Math. Program., Ser. B (2019) 174:253–292
https://doi.org/10.1007/s10107-018-1297-x

FULL LENGTH PAPER

On variance reduction for stochastic smooth convex optimization with multiplicative noise

Alejandro Jofré[1] · **Philip Thompson**[2]

Received: 24 May 2017 / Accepted: 11 May 2018 / Published online: 5 June 2018

Abstract We propose dynamic sampled stochastic approximation (SA) methods for stochastic optimization with a heavy-tailed distribution (with finite 2nd moment). The objective is the sum of a smooth convex function with a convex regularizer. Typically, it is assumed an oracle with an upper bound σ^2 on its variance (OUBV). Differently, we assume an oracle with *multiplicative noise*. This rarely addressed setup is more aggressive but realistic, where the variance may not be uniformly bounded. Our methods achieve optimal iteration complexity and (near) optimal oracle complexity. For the smooth convex class, we use an accelerated SA method a la FISTA which achieves, given tolerance $\varepsilon > 0$, the optimal iteration complexity of $\mathcal{O}(\varepsilon^{-\frac{1}{2}})$ with a near-optimal oracle complexity of $\mathcal{O}(\varepsilon^{-2})[\ln(\varepsilon^{-\frac{1}{2}})]^2$. This improves upon Ghadimi and Lan (Math Program 156:59–99, 2016) where it is assumed an OUBV. For the strongly convex class, our method achieves optimal iteration complexity of $\mathcal{O}(\ln(\varepsilon^{-1}))$ and optimal oracle complexity of $\mathcal{O}(\varepsilon^{-1})$. This improves upon Byrd et al. (Math Program 134:127–155, 2012) where it is assumed an OUBV. In terms of variance, our bounds are local: they depend on variances $\sigma(x^*)^2$ at solutions x^* and the per unit distance multiplicative variance σ_L^2. For the smooth convex class, there exist policies such that our bounds resemble, up to absolute constants, those obtained in the mentioned papers if it was assumed an OUBV with $\sigma^2 := \sigma(x^*)^2$. For the strongly convex class such property is obtained exactly if the condition number is estimated or in the limit for better conditioned problems or for larger initial batch sizes. In

✉ Philip Thompson
philipthomp@gmail.com; Philip.THOMPSON@ensae.fr

Alejandro Jofré
ajofre@dim.uchile.cl

1 Center for Mathematical Modeling (CMM) and DIM, Universidad de Chile, Santiago, Chile

2 Center for Mathematical Modeling (CMM), Universidad de Chile, Santiago, Chile

any case, if it is assumed an OUBV, our bounds are thus sharper since typically $\max\{\sigma(x^*)^2, \sigma_L^2\} \ll \sigma^2$.

Keywords Stochastic approximation · Smooth convex optimization · Composite optimization · Multiplicative noise · Acceleration · Dynamic sampling · Variance reduction · Complexity

Mathematics Subject Classification 65K05 · 62L20 · 90C25 · 90C15 · 68Q25

1 Introduction

We consider methods for convex optimization problems where only noisy first-order information is assumed. This setting includes problems in signal processing and empirical risk minimization for machine learning [5,8,57], stochastic optimization and finance [42,56] and simulation optimization [18]. In such problems, we have a closed convex set $X \subset \mathbb{R}^d$, a distribution $\mathbf{P}$ over a sample space Ξ and a measurable function $F : X \times \Xi \to \mathbb{R}$ satisfying

$$f(x) := \mathbb{E}F(x,\xi) = \int_{\Xi} F(x,\xi)\,\mathrm{d}\,\mathbf{P}(\xi), \quad (x \in X), \tag{1}$$

where for almost every (a.e.) $\xi \in \Xi$, $F(\cdot,\xi)$ is a continuously differentiable convex function for which $F(x,\cdot)$ and $\nabla F(x,\cdot)$ are integrable. The stochastic optimization problem is then to solve

$$\min_{x\in X} f(x). \tag{2}$$

The challenge aspect of stochastic optimization, when compared to deterministic optimization, is that the expectation (1) cannot be evaluated.[1] However, a practical assumption is that the decision maker have access to samples drawn from the distribution $\mathbf{P}$.

Two different methodologies exist for solving (1)–(2) when samples $\{\xi_j\}_{j=1}^N$ of the distribution $\mathbf{P}$ is available. The *sample average approximation* (SAA) methodology is to solve the problem

$$\min_{x\in X}\left\{\widehat{F}_N(x) := \frac{1}{N}\sum_{j=1}^{N} F(\xi_j, x)\right\}, \tag{3}$$

by resorting to a chosen algorithm. See for instance [42,56] for such kind of approach in stochastic optimization based on Monte Carlo simulation. Such methodology is also the case of empirical risk minimization (ERM) in statistical machine learning where $\mathbf{P}$

[1] Typical reasons are: a sample space with high dimension requiring Monte Carlo evaluation, no knowledge of the distribution $\mathbf{P}$ or, even worse, no knowledge of a close form for F.

is unknown and a limited number of samples is acquired by measurements. Note that (3) itself is of the form (1)–(2) with the empirical distribution $\widehat{\mathbf{P}}_N := \frac{1}{N}\sum_{j=1}^{N}\delta_{\xi_j}$, where δ_ξ denotes the Dirac measure at the point $\xi \in \Xi$.

