$R(x_*) + \langle -\nabla f(x_*), x - x_*\rangle \leq R(x).$$

Rearranging gives

$$-\langle \nabla f(x_*), x - x_*\rangle \leq R(x) - R(x_*).$$

Adding $f(x) - f(x_*)$ to both sides we have,

$$f(x) - f(x_*) - \langle \nabla f(x_*), x - x_*\rangle \leq f(x) + R(x) - (f(x_*) + R(x_*)) = F(x) - F(x_*).$$

Now note that the on the left-hand side we have the Bregman divergence $D_f(x, x_*)$. □

Definition C.1 Given a stepsize $\gamma > 0$, the prox-grad mapping is defined as:

$$T_\gamma(x) \overset{\text{def}}{=} \mathrm{prox}_{\gamma R}\left(x - \gamma\nabla f(x)\right).$$

For the ease of exposition, we restate Theorem 3.1.

Theorem C.1 *Suppose that Assumptions 2 and 1 hold. Let* $M \overset{def}{=} B/\rho$ *and let* $(\gamma_k)_{k\geq 0}$ *be a decreasing, strictly positive sequence of step sizes chosen such that*

$$0 < \gamma_0 < \frac{1}{2(A + MC)}.$$

The iterates given by (2) converge according to

$$\mathbb{E}\left[F(\bar{x}_t) - F(x_*)\right] \leq \frac{V_0 + 2\gamma_0\delta_0 + 2\left(D_1 + 2MD_2\right)\sum_{k=0}^{t-1}\gamma_k^2}{2\sum_{i=0}^{t-1}\left(1 - 2\gamma_i\left(A + MC\right)\right)\gamma_i},$$

where $\bar{x}_t \overset{def}{=} \sum_{k=0}^{t-1}\frac{(1-\gamma_k\eta)\gamma_k}{\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i}x_k$ *and* $V_0 \overset{def}{=} \|x_0 - x_*\|^2 + 2\gamma_0^2 M\sigma_0^2$ *and* $\delta_0 \overset{def}{=} F(x_0) - F(x_*)$.

Proof Let x_* be a minimizer of F. Using (33) from Lemma C.2 with $y = x_k - \gamma_k g_k$, $x = x_k$ and $\gamma = \gamma_k$ gives

$$\begin{aligned}-2\gamma_k\left(F(x_{k+1}) - F(x_*)\right) \geq \|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \\ + 2\gamma_k\left\langle x_{k+1} - x_*, g_k - \nabla f(x_k)\right\rangle.\end{aligned}$$

Multiplying both sides by -1 results in

$$\begin{aligned}2\gamma_k\left(F(x_{k+1}) - F(x_*)\right) \leq \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 \\ + 2\gamma_k\left\langle x_{k+1} - x_*, \nabla f(x_k) - g_k\right\rangle.\end{aligned} \tag{35}$$

Now focusing on the last term in the above and consider the straightforward decomposition

$$\begin{aligned}\left\langle x_{k+1} - x_*, \nabla f(x_k) - g_k\right\rangle = \left\langle x_{k+1} - T_{\gamma_k}(x_k), \nabla f(x_k) - g_k\right\rangle \\ + \left\langle T_{\gamma_k}(x_k) - x_*, \nabla f(x_k) - g_k\right\rangle.\end{aligned} \tag{36}$$

By Cauchy Schwartz we have that

$$\left\langle x_{k+1} - T_{\gamma_k}(x_k), \nabla f(x_k) - g_k\right\rangle \leq \left\|x_{k+1} - T_{\gamma_k}(x_k)\right\|\left\|g_k - \nabla f(x_k)\right\|. \tag{37}$$

Now using the nonexpansivity of the proximal operator

$$\begin{aligned}\left\|x_{k+1} - T_{\gamma_k}(x_k)\right\| &= \left\|\operatorname{prox}_{\gamma_k R}\left(x_k - \gamma_k g_k\right) - \operatorname{prox}_{\gamma_k R}\left(x_k - \gamma_k\nabla f(x_k)\right)\right\| \\ &\leq \left\|(x_k - \gamma_k g_k) - (x_k - \gamma_k\nabla f(x_k))\right\| = \gamma_k\left\|g_k - \nabla f(x_k)\right\|.\end{aligned}$$

Using this in (37), we have

$$\left\langle x_{k+1} - T_{\gamma_k}(x_k), \nabla f(x_k) - g_k\right\rangle \leq \gamma_k\left\|g_k - \nabla f(x_k)\right\|^2. \tag{38}$$

Using (38) in (36) and taking expectation conditioned on x_k, and using $\mathbb{E}_k[\cdot] \overset{\text{def}}{=} \mathbb{E}[\cdot \mid x_k]$ for shorthand, we have

$$\begin{aligned} \mathbb{E}_k\left[\langle x_{k+1} - x_*, g_k - \nabla f(x_k)\rangle\right] &\le \gamma_k \cdot \mathbb{E}_k\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &\quad + \left\langle T_{\gamma_k}(x_k) - x_*, \underbrace{\mathbb{E}_k\left[\nabla f(x_k) - g_k\right]}_{=0}\right\rangle \qquad (39) \\ &= \gamma_k \cdot \mathbb{E}_k\left[\|g_k - \nabla f(x_k)\|^2\right]. \end{aligned}$$