A different methodology is the *stochastic approximation* (SA) approach where the samples are accessed in an iterative and online fashion: a deterministic version of an algorithm is chosen and samples are used whenever the algorithm requires gradients at the current or previous iterates [36,42]. In this setting the mechanism to access F via samples of $\mathbf{P}$ is usually named a *stochastic oracle* (SO). Precisely, given an input $x \in X$ and an independent identically distributed (i.i.d.) sample $\{\xi_j\}$ drawn from $\mathbf{P}$ (also independent of the input x), the SO outputs unbiased gradients $\{\nabla F(x, \xi_j)\}$, that is, satisfying $\mathbb{E}[\nabla F(x, \xi_j)] = \nabla f(x)$ for all j.

The SA methodology was first proposed by Robbins and Monro in [52] for problem (1)–(2) when f is a smooth strongly convex function under specific conditions. In the unconstrained case, the method takes the form

$$x^{t+1} := x^t - \alpha_t \nabla F(x^t, \xi^t), \tag{4}$$

where α_t is a positive stepsize and ξ^t is a sample from $\mathbf{P}$. The above method is one SA version of the gradient descent method, known as stochastic gradient method (SG). This methodology was then extensively explored in numerous works spanning the communities of statistics and stochastic approximation, stochastic optimization and machine learning. We refer, e.g., to [5,10,13,14,20,21,42,57] for further references. More recently, the SA methodology was also analyzed for stochastic variational inequalities (VI). See e.g., [6,26–28,31,34] and references therein. VI is a framework which generalizes unconstrained system of equations and the first order necessary condition of constrained minimization (including a broader class of problems such as equilibrium, saddle-point and complementarity problems).

In this work we consider SA methods for solving constrained and regularized stochastic *smooth convex* optimization problems (CO) of the form

$$g^* := \min_{x\in X} \{g(x) := f(x) + \varphi(x)\}, \tag{5}$$

where $X \subset \mathbb{R}^d$ is closed and convex, $f : X \to \mathbb{R}$ is a smooth convex function satisfying (1) and $\varphi : X \to \mathbb{R}^d$ is a (possibly nonsmooth) convex function. By smoothness we mean f is a differentiable function satisfying, for some $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in X, \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm. The function φ is suppose to be a known simple function so that proximal evaluations are easy to compute (see Sect. 2). The above set-up includes many problems in stochastic optimization, simulation optimization and, in particular, inference problems in machine learning and signal processing [7,8,42]. The function φ is often used in applications to induce regularization and parsimony (such as sparsity). Examples include $\varphi := \lambda\|\cdot\|^2$ (as in *ridge regression*), $\varphi := \lambda\|\cdot\|_1$ (as in the Lasso estimator) or $\varphi := \lambda\|\cdot\|^2 + \gamma\|\cdot\|_1$ (as in *elastic net*) for some $\lambda, \gamma > 0$

where $\|\cdot\|_1$ denotes the the ℓ_1-norm (see, e.g., [60]). We will denote the solution set of (5) by X^*.

We will also consider the special subclass where f is *smooth c-strongly convex*, i.e., satisfying (6) and for some $c > 0$,

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{c}{2}\|y - x\|^2 \leq f(y), \quad \forall x, y \in X. \tag{7}$$

In this case, problem (5) has an unique solution. In order to present the main ideas with respect to variance reduction, we refrain from considering non-Euclidean geometries which will be treated in future work. As usual, we make the following assumption on the sampling resource.

Assumption 1 (i.i.d. sampling) In all our methods, the samples drawn from the distribution **P** and used along the chosen algorithm are i.i.d.

1.1 Oracle assumptions

In stochastic optimization, the assumption on the *stochastic oracle's variance* is as important as the class of smoothness of the objective since both have consequences in obtaining surrogate models and on condition numbers of the problem. Define, for every $x \in X$, the pointwise oracle's standard deviation by $\sigma(x) := \|\nabla F(x, \xi) - \nabla f(x)\|_2$, where $|q(\xi)|_2 := \sqrt{\mathbb{E}[q(\xi)^2]}$ denotes the L^2-norm of the random variable $q : \Xi \to \mathbb{R}$.

A reasonable hypothesis, used since the seminal work of Robbins and Monro [52], is to assume a SO with an *uniformly bounded variance* over the feasible set X, i.e., that there exists $\sigma > 0$, such that

$$\sup_{x \in X} \mathbb{E}\left[\|\nabla F(x, \xi) - \nabla f(x)\|^2\right] \leq \sigma^2. \tag{8}$$

Condition (8) is a *global variance* assumption on the noise when using stochastic approximation. Assumption (8) is valid in important instances, e.g., when the feasible set X is compact or when an *uniform additive noise* is assumed, that is, for some centered random variable $\varepsilon \in \mathbb{R}^d$ with $\mathbb{E}[\|\varepsilon\|^2] \leq \sigma^2$, for a.e. $\xi \in \Xi$,

$$\nabla F(\xi, x) = \nabla f(x) + \varepsilon(\xi), \quad \forall x \in X. \tag{9}$$

Property (9) is a reasonable assumption in many important ERM problems, such as instances of the least squares regression problem (LSR). Note that property (9), although more structured, allows unconstrained optimization ($X = \mathbb{R}^d$).

Property (9) is not a reasonable assumption in problems where the noise is dependent on the point of the feasible set. In that case, property (8) is an important generalization of (9) assumed in most stochastic optimization methods. However, it implicitly assumes compactness of X. This has two drawbacks. The first is that it rules out unconstrained minimization problems where (9) is not satisfied. This includes many stochastic and simulation optimization problems as well as LSR without additive noise,

a more aggressive but relevant condition in ERM. The second is that, even if (8) does hold, σ^2 may be very large. The reason is that, in case of multiplicative noise, $\sigma(\cdot)^2$ *is typically coercive over* X and, hence, σ^2 grows with the diameter of X (see Example 1 in the following).

In this work we will consider the following assumption.

Assumption 2 *(Oracle with multiplicative noise)* There exist $x^0 \in X$ such that $\sigma(x^0) < \infty$ and $\sigma_L > 0$ such that[2]

$$\sigma(x) \leq \sigma(x^*) + \sigma_L \|x - x^*\|, \quad \forall x, x^* \in X. \tag{10}$$

The number σ_L bounds the per unit distance multiplicative spread of the standard deviation of the oracle relative to the reference point $x^* \in X$. Precisely, $\sigma(x) - \sigma(x^*) \leq \sigma_L$, if $\|x - x^*\| \leq 1$. Condition (10) is much weaker than (8) since it allows the more aggressive setting where

$$\sup_{x \in X} \mathbb{E}\left[\|\nabla F(x, \xi) - \nabla f(x)\|^2\right] = \infty, \tag{11}$$

and, in case X is compact, it implies (8) with $\sigma^2 := 2\sigma(x^*)^2 + 2\sigma_L^2 \mathscr{D}(X)^2$, where $\mathscr{D}(X)$ is the diameter of X. However, we note the quadratic dependence on $\mathscr{D}(X)$ and that $\sigma^2 \gg \min\{\sigma^2(x^*), \sigma_L^2\}$ if $\sqrt{\mathscr{D}(X)}$ is large. In this sense, condition (10) exploits *local variance* behavior of the noise when using stochastic approximation. One of our main objectives in this work is to consider the more general condition (10) in stochastic approximation algorithms for smooth convex optimization. We refer the reader to [4], where *local strong convexity* is exploited in order to ensure improved error bounds of order $\mathscr{O}(t^{-1})$ in terms of better constants.[3]

The terminology "multiplicative noise" and generality of (10) are explained in the following lemma whose proof is in the "Appendix".

Lemma 1 *Suppose the random variable* $\mathsf{L} : \Xi \to \mathbb{R}_+$ *has finite variance and for a.e.* $\xi \in \Xi$,

$$\|\nabla F(x, \xi) - \nabla F(y, \xi)\| \leq \mathsf{L}(\xi)\|x - y\|, \quad x, y \in X. \tag{12}$$

Then (6) *holds with* $L := |\mathsf{L}(\xi)|_2$ *and Assumption 2 holds with* $\sigma_L = 2L$.

Hence, condition (12) defines the smoothness of f *and* the variance of the oracle as in Assumption 2. We note that (12) is a standard assumption in stochastic optimization [56]. In the sense of (12), the random variable L is indeed a multiplicative noise on the first order SO. Under the assumptions of Lemma 1, Assumption 2 is merely a *finite*

[2] The convergence theory we present would work if (10) is satisfied for just one $x^* \in X$. However, more uniform bounds are obtained under Assumption 2 which is satisfied when (12) holds.

[3] The local strong convexity modulus around the unique solution x^* is potentially much higher than the global strong convexity modulus c. We remark that, although our presented methods adapt to local variances, they still depend on the global strong convexity constant (in case the objective is strongly convex).

second moment assumption of $\|\nabla F(\xi, x^*) - \nabla f(x^*)\|$ for some $x^* \in X$ and $\mathsf{L}(\xi)$. We next show a typical example.