Let $r_k \overset{\text{def}}{=} x_k - x_*$. Taking expectation conditioned on x_k in (35) and using (39), we have

$$2\gamma_k \mathbb{E}_k\left[F(x_{k+1} - F(x_*))\right] \le \|r_k\|^2 - \mathbb{E}_k\left[\|r_{k+1}\|^2\right] + 2\gamma_k^2 \mathbb{E}_k\left[\|g_k - \nabla f(x_k)\|^2\right].$$

Using (8) from Assumption 2, we have

$$\begin{aligned} &2\gamma_k \mathbb{E}_k\left[F(x_{k+1}) - F(x_*)\right] \\ &\quad \le \|r_k\|^2 - \mathbb{E}_k\left[\|r_{k+1}\|^2\right] + 2\gamma_k^2\left(2AD_f(x_k, x_*) + B\sigma_k^2 + D_1\right). \end{aligned}$$

Let $V_k \overset{\text{def}}{=} \|r_k\|^2 + 2\,M\gamma_k^2\sigma_k^2$ where $M = \frac{B}{\rho}$, then

$$\begin{aligned} 2\gamma_k \mathbb{E}_k\left[F(x_{k+1}) - F(x_*)\right] &\le V_k - \mathbb{E}_k\left[V_{k+1}\right] + 4\gamma_k^2 AD_f(x_k, x_*) + 2\gamma_k^2 D_1 \\ &\quad + \gamma_k^2\left(2B - 2M\right)\sigma_k^2 + 2M\gamma_{k+1}^2 \mathbb{E}\left[\sigma_{k+1}^2\right]. \end{aligned} \tag{40}$$

Since $\gamma_{k+1} \le \gamma_k$, we have that

$$\begin{aligned} 2\gamma_{k+1} \mathbb{E}_k\left[F(x_{k+1}) - F(x_*)\right] &\le V_k - \mathbb{E}_k\left[V_{k+1}\right] + 4\gamma_k^2 AD_f(x_k, x_*) + 2\gamma_k^2 D_1 \\ &\quad + \gamma_k^2\left(2B - 2M\right)\sigma_k^2 + 2M\gamma_k^2 \mathbb{E}\left[\sigma_{k+1}^2\right]. \end{aligned}$$

Using (9) from Assumption 2, we have

$$\begin{aligned} 2\gamma_k^2\left(B - M\right)\sigma_k^2 + 2M\gamma_k^2 \mathbb{E}_k\left[\sigma_{k+1}^2\right] &\le 2\gamma_k^2\left(B - M + M(1-\rho)\right)\sigma_k^2 + 4M\gamma_k^2 CD_f(x_k, x_*) \\ &\quad + 2M\gamma_k^2 D_2 \\ &= 2\gamma_k^2 \underbrace{\left(B - \rho M\right)}_{=0}\sigma_k^2 + 4M\gamma_k^2 CD_f(x_k, x_*) + 2M\gamma_k^2 D_2 \\ &\le 4M\gamma_k^2 CD_f(x_k, x_*) + 2M\gamma_k^2 D_2. \end{aligned} \tag{41}$$

Using (41) in (40) gives

$$2\gamma_{k+1}\mathbb{E}_k\left[F(x_{k+1})-F(x_*)\right]\leq V_k-\mathbb{E}_k\left[V_{k+1}\right]+2\gamma_k^2\left(2A+2MC\right)D_f(x_k,x_*) \\ +2\gamma_k^2\left(D_1+MD_2\right). \tag{42}$$

Let $\eta \overset{\text{def}}{=} 2A+2MC$. Using (34) in (42) we have,

$$2\gamma_{k+1}\mathbb{E}_k\left[F(x_{k+1})-F(x_*)\right] \\ \leq V_k-\mathbb{E}_k\left[V_{k+1}\right]+2\gamma_k^2\eta\left(F(x_k)-F(x_*)\right)+2\gamma_k^2\left(D_1+MD_2\right).$$

Using the abbreviation $\delta_k=F(x_k)-F(x_*)$ gives

$$2\gamma_{k+1}\mathbb{E}_k\left[\delta_{k+1}\right]\leq V_k-\mathbb{E}_k\left[V_{k+1}\right]+2\gamma_k^2\eta\delta_k+2\gamma_k^2\left(D_1+MD_2\right).$$

Taking expectation,

$$2\gamma_{k+1}\mathbb{E}\left[\delta_{k+1}\right]\leq\mathbb{E}\left[V_k\right]-\mathbb{E}\left[V_{k+1}\right]+2\gamma_k^2\eta\mathbb{E}\left[\delta_k\right]+2\gamma_k^2\left(D_1+MD_2\right),$$

summing over $k=0,\ldots,t-1$ and using telescopic cancellation gives

$$2\sum_{k=1}^{t}\gamma_k\mathbb{E}\left[\delta_k\right]\leq V_0-\mathbb{E}\left[V_t\right]+2\eta\sum_{k=0}^{t-1}\gamma_k^2\mathbb{E}\left[\delta_k\right]+2\left(D_1+MD_2\right)\sum_{k=0}^{t-1}\gamma_k^2.$$

Adding $2\gamma_0\delta_0$ to both sides of the above inequality and rearranging,

$$2\sum_{k=0}^{t-1}\gamma_k(1-\eta\gamma_k)\mathbb{E}\left[\delta_k\right]\leq V_0-\mathbb{E}\left[V_t\right]+2\gamma_0\delta_0+2\left(D_1+MD_2\right)\sum_{k=0}^{t-1}\gamma_k^2$$

where we also used that $V_t\geq 0$ and $\delta_t\geq 0$. By the choice of γ_0 we have $1-\gamma_0\eta>0$, and since $(\gamma_i)_i$ is a decreasing sequence, we have $1-\gamma_i\eta>0$ for all i. Hence, dividing both sides by $2\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i$, we have

$$\sum_{k=0}^{t-1}w_k\mathbb{E}\left[\delta_k\right]\leq\frac{V_0+2\gamma_0\delta_0}{2\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i}+\left(D_1+2MD_2\right)\frac{\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{i=1}^{t}(1-\gamma_i\eta)\gamma_i},$$

where $w_k \overset{\text{def}}{=} \frac{(1-\gamma_k\eta)\gamma_k}{\sum_{i=0}^{t-1}(1-\gamma_i\eta)\gamma_i}$ for all $k \in \{0, \ldots, t-1\}$. Note that $\sum_{k=0}^{t-1} w_k = 1$ and $w_k \geq 0$ for all $k \in \{0, \ldots, t-1\}$. Hence, since F is convex, we can use Jensen's inequality to conclude

$$\mathbb{E}\left[F(\bar{x}^k) - F(x_*)\right] = \mathbb{E}\left[F\left(\sum_{k=0}^{t-1} w_k x_k\right) - F(x_*)\right]$$
$$\leq \sum_{k=0}^{t-1} w_k \mathbb{E}\left[\delta_k\right] \leq \frac{V_0 + 2\gamma_0\delta_0}{2\sum_{i=0}^{t-1}\left(1-\gamma_i\eta\right)\gamma_i} + \frac{(D_1 + 2MD_2)\sum_{k=0}^{t-1}\gamma_k^2}{\sum_{i=0}^{t-1}\left(1-\gamma_i\eta\right)\gamma_i}.$$

Writing out the definition of δ_0 yields the theorem's statement. □

Appendix D: Proofs for Sect. 4

Appendix D.2: Proof of Corollary 4.2

Proof Note that, using the integral bound, we have:

$$\sum_{k=0}^{t-1} \gamma_k^2 \leq \gamma^2\left(\log(t) + 1\right)$$
$$\sum_{k=0}^{t-1} \gamma_k \geq 2\gamma\left(\sqrt{t} - 1\right).$$

Moreover, note that since $\gamma_k \leq \frac{1}{4(A+MC)}$, we have $1 - 2\gamma_k(A + MC) \geq \frac{1}{2}$ for all $k \in \mathbb{N}$. Thus

$$\frac{1}{2\sum_{k=0}^{t-1}\gamma_k\left(1-\eta\gamma_k\right)} \leq \frac{1}{2\gamma\left(\sqrt{t}-1\right)},$$

where $\eta \overset{\text{def}}{=} 2(A + MC)$. Corollary 4.2 follows from using these bounds in Equation (10). □

Appendix E: Roofs for Sect. 5

Appendix E.1: Proof of Proposition 5.1

Proof We start by expanding the square:

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\gamma\left\langle g_k, x_k - x_*\right\rangle + \gamma^2\|g_k\|^2.$$

Thus, taking expectation conditioned on x_k, and using $\mathbb{E}_k[\cdot] \overset{\text{def}}{=} \mathbb{E}[\cdot \mid x_k]$ for shorthand, we have

$$\begin{aligned}\mathbb{E}_k\left[\|x_{k+1}-x_*\|^2\right] &= \|x_k-x_*\|^2 - 2\gamma\left\langle \nabla f(x_k), x_k-x_*\right\rangle + \gamma^2\mathbb{E}_k\left[\|g_k\|^2\right] \\ &\overset{(6)+(7)+(8)}{\leq} \|x_k-x_*\|^2 - 2\gamma(1-2\gamma A)\left(f(x_k)-f(x_*)\right) + B\sigma_k^2.\end{aligned}$$

Thus, using (9),

$$\begin{aligned}&\mathbb{E}_k\left[\|x_{k+1}-x_*\|^2\right] + 2M\gamma^2\mathbb{E}_k\left[\sigma_{k+1}^2\right] \\ &\quad\leq \|x_k-x_*\|^2 - 2\gamma(1-2\gamma(A+MC))\left(f(x_k)-f(x_*)\right) + 2M\gamma^2\sigma_k^2.\end{aligned}$$

Thus, rearranging and taking the expectation, we have:

$$\begin{aligned}2\gamma(1-2\gamma(A+MC))\mathbb{E}\left[f(x_k)-f(x_*)\right] \leq \mathbb{E}\left[\|x_k-x_*\|^2\right] - \mathbb{E}\left[\|x_{k+1}-x_*\|^2\right] \\ +2M\gamma^2\left(\mathbb{E}\left[\sigma_k^2\right] - \mathbb{E}\left[\sigma_{k+1}^2\right]\right).\end{aligned}$$