Example 1 (Stochastic quadratic optimization with a random matrix) Consider the stochastic *quadratic optimization* problem

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{2} \langle x, \mathbb{E}\left[A(\xi)\right] x\rangle + \langle \mathbb{E}\left[b(\xi)\right], x\rangle \right\},$$

where $A : \Xi \to \mathbb{R}^{d\times d}$ is a semi-positive definite *random matrix* with nonnull mean and finite second moment and $b : \Xi \to \mathbb{R}^d$ is an integrable random vector. This is the case, e.g., of Least-Squares Regression [12,15]. After some straightforward calculation, we can derive the expression $\sigma(x)^2 = \langle x, Bx\rangle$, for any $x \in \mathbb{R}^d$, where $B := \sum_{i=1}^d \mathrm{cov}[A_i(\xi)]$, the vectors $A_1(\xi), \ldots, A_d(\xi)$ are the rows of $A(\xi)$ and $\mathrm{cov}[q(\xi)]$ defines the covariance matrix of the random vector $q : \Xi \to \mathbb{R}^d$. Let $N(B)$ denote the kernel of B and $N(B)^\perp$ denote its orthogonal complement. Given $x \in \mathbb{R}^d$, we denote by x_B the orthogonal projection of x onto $N(B)^\perp$. Since B is semi-positive definite, from the spectral theorem we get

$$\sigma(x)^2 \geq \lambda_+(B)\|x_B\|^2, \quad \forall x \in \mathbb{R}^d,$$

where $\lambda_+(B)$ is the smallest nonnull eigenvalue of B. This shows $\sigma(\cdot)^2$ *is quadratically coercive on the linear subspace* $N(B)^\perp$. In particular, if B is positive definite, $\sigma(x)^2 \geq \lambda_+(B)\|x\|^2$, for all $x \in \mathbb{R}^d$. We thus conclude that (8) is not valid if $X \cap N(B)^\perp$ is unbounded. In particular, it is not true in the unconstrained case ($X = \mathbb{R}^d$). Nevertheless, (12) holds with $\mathsf{L}(\xi) := \sup\{\|A(\xi)x\| : \|x\| = 1\}$.

1.2 Related work and contributions

The performance of a SA method can be measured by its *iteration* and *oracle complexities* given a tolerance $\varepsilon > 0$ on the mean optimality gap. The first is the total number of iterations, a measure for the optimization error, while the second is the total number of samples and oracle calls, a measure for the estimation error. Statistical and optimization lower bounds [2,44,46] show that, *in terms of the tolerance* ε, the optimal oracle complexities are $\mathcal{O}(\varepsilon^{-2})$ for the smooth convex class and $\mathcal{O}(\varepsilon^{-1})$ for the smooth strongly convex class.[4] Anyhow, an important question that remains is (Q): *What is the optimal iteration complexity such that a (near) optimal oracle complexity is respected*? Related to such question is the ability of method to treat the *oracle's assumptions and variance* with *sampling efficiency*. We review some literature related to these observations.

[4] We emphasize that the lower bounds in [2,44] assume the feasible set is compact, the objective is Lipschitz continuous, the oracle's variance is uniformly bounded and the oracle can only be called once per iteration. This includes the smooth class if we assume that the compact feasible set is contained in the domain of the objective and we pay attention to the tolerance level ε.

(a) *Polyak–Ruppert's iterate averaging in stochastic approximation* the initial version of the SG method uses one oracle call per iteration with stepsizes satisfying $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$, typically $\alpha_t = \mathcal{O}(t^{-1})$. A fundamental improvement with respect to the estimation error was Polyak–Ruppert's *iterate averaging* scheme [43,50,51,54], where *longer stepsizes* $\alpha_t = \mathcal{O}(t^{-\frac{1}{2}})$ are used with a subsequent *final average* of the iterates with weights $\{\alpha_t\}$ (this is sometimes called *ergodic average*). *If one oracle call per iteration is assumed* (so that the iteration complexity is equal to the oracle complexity) such scheme obtains optimal iteration and oracle complexities of $\mathcal{O}(\varepsilon^{-2})$ for the smooth convex class and of $\mathcal{O}(\varepsilon^{-1})$ for smooth strongly convex class [2,44]. These are also the size of the final ergodic average, a measure of the additional *averaging effort* implicitly required in iterate averaging schemes. Such methods, hence, are efficient in terms of oracle complexity. Iterate averaging was then extensively explored (see e.g., [33,35,38,42,48,49,59]). The important work [42] exploits the robustness of iterate averaging in SA methods and shows that such schemes can outperform the SAA approach on relevant convex problems. On the strongly convex class, [5] gives a detailed non-asymptotic robust analysis of Polyak–Ruppert averaging scheme. See also [24,41,53] for improvements.

(b) *Nesterov's acceleration in composite and stochastic problems* in the seminal work [45] of Nesterov, a novel accelerated scheme for deterministic unconstrained smooth convex optimization with an exact oracle is presented obtaining the optimal rate $\mathcal{O}(Lt^{-2})$. This improves upon $\mathcal{O}(Lt^{-1})$ of standard methods. Motivated by the importance of regularized problems, this result was further generalized for composite convex optimization in [7,47,58] and in [36], where the stochastic oracle was considered. Assuming (8) and one oracle call per iteration, [36] obtains iteration and oracle complexities of $\mathcal{O}\left(\sqrt{L\varepsilon^{-1}} + \sigma^2\varepsilon^{-2}\right)$, allowing larger values of L. See also [25,59] for the stochastic case. In [20,21], strongly convex problems with a stochastic oracle were considered and iteration and oracle complexities of $\mathcal{O}\left(\sqrt{Lc^{-1}}\ln(\varepsilon^{-1}) + \sigma^2 c^{-1}\varepsilon^{-1}\right)$ were obtained, allowing for larger values of the condition number L/c. See [38,40,59] for considerations on sparse solutions.

(c) *Variance reduction in stochastic approximation* we observe that a vanishing stepsize policy with iterate averaging is itself a variance reduction procedure in the sense that the stepsize has to be small enough to reduce the variance and large enough to guarantee convergence. Schemes with vanishing stepsizes typically require just one oracle call per iteration. However, even in the deterministic setting, it is known that a vanishing stepsize policy entails a slower rate when compared to a constant stepsize policy. Motivated by question (Q), a rapidly growing line of research is the development of methods which use *more than one oracle call per iteration* (a viable assumption in many problem instances). Current examples are the *gradient aggregation* and *dynamic sampling* methods (see [8, Section 5]). Designed for *finitely supported distributions* of the form (3) (as found in the ERM problem of machine learning), gradient aggregation methods reduce the variance by combining in a specific manner eventual exact computation (or storage) of gradients and eventual iterate averaging (or randomization schemes) with frequent gradient sampling. Their complexity bounds typically grow with the sample size N in (3). See e.g., [1,11,23,32,37,39,55,60,61] and references therein. Designed for solving problem (1)–(2) with an online data acquisition (as is the

case in many stochastic and simulation optimization problems based on Monte Carlo methods), dynamic sampling methods reduce variance by estimating the gradient via an *empirical average* associated to a sample whose size (*mini-batch*) is increased at every iteration. Their complexity bounds typically grow with the oracle variance σ^2 in (8). See [9,16,22,27,28,62] and references therein. An essential point related to (Q) is if such increased effort in computation per iteration used is worthwhile. A nice fact is that current gradient aggregation and dynamic sampling methods achieve the order of the deterministic optimal iteration complexity with the same (near) optimal oracle complexity and averaging effort of standard iterate averaging schemes. *In this sense*[5] and assuming that it is possible to call the oracle more than once per iteration, gradient aggregation and dynamic sampling methods can be a more efficient option than the classical iterate averaging scheme.

The contributions of this work may be summarized as follows:

We show the standard global assumption (8) *used to obtain non-asymptotic convergence can be replaced by the significantly weaker local assumption* (10), *provided dynamic sampling is possible. The bounds derived under* (10) *and dynamic sampling depend on the local variances* $\sigma(x^*)^2$ *and* σ_L^2, *for* $x^* \in X^*$. *Moreover, such bounds can be tuned so to resemble, up to absolute numerical constants, previous bounds obtained in* [9,22] *if it was supposed that* (8) *holds but replacing* σ^2 *with* $\sigma(x^*)^2$. *In particular, if* (8) *does hold then our bounds are, up to absolute constants, sharper than the ones in* [9,22] *since typically* $\sigma(x^*)^2 \ll \sigma^2$. *These type of results can be seen as a* variance localization property.

We next state our contributions more precisely.

(i) *Stochastic smooth non-strongly convex optimization* for this class of problems, we propose Algorithm 1 stated in Sect. 3 which is a dynamic sampled SA version of the accelerated method FISTA [7]. We show Algorithm 1 achieves the optimal iteration complexity of $\mathcal{O}(\varepsilon^{-\frac{1}{2}})$ with a near-optimal oracle complexity and average effort of $\mathcal{O}(\varepsilon^{-2})[\ln(\varepsilon^{-\frac{1}{2}})]^2$ under the more general Assumption 2 of *multiplicative noise*. These are online bounds.[6] The factor $[\ln(\varepsilon^{-\frac{1}{2}})]^2$ can be removed for offline bounds. Precisely, if $\sigma_L = \mathcal{O}(L)$ (see Lemma 1), Algorithm 1 can be tuned so to have a convergence rate of the form

$$\mathbb{E}\left[g(z^{t+1}) - g^*\right] \leq \mathcal{O}\left(t^{-2}\right)\left(L\|z^0 - x^*\|^2 + \frac{\sigma(x^*)^2}{L}\right),$$

where z^0 is the initial iterate and $x^* \in X^*$. This is achieved with a sampling rate $N_t = \mathcal{O}(1)t^3(\ln t)^2$ and it guarantees the oracle complexity $\sum_{\tau=1}^{t} N_\tau \lesssim \mathcal{O}(\varepsilon^{-2})[\ln(\varepsilon^{-\frac{1}{2}})]^2$ after t iterations. This improves upon [22], where accelerated dynamic sampling

[5] Here "optimality" refers to the order in $\mathcal{O}(\varepsilon^{-1})$. The precise complexity comparison, considering constants appearing in *statistical minimax bounds* which depend on *endogenous parameters*, is still a further question of research.