Summing over $k = 0, \ldots, t-1$ and using telescopic cancellation gives

$$\begin{aligned}2\gamma(1-2\gamma(A+MC))\sum_{k=0}^{t-1}\mathbb{E}\left[f(x_k)-f(x_*)\right] \leq \|x_0-x_*\|^2 - \mathbb{E}\left[\|x_k-x_*\|^2\right] \\ +2M\gamma^2\left(\mathbb{E}\left[\sigma_0^2\right] - \mathbb{E}\left[\sigma_{k+1}^2\right]\right).\end{aligned}$$

Ignoring the negative terms in the upper bound, and using Jensen's inequality, we have

$$\mathbb{E}\left[f(\bar{x}_t)-f(x_*)\right] \leq \frac{\|x_0-x_*\|^2 + 2M\gamma^2\sigma_0^2}{2\gamma(1-2\gamma(A+MC))t}.$$

Moreover, notice that if $\gamma \leq \frac{1}{4(A+MC)}$, then $2(1-2\gamma(A+MC)) \geq 1$, which gives (13). □

Appendix E.2: Optimal Minibatch Size for *b*-SAGA (Algorithm 1)

In this section, we present the proofs for Sect. 5.1.

Appendix E.2.1: Proof of Lemma 5.1

Proof For constant A, B, ρ, C, D_1, D_2, see Lemma B.5. Moreover,

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{nb}\frac{n-b}{n-1}\|\nabla H(x_0)-\nabla H(x_*)\|_{\mathrm{Tr}}^2 = \frac{1}{nb}\frac{n-b}{n-1}\sum_{i=1}^n \|\nabla f_i(x_0)-\nabla f_i(x_*)\|^2 \\
&= \frac{1}{b}\frac{n-b}{n-1}\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_0)-\nabla f_i(x_*)\|^2 \\
&\overset{(52)}{\leq} \frac{1}{b}\frac{n-b}{n-1}L_{\max}\left(f(x_0)-f(x_*)\right) \\
&\overset{(5)+(18)}{\leq} \zeta(b)L\|x_0-x_*\|^2.
\end{aligned}$$

Thus, (14) holds with $G=\zeta(b)$. □

Appendix E.2.2: Proof of Proposition 5.2

Proof First, since $\frac{\|x_0-x_*\|^2}{\epsilon}$ does not depend on b, the variations of $K(b)$ are the same as those of

$$Q(b)=\frac{4\left(3(n-b)L_{\max}+2n(b-1)L\right)}{b(n-1)}+\frac{n(n-b)L_{\max}L}{2b\left(3(n-b)L_{\max}+2n(b-1)L\right)}.$$

Let's determine the sign of $Q^{'}(b)$. We have:

$$Q^{'}(b)=\frac{W_1b^2+W_2b+W_3}{4(n-1)\left((2nL-3L_{\max})\,b+\left(\frac{3L_{\max}}{2}-L\right)n\right)^2},$$

where

$$\begin{aligned}
W_1 &= 4\left(2nL-3L_{\max}\right)^3, \\
W_2 &= 8n\left(3L_{\max}-2L\right)\left(2nL-3L_{\max}\right)^2, \\
W_3 &= n^2\left(-108L_{\max}^3+72\left(n+2\right)L_{\max}^2L-\left(n^2+94n+49\right)L^2L_{\max}+32nL^3\right).
\end{aligned}$$

And we have:

$$W_2^2-4W_1W_2=16n^2(n-1)^2L^2L_{\max}\left(2nL-3L_{\max}\right)^3.$$

Case 1 $L_{\max}>\frac{2nL}{3}$. We have $2nL-3L_{\max}<0$. Hence, $W_2^2-4W_1W_2<0$.
Moreover, since $W_1<0$, we have

$$L_{\max}>\frac{2nL}{3} \implies K'(b)<0.$$

Thus,

$$L_{\max} > \frac{2nL}{3} \implies b^* = n$$

Case 2 $L_{\max} \leq \frac{2nL}{3}$. Then, $W_2^2 - 4W_1W_2 \geq 0$ and $K'(b) = 0$ has at least one solution. We are now going to examine whether or not $K(b)$ is convex. We have:

$$Q^{''}(b) = \frac{2n^2(n-1)L_{\max}L^2\left(2nL - 3L_{\max}\right)}{\left(\left(2nL - 3L_{\max}\right)b + \left(3L_{\max} - 2L\right)n\right)^3} \geq 0.$$

Thus, $K(b)$ is convex. $K^{'}(b) = 0$ has two solutions:

$$\begin{aligned} b_1 &= \frac{n\left((n-1)L\sqrt{L_{\max}} - 2\sqrt{2nL - 3L_{\max}}(3L_{\max} - 2L)\right)}{2(2nL - 3L_{\max})^{\frac{3}{2}}}, \\ b_2 &= \frac{-n\left((n-1)L\sqrt{L_{\max}} + 2\sqrt{2nL - 3L_{\max}}(3L_{\max} - 2L)\right)}{2(2nL - 3L_{\max})^{\frac{3}{2}}}. \end{aligned}$$

But since $b_2 \leq 0$, we have that:

$$L_{\max} \leq \frac{2nL}{3} \implies b^* = \begin{cases} 1 & \text{if } b_1 < 2 \\ \lfloor b_1 \rfloor & \text{if } 2 \leq b_1 < n \\ n & \text{if } b_1 \geq n \end{cases}.$$

□

Appendix E.3: Optimal Minibatch Size for *b*-L-SVRG (Algorithm 2)

In this section, we present a detailed analysis of the optimal minibatch size derived in Sect. 5.2.