[6] That is, without an a priori known number of iterations.

schemes obtain, up to absolute constants, a similar complexity but with the more restrictive assumption (8) of an oracle with *uniformly bounded variance*. Hence, we do not implicitly require an additive noise model nor boundedness of X. Our rates depend on the local variances $\sigma(x^*)^2$ for $x^* \in X^*$ and σ_L^2. Interestingly, the stepsize and sampling rate policies can be tuned so that our bounds resemble, up to absolute constants, those obtained in [22] if *it was supposed* (8) *holds* but with σ^2 replaced by $\sigma(x^*)^2$ for some $x^* \in X^*$. Hence, in case (8) indeed holds, our bounds are sharper than the ones in [22] since typically $\sigma(x^*)^2 \ll \sigma^2$. See Example 1 in Sect. 1.1 and Theorem 1, Corollary 1 and Remark 2 in Sect. 3. Additionally, Algorithm 1 is sparsity preserving since it is based on FISTA. Differently, the methods in [22] are not sparsity preserving since they are based on the AC-SA method of [36] which uses iterate averaging (see [38,40,59] for further observations).

(ii) *Stochastic smooth strongly convex optimization* for this class of problems we propose Algorithm 2 given in Sect. 4 which is a dynamic sampled version of the stochastic proximal gradient method. We show Algorithm 2 achieves the optimal iteration complexity of $\mathscr{O}(\ln(\varepsilon^{-1}))$ with an optimal oracle complexity and average effort of $\mathscr{O}(\varepsilon^{-1})$ under the more general Assumption 2. Precisely, if $\sigma_L = \mathscr{O}(L)$ (see Lemma 1), Algorithm 2 can be tuned so to have a convergence rate of the form

$$\mathbb{E}\left[g(x^{t+1}) - g^*\right] \leq \mathscr{O}\left(\rho^{t-1}\right)\left(L\|x^1 - x^*\|^2 + \frac{\sigma(x^*)^2}{c}\right),$$

where $\rho = 1 - \mathscr{O}\left(\frac{1}{\kappa}\right) \in (0,1)$, $\kappa := \frac{L}{c}$ is the condition number, x^1 is the initial iterate and x^* is the unique solution. This is achieved with a sampling rate $N_t = \mathscr{O}(\kappa)\rho^{-t}$ and it guarantees the oracle complexity $\sum_{\tau=1}^{t} N_\tau \lesssim \mathscr{O}(\varepsilon^{-1})$ after t iterations. This improves upon [9] where, up to absolute constants, a similar complexity is obtained for dynamic sampling schemes but with the more restrictive assumption (8). Also, no regularization nor constraints are addressed in [9] (i.e. $\varphi \equiv 0$ and $X = \mathbb{R}^d$). Again, a consequence of our results is that no boundedness or additive noise assumptions are required. Our bounds depend on the local variance $\sigma(x^*)^2$. In case a precise tunning is not used, they also depend on $\mathsf{Q}(x^*, t_0) := \sigma_L^2 \max_{1\leq t\leq t_0-1} \mathbb{E}[\|x^\tau - x^*\|^2]$, a quantity which depends on the (local) per unit distance multiplicative variance σ_L^2. Here t_0 is a small number depending logarithmic on κ/N_0, where N_0 is the initial batch size. Our bounds show an interesting property: if an upper bound on κ is known, the stepsize and sampling rate policies can be tuned so that our bounds resemble, up to absolute constants, those obtained in [9] if *it was supposed* (8) *holds* but with σ^2 replaced by $\sigma(x^*)^2$ (without dependence on $\mathsf{Q}(x^*, t_0)$). We thus conclude a *regularization property* for strongly convex functions: the more *well-conditioned* the optimization problem is or the more aggressive the *initial batch size* is, the more our error bounds approach, up to absolute constants, those obtained in [9] if it was supposed (8) holds but with σ^2 replaced by $\sigma(x^*)^2$ (with this bound exactly achieved if an upper bound on κ is known). In case (8) indeed holds, our bounds are thus sharper than the ones in [9] since typically $\sigma(x^*)^2 \ll \sigma^2$ and[7] $\mathsf{Q}(t_0, x^*) \ll \sigma^2$ if $\max_{1\leq t\leq t_0-1} \mathbb{E}[\|x^\tau - x^*\|^2] \ll \mathscr{D}(X)^2$. See Example 1 in Sect. 1.1 and Theorem 2, Corollary 2 and Remark 3 in Sect. 4.

[7] We remark that $\max_{1\leq\tau\leq T} \mathbb{E}[\|x^\tau - x^*\|^2] \ll \mathscr{D}(X)^2$ always holds after a *finite* number T of iterations.