Lemma E.1 *We have that the iterates of Algorithm 2 satisfy Assumption 2 and Eq. (14) with*

$$\sigma_k^2 = \mathbb{E}\left[\left\| \|\nabla f_B(w_k) - \nabla f_B(x_*) - (\nabla f(w_k) - \nabla f(x_*))\|^2 \right\|^2\right],$$

and constants

$$A = 2\mathcal{L}(b),\ \ B = 2,\ \ \rho = p,\ \ C = p\mathcal{L}(b),\ \ D_1 = D_2 = 0,\ \ G = \mathcal{L}(b)L, \quad (43)$$

where $\mathcal{L}(b)$ is defined in (17).

Proof For constant A, B, ρ, C, D_1, D_2, see Lemma B.4 and Corollary B.4.

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f_{v_0}(x_0)-\nabla f_{v_0}(x_*)-(\nabla f(x_0)-\nabla f(x_*))\right\|^2\right] &\le \mathbb{E}\left[\left\|\nabla f_{v_0}(x_0)-\nabla f_{v_0}(x_*)\right\|^2\right]\\ &\overset{(25)}{\le} 2\mathcal{L}(b)D_f(x_0,x_*)\\ &\overset{(5)}{\le} \mathcal{L}(b)L\|x_0-x_*\|^2.\end{aligned}$$

where we used in the first inequality that $\mathbb{E}\left[\|X-\mathbb{E}[X]\|^2\right] = \mathbb{E}\left[\|X\|^2\right] - \|\mathbb{E}[X]\|^2 \le \mathbb{E}\left[\|X\|^2\right]$. Thus, (14) holds with $G=\mathcal{L}(b)L$. □

In the next corollary, we will give the iteration complexity for Algorithm 2 in the case where $p=1/n$, which is the usual choice for p in practice. A justification for this choice can be found in [23, 37].

Corollary E.1 (Iteration complexity of L-SVRG) *Consider the iterates of Algorithm 2. Let $p=1/n$ and $\gamma=\frac{1}{12\mathcal{L}(b)}$. Given the constants obtained for Algorithm 2 in (43), we have, using Corollary 5.1, that if*

$$k \ge \left(12\mathcal{L}(b)+\frac{nL}{6}\right)\frac{\|x_0-x_*\|^2}{\epsilon},$$

then, $\mathbb{E}\left[f(\bar{x}_k)-f(x_*)\right]\le\epsilon$.

The usual definition for the total complexity is the expected number of gradients computed per iteration, times the iteration complexity, required to reach an ϵ-approximate solution in expectation. However, since L-SVRG computes the full gradient every n iterations in expectation, we can say that L-SVRG computes roughly $2b+1$ gradients every iteration, so that after n iteration, it will have computed $n+2bn$ gradient. Thus, the total complexity for SVRG is:

$$\begin{aligned}K(b) &\overset{\text{def}}{=} (1+2b)\left(12\mathcal{L}(b)+\frac{nL}{6}\right)\frac{\|x_0-x_*\|^2}{\epsilon} &(44)\\ &= (1+2b)\left(\frac{12\left((n-b)L_{\max}+n(b-1)L\right)}{b(n-1)}+\frac{nL}{6}\right)\frac{\|x_0-x_*\|^2}{\epsilon}. &(45)\end{aligned}$$

Appendix E.3.1: Proof of Proposition 5.3

Proof Since the factor $\frac{\|x_0-x_*\|^2}{\epsilon}$ which appears in (44) does not depend on the minibatch size, minimizing the total complexity in the minibatch size corresponds to minimizing the following quantity:

$$Q(b)=(1+2b)\left(12\mathcal{L}(b)+\frac{nL}{6}\right).$$

We have

$$(n-1)Q(b) = 12(n-1)\mathcal{L}(b) + 24(n-1)b\mathcal{L}(b) + \frac{n(n-1)Lb}{3} + \frac{nL}{6}$$
$$= \frac{12n(L_{\max}-L)}{b} + \left(24(nL-L_{\max}) + \frac{n(n-1)L}{3}\right)b + \xi,$$

where ξ is a constant independent of b. Differentiating, we have:

$$(n-1)Q'(b) = -\frac{12n(L_{\max}-L)}{b^2} + 24(nL-L_{\max}) + \frac{n(n-1)L}{3}.$$

Since $L_{\max} \geq L$ and $nL \geq L_{\max}$ (see for example Lemma A.6 in [37]), $C(b)$ is a convex function of b. Thus, $Q(b)$ is minimized when $Q'(b) = 0$. Hence:

$$b^* = 6\sqrt{\frac{n\,(L_{\max}-L)}{72\,(nL-L_{\max}) + n(n-1)L}}.$$

Since $L_{\max}$ can take any value in the interval $[L, nL]$, we have $b^* \in [0, 6]$. □

Appendix E.4: Optimal Miniblock Size for *b*-SEGA (Algorithm 6)

In this section, we define for any $j \in [d]$ the matrix $I_j \in \mathbb{R}^{d\times d}$ such that

$$(I_j)_{pq} \overset{\text{def}}{=} \begin{cases} 1 \text{ if } p = q = j \\ 0 \text{ otherwise} \end{cases},$$

and we consequently define for any subset $B \subseteq [d]$,

$$I_B \overset{\text{def}}{=} \sum_{j\in B} I_j.$$

Algorithm 6 *b*-SEGA

```
Parameters step size $\gamma$, block size $b \in [d]$
Initialization $x_0 \in \mathbb{R}^d$, $h^0 = 0$
for $k = 1, 2, \dots$ do
    Sample a miniblock $B_k \subseteq [d]$ s.t. $|B_k| = d$
    $h^{k+1} = h^k + I_{B_k}\left(\nabla f(x_k) - h_k\right)$
    $g_k = \frac{d}{b} I_{B_k}\left(\nabla f(x_k) - h_k\right) + h_k$
    $x_{k+1} = \mathrm{prox}_{\gamma R}\left(x_k - \gamma g_k\right)$
end for
```

Corollary E.2 *From Lemma B.8, we have that the iterates of Algorithm 6 satisfy Assumption 2 and Eq.* (14) *with*

$$\sigma_k^2 = \|h_k - \nabla f(x_*)\|^2$$

and constants:

$$A = \frac{2dL}{b}, \; B = 2\left(\frac{d}{b} - 1\right), \; \rho = \frac{b}{d}, \; C = \frac{bL}{d}, \; D_1 = D_2 = 0, \; G = 0. \quad (46)$$

Proof For the constants A, B, ρ, C, D_1, D_2, see Lemma B.8. Moreover, in Algorithm 6, $h_0 = 0$. Thus, $\sigma_0^2 = \|h_0\|^2 = 0$. Thus, (14) holds with $G = 0$. □

In the next corollary, we will give the iteration complexity for Algorithm 6.

Corollary E.3 (Iteration complexity of b-SEGA) *Consider the iterates of Algorithm 6. Let* $\gamma = \frac{b}{4(3d-b)L}$. *Given the constants obtained for Algorithm 6 in* (46)*, we have, using Corollary 5.1, that if*

$$k \geq \frac{4(3d-b)L}{b} \frac{\|x_0 - x_*\|^2}{\epsilon},$$

then, $\mathbb{E}\left[F(\bar{x}_k) - F(x_*)\right] \leq \epsilon$.

Here, we define the total complexity as the number of coordinates of the gradient that we sample at each iteration times the iteration complexity. Since at each iteration, we sample b coordinates of the gradient, the total complexity for Algorithm 6 to reach an ϵ-approximate solution is

$$K(b) \overset{\text{def}}{=} 4\,(3d - b)\,L \frac{\|x_0 - x_*\|^2}{\epsilon}. \quad (47)$$

Thus, we immediately have the following proposition.

Proposition E.1 *Let* $b^* = \underset{b \in [d]}{\operatorname{argmin}} K(b)$, *where* $K(b)$ *is defined in* (47). *Then,*

$$b^* = d.$$

The consequence of this proposition is that when using Algorithm 6, one should always use as big a miniblock as possible if the cost of a single iteration is proportional to the miniblock size.

Appendix F: Auxiliary Lemmas

Appendix F.1: Smoothness and Convexity Lemma

We now develop an immediate consequence of each f_i being convex and smooth based on the follow lemma.

Lemma F.1 *Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex function*

$$g(z) - g(x) \leq \langle \nabla g(z), z - x \rangle, \quad \forall x, z \in \mathbb{R}^d, \tag{48}$$

and L_g-smooth

$$g(z) - g(x) \leq \langle \nabla g(x), z - x \rangle + \frac{L_g}{2} \|z - x\|_2^2, \quad \forall x, z \in \mathbb{R}^d. \tag{49}$$

It follows that

$$\|\nabla g(x) - \nabla g(z)\|^2 \leq L_g(g(x) - g(z) - \langle \nabla g(z), x - z \rangle), \quad \forall x \in \mathbb{R}^d. \tag{50}$$

Proof Fix $i \in \{1, \ldots, n\}$ and let

$$z = x - \frac{1}{L_g}(\nabla g(x) - \nabla g(x^*)).$$

To prove (50), it follows that

$$\begin{aligned} g(x^*) - g(x) &= g(x^*) - g(z) + g(z) - g(x) \\ &\overset{(48)+(49)}{\leq} \left\langle \nabla g(x^*), x^* - z \right\rangle + \langle \nabla g(x), z - x \rangle + \frac{L_g}{2} \|z - x\|_2^2 . \end{aligned} \tag{51}$$

Substituting this in z into (51) gives

$$\begin{aligned} g(x^*) - g(x) &= \left\langle \nabla g(x^*), x^* - x + \frac{1}{L_g}(\nabla g(x) - \nabla g(x^*)) \right\rangle \\ &\quad - \frac{1}{L_g} \left\langle \nabla g(x), \nabla g(x) - \nabla g(x^*) \right\rangle + \frac{1}{2L_g} \left\| \nabla g(x) - \nabla g(x^*) \right\|_2^2 \\ &= \left\langle \nabla g(x^*), x^* - x \right\rangle \\ &\quad - \frac{1}{L_g} \left\| \nabla g(x) - \nabla g(x^*) \right\|_2^2 + \frac{1}{2L_g} \left\| \nabla g(x) - \nabla g(x^*) \right\|_2^2 \\ &= \left\langle \nabla g(x^*), x^* - x \right\rangle - \frac{1}{2L_g} \left\| \nabla g(x) - \nabla g(x^*) \right\|_2^2 . \end{aligned}$$