Let's call a SA method (near) optimal if it achieves, in terms of the tolerance ε, the order of the deterministic optimal iteration complexity with (near) optimal oracle complexity and average cost. To the best of our knowledge, we give for the first time (near) optimal SA methods for stochastic smooth CO with an oracle with *multiplicative noise* and a *general heavy-tailed distribution* (with finite 2nd moment). This is an important improvement since, in principle, it is not obvious that SA methods can converge if the oracle has an unbounded variance satisfying (11). This is even less obvious when using acceleration a la FISTA since, in that case, an extrapolation is computed with divergent weights $\beta_t = \mathcal{O}(t)$ (see Theorem 1 and Remark 2). Also, the introduction of a regularization term φ and constraints is nontrivial in the setting of a multiplicative noise. The main reason for obtaining non-asymptotic convergence under Assumption 2 and unbounded gradients is the use of dynamic sampling with a careful sampling rate, i.e., one that uses the order of the oracle complexity of the standard iterate averaging scheme. In these methods typically *one ergodic average of iterates* with size T at the *final T-th iteration* is used. Differently, in dynamic sampling schemes *local empirical averages of gradients* with smaller but increasing sizes are distributed *along iterations*.[8] This is also the reason our bounds depend on local variances at points of the trajectory of the method and at points of X^* (but not on the whole X). Such results are not shared by the SA method with constant N_t for ill-conditioned problems nor for the SAA method.

We review some works besides [9,22]. A variation of Assumption 2 was proposed by Iusem, Jofré, Oliveira and Thompson in [27], but their method is tailored at solving monotone variational inequalities with the extragradient method. Hence, on the class of smooth convex functions, the suboptimal iteration complexity of $\mathcal{O}(\varepsilon^{-1})$ is achieved with the use of an additional proximal step (not required for optimization). In [12,15], the assumption of multiplicative noise is also analyzed but their rate analysis is for the special class of stochastic convex quadratic objectives as in Example 1, where the main motivation is to study linear LSR problems. They obtain offline iteration and oracle complexities of $\mathcal{O}(\varepsilon^{-1})$ on the class of quadratic convex functions using Tykhonov regularization with iterate averaging. Differently, our analysis is optimal on the general class of smooth convex functions. See also the recent works [29,30] on LSR problems. Finally, we mention [3] which considers stochastic proximal gradient methods for composite problems still assuming an uniformly bounded variance (hypothesis H5, page 8). It should be noted, however, that such work allows non i.i.d. sampling and presents an interesting application in high-dimensional graphical models.

We focus now on the class of smooth strongly convex functions. In [16] the dynamic sampled SG method is also analyzed. However, their analysis strongly relies on finitely supported distributions as in aggregated methods, no oracle complexity bounds are provided and it is assumed no regularization nor constraints ($\varphi \equiv 0$ and $X = \mathbb{R}^d$). In [62], a method using $\ln(\varepsilon^{-1})$ projections for smooth strongly CO is proposed but still assuming boundedness of the oracle and obtaining the suboptimal iteration complexity $\mathcal{O}(\varepsilon^{-1})$. The works [5,19,41,53] do not require the assumption (8) and cope with multiplicative noise. However, their iteration and oracle complexities are

[8] The possibility of distributing the empirical averages along iterations is possible due to the on-line nature of the SA method. This is not shared by the SAA methodology which is an off-line procedure.

$\mathscr{O}(\varepsilon^{-1})$ and, hence, suboptimal when compared to our method (if we assume that the oracle can be accessed more than once per iteration). We remark that [19] allows the objective to be non-strongly convex (but requires it to be twice differentiable and to satisfy the Kurdyka-Lojiasewicz's condition).

Finally, we compare the results of item (ii) with [17], where also dynamic sampled optimal methods for stochastic smooth strongly convex optimization are derived (with knowledge of κ). The class of smooth functions analyzed in [17] are smaller, requiring them to be *twice differentiable*. With respect to statistical assumptions, their analysis allow multiplicative noise but it uses a condition much stronger than Assumption 2. Indeed, they assume a.e. $\nabla F(\cdot, \xi)$ is $\mathsf{L}(\xi)$-Lipschitz continuous with a *bounded Lipschitz modulus*, that is, $\sup_{\xi \in \Xi} \mathsf{L}(\xi) < \infty$. As a consequence, $\mathsf{L}(\cdot) - \mathbb{E}[\mathsf{L}(\xi)]$ is a sub-Gaussian random variable (whose tail decreases exponentially fast).[9] From (12), this implies that $\nabla F(x, \cdot)$ is also sub-Gaussian. In our Assumption 2, we allow heavy-tailed distributions (with finite 2nd moment). Precisely, for any $x \in X$, we only require that $\mathbb{E}[\|\nabla F(x, \cdot)\|^2] < \infty$ so that the fluctuations of $\nabla F(x, \cdot)$ can be much heavier than a Gaussian random variable. The work in [17] does not consider regularization nor constraints (i.e., $\varphi \equiv 0$ and $X = \mathbb{R}^d$), ignoring effects of φ and X on the oracle's variance.[10] Finally, their methods differ significantly from ours. Indeed, the methods in [17] are SA versions of the SVRG method of [32,60] originally designed for finite-sum objectives. Hence, besides dynamic mini-batching they require in every iteration an inner loop of m iterations to further reduce the variance of the gradients. This implies the need of a randomization scheme and, at least, an additional $m \geq 400 \cdot 3^6 \kappa$ number of samples per iteration (which can be large if $\kappa \gg 1$. See [17, Corollary 4]). Our method is solely based on the simple stochastic gradient method and we only use dynamic mini-batching to reduce variance with no additional randomization and sampling.

In Sect. 2 preliminaries and notation are presented while in Sects. 3 and 4 our convergence theory is presented for non-strongly and strongly convex problems respectively. Technical results are proved in the "Appendix".

2 Preliminaries and notation

For $x, y \in \mathbb{R}^n$, we denote $\langle x, y \rangle$ the standard inner product, and $\|x\| = \sqrt{\langle x, x \rangle}$ the correspondent Euclidean norm. Given $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we use the notation $\mathrm{d}(x, C) := \inf\{\|x - y\| : y \in C\}$. Given sequences $\{x^k\}$ and $\{y^k\}$, we use the notation $x^k = \mathscr{O}(y^k)$ or $\|x^k\| \lesssim \|y^k\|$ to mean that there exists a constant $C > 0$ such that $\|x^k\| \leq C\|y^k\|$ for all k. The notation $\|x^k\| \sim \|y^k\|$ means that $\|x^k\| \lesssim \|y^k\|$ and $\|y^k\| \lesssim \|x^k\|$. Given a σ-algebra $\mathscr{F}$ and a random variable ξ, we denote by $\mathbb{E}[\xi]$ and

[9] We say a centered random variable $q : \Xi \to \mathbb{R}$ is sub-Gaussian with parameter σ^2 if $\mathbb{E}\left\{e^{uq(\xi)}\right\} \leq e^{\frac{\sigma^2 u^2}{2}}$ for all $u \in \mathbb{R}$. We note that a centered Gaussian variable N with variance σ^2 satisfies $\mathbb{E}\left\{e^{uN(\xi)}\right\} = e^{\frac{\sigma^2 u^2}{2}}$ for all $u \in \mathbb{R}$.

[10] In the case that $X = \mathbb{R}^d$ and $\varphi \equiv 0$, the unique solution x^* satisfies $\nabla f(x^*) = 0$ so that $\sigma(x^*)^2 = \mathbb{E}[\|\nabla F(x^*, \xi)\|^2]$.

$\mathbb{E}[\xi|\mathscr{F}]$ the expectation and conditional expectation, respectively. We write $\xi \in \mathscr{F}$ for "ξ is $\mathscr{F}$-measurable" and $\xi \perp\!\!\!\perp \mathscr{F}$ for "ξ is independent of $\mathscr{F}$". We denote by $\sigma(\xi_1, \ldots, \xi_k)$ the σ-algebra generated by the random variables $\xi_1, \ldots, \xi_k$. Given the random variable ξ and $p \geq 1$, $|\xi|_p$ is the L^p-norm of ξ and $|\xi|_{\mathscr{F}}|_p := \sqrt[p]{\mathbb{E}[|\xi|^p|\mathscr{F}]}$ is the L_p-norm of ξ conditional to the σ-algebra $\mathscr{F}$. By $\lfloor x \rfloor$ and $\lceil x \rceil$ we mean the lowest integer greater and the highest integer lower than $x \in \mathbb{R}$, respectively. We use the notation $[m] := \{1, \ldots, m\}$ for $m \in \mathbb{N}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Given $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$.

Given a function $p : X \to \mathbb{R}$, $y \in X$ and $v \in \mathbb{R}^d$, we define

$$z \mapsto \ell_p(y, v; z) := p(y) + \langle v, z - y \rangle, \tag{13}$$

i.e., the linearization of p at the point y and direction v. If moreover p is differentiable, we define

$$z \mapsto \ell_p(y; z) := p(y) + \langle \nabla p(y), z - y \rangle, \tag{14}$$

i.e., the linearization of p at y. The *prox-mapping* with respect to X and a given convex function $\varphi : X \to \mathbb{R}$ is defined as $P_y^{\varphi}[u] := \operatorname{argmin}_{x \in X} \left\{ \langle u, x - y \rangle + \frac{1}{2}\|x - y\|^2 + \varphi(x) \right\}$. The following two properties are well known.

Lemma 2 *Let $p : X \to \mathbb{R}$ be a convex function and $\alpha > 0$. For any $y \in X$, if $z \in \operatorname{argmin}_{x \in X} \left\{ p(x) + \frac{1}{2\alpha}\|x - y\|^2 \right\}$, then*

$$p(z) + \frac{1}{2\alpha}\|z - y\|^2 \leq p(x) + \frac{1}{2\alpha}\|x - y\|^2 - \frac{1}{2\alpha}\|x - z\|^2, \quad \forall x \in X.$$

Proof See Lemma 1 in [36]. □

Lemma 3 *Let $p : X \to \mathbb{R}$ be a smooth convex function with L-Lipschitz gradient. Set $c_p := c > 0$ if p is c-strongly convex