□

Lemma F.2 *Suppose that for all $i \in [n]$, f_i is convex and L_i-smooth, and let $L_{\max} = \max_{i \in [n]} L_i$. Then*

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L_{\max} \left(f(x) - f(x_*) \right). \tag{52}$$

Proof From (50), we have for all $i \in [n]$,

$$\|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L_i \left(f_i(x) - f_i(x_*) - \langle \nabla f_i(x_*), x - x_*\rangle\right)$$

Thus,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 &\leq 2L_{\max} \left(f(x) - f(x_*) - \langle \nabla f(x_*), x - x_*\rangle\right) \\ &= 2L_{\max} \left(f(x) - f(x_*)\right).\end{aligned}$$

□

Appendix F.2: Proximal Lemma

Lemma F.3 *Let $R : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex lower semi-continuous function. For $z, y \in \mathbb{R}^d$ and $\gamma > 0$. With $p = \operatorname{prox}_{\gamma g}(y)$ we have that for*

$$g(p) - g(z) \leq -\frac{1}{\gamma}\langle p - y, p - z\rangle. \tag{53}$$

Proof This is classic result, see, for example, the “Second Prox Theorem” in Section 6.5 in [6]. □

Appendix F.3: Proof of Lemma C.2

This proof follows the proof of Lemma 8 in [5], and we reproduce it for completeness. Indeed, using the convexity of f

$$f(x) - f(x_*) \geq -\langle \nabla f(x), x_* - x\rangle$$

in combination with (53) where $z = x_*$ gives

$$f(x) + g(p) - F(x_*) \leq -\frac{1}{\gamma}\langle p - y, p - x_*\rangle - \langle \nabla f(x), x_* - x\rangle.$$

Now using smoothness

$$f(p) - f(x) \leq \langle \nabla f(x), p - x\rangle + \frac{1}{2\gamma}\|p - x\|^2,$$

gives

$$\begin{aligned} F(p) - F(x_*) &\leq -\frac{1}{\gamma}\left\langle p - y, p - x_* \right\rangle - \left\langle \nabla f(x), x_* - x \right\rangle + \left\langle \nabla f(x), p - x \right\rangle \\ &\quad + \frac{1}{2\gamma}\left\| p - x \right\|^2 \\ &= -\frac{1}{\gamma}\left\langle p - y, p - x_* \right\rangle + \left\langle \nabla f(x), p - x_* \right\rangle + \frac{1}{2\gamma}\left\| p - x \right\|^2 \\ &= -\frac{1}{\gamma}\left\langle p - \gamma\nabla f(x) - y, p - x_* \right\rangle + \frac{1}{2\gamma}\left\| p - x \right\|^2 \\ &= -\frac{1}{\gamma}\left\langle p - x + x - \gamma\nabla f(x) - y, p - x_* \right\rangle + \frac{1}{2\gamma}\left\| p - x \right\|^2 \\ &= -\frac{1}{\gamma}\left\langle p - x, p - x_* \right\rangle - \frac{1}{\gamma}\left\langle x - \gamma\nabla f(x) - y, p - x_* \right\rangle + \frac{1}{2\gamma}\left\| p - x \right\|^2 . \end{aligned} \tag{54}$$

Using that

$$-2\left\langle p - x, p - x_* \right\rangle + \left\| p - x \right\|^2 = -\left\| p - x_* \right\|^2 + \left\| z - x \right\|^2 ,$$

in combination with (54) gives

$$F(p) - F(x_*) \leq -\frac{1}{2\gamma}\| p - x_* \|^2 - \frac{1}{\gamma}\left\langle x - \gamma\nabla f(x) - y, p - x_* \right\rangle + \frac{1}{2\gamma}\| x_* - x \|^2.$$

Now it remains to multiply both sides by -2γ to arrive at (33).

References

1. Alistarh, D., Grubic, D., Li, J., Tomioka, R., Vojnovic, M.: QSGD: communication-efficient SGD via gradient quantization and encoding. Adv. Neural Inf. Process. Syst. **30**, 1709–1720 (2017)
2. Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., Renggli, C.: The convergence of sparsified gradient methods. Adv. Neural Inf. Process. Syst. **31**, 5977–5987 (2018)
3. Allen-Zhu, Z.: Katyusha: the first direct acceleration of stochastic gradient methods. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC, pp. 1200–1205 (2017)
4. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. In: Proceedings of the 33nd International Conference on Machine Learning (2016)
5. Atchadé, Y.F., Fort, G., Moulines, E.: On perturbed proximal gradient algorithms. J. Mach. Learn. Res. **18**(1), 310–342 (2017)
6. Beck, A.: First-Order Methods in Optimization. MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia (2017)
7. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
8. Defazio, A., Bach, F.R., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. Adv. Neural Inf. Process. Syst. **27**, 1646–1654 (2014)
9. Gazagnadou, N., Gower, R.M., Salmon, J.: Optimal mini-batch and step sizes for SAGA. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 2142–2150 (2019)
10. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**(4), 2341–2368 (2013)

11. Gorbunov, E., Hanzely, F., Richtárik, P.: A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent. In: AISTATS (2020)
12. Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., Richtárik, P.: SGD: general analysis and improved rates. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 5200–5209 (2019)
13. Gower, R.M., Richtárik, P., Bach, F.: Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. Math. Program. **188**, 135–192 (2020)
14. Grimmer, B.: Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. SIAM J. Optim. **29**(2), 1350–1365 (2019)
15. Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P.: Deep learning with limited numerical precision. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 1737–1746 (2015)
16. Hanzely, F., Mishchenko, K., Richtárik, P.: SEGA: variance reduction via gradient sketching. Adv. Neural Inf. Process. Syst. **31**, 2086–2097 (2018)
17. Hofmann, T., Lucchi, A., Lacoste-Julien, S., McWilliams, B.: Variance reduced stochastic gradient descent with neighbors. Adv. Neural Inf. Process. Syst. **28**, 2305–2313 (2015)
18. Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., Richtárik, P.: Stochastic distributed learning with gradient quantization and variance reduction. arXiv:1904.05115 (2019)
19. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. **26**, 315–323 (2013)
20. Khaled, A., Richtárik, P.: Better theory for SGD in the nonconvex world. arXiv:2002.03329 (2020)
21. Konečný, J., Liu, J., Richtárik, P., Takác, M.: Mini-batch semi-stochastic gradient descent in the proximal setting. J. Sel. Top. Signal Process. **10**(2), 242–255 (2016)
22. Konečný, J., Richtárik, P.: Randomized distributed mean estimation: accuracy vs communication. arXiv:1611.07555 (2016)
23. Kovalev, D., Horváth, S., Richtárik, P.: Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In: Proceedings of the 31st International Conference on Algorithmic Learning Theory, vol. 117, pp. 451–467 (2020)
24. Lei, Y., Ting, H., Li, G., Tang, K.: Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. IEEE Trans. Neural Netw. Learn. Syst. **31**(10), 4394–4400 (2020)
25. Loizou, N., Vaswani, S., Laradji, I.H., Lacoste-Julien, S.: Stochastic polyak step-size for SGD: an adaptive learning rate for fast convergence. In: Banerjee, A., Fukumizu, K. (eds.) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pp. 1306–1314. PMLR (2021)
26. Mishchenko, K., Gorbunov, E., Takáč, M., Richtárik, P.: Distributed learning with compressed gradient differences. arXiv:1901.09269 (2019)
27. Mishchenko, K., Hanzely, F., Richtárik, P.: 99% of worker-master communication in distributed optimization is not needed. In: Adams, R.P., Gogate, V. (eds.) Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3–6, 2020, volume 124 of Proceedings of Machine Learning Research, pp. 979–988. AUAI Press (2020)
28. Needell, D., Srebro, N., Ward, R.: Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. Math. Program. Ser. A **155**(1), 549–573 (2016)
29. Nemirovski, A., Juditsky, A.B., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
30. Nesterov, Y.E.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
31. Nguyen, L., Nguyen, P.H., van Dijk, M., Richtárik, P., Scheinberg, K., Takáč, M.: SGD and Hogwild! Convergence without the bounded gradients assumption. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 3750–3758 (2018)
32. Nguyen, L.M., Liu, J., Scheinberg, K., Takáč, M.: SARAH: a novel method for machine learning problems using stochastic recursive gradient. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 2613–2621 (2017)
33. Ravikumar, P., Wainwright, M.J., Lafferty, J.D.: High-dimensional Ising model selection using ℓ_1-regularized logistic regression. Ann. Stat. **38**(3), 1287–1319 (2010)
34. Reddi, S.J., Hefny, A., Sra, S., Póczos, B., Smola, A.J.: Stochastic variance reduction for nonconvex optimization. In: Proceedings of the 33nd International Conference on Machine Learning, vol. 48, pp. 314–323 (2016)

35. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
36. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)
37. Sebbouh, O., Gazagnadou, N., Jelassi, S., Bach, F., Gower, R.M.: Towards closing the gap between the theory and practice of SVRG. Adv. Neural Inf. Process. Syst. **32**, 646–656 (2019)
38. Seide, F., Fu, H., Droppo, J., Li, G., Yu, D.: 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNS. In: INTERSPEECH, 15th Annual Conference of the International Speech Communication Association, pp. 1058–1062 (2014)
39. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
40. Stich, S.U.: Unified optimal analysis of the (stochastic) gradient method. arXiv:1907.04232 (2019)
41. Tibshirani, R.J.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
42. Vaswani, S., Bach, F., Schmidt, M.: Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In: The 22nd International Conference on Artificial Intelligence and Statistics, vol. 89, pp. 1195–1204 (2019)
43. Wangni, J., Wang, J., Liu, J., Zhang, T.: Gradient sparsification for communication-efficient distributed optimization. Adv. Neural Inf. Process. Syst. **31**, 1306–1316 (2018)
44. Wright, S.J.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)
45. Zhang, H., Li, J., Kara, K., Alistarh, D., Liu, J., Zhang, C.: Zipml: training linear models with end-to-end low precision, and a little bit of deep learning. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 4035–4043 (2017)
46. Zhao, P., Zhang, T.: Stochastic optimization with importance sampling for regularized loss minimization. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 1–9 (2015)
47. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(2), 301–320 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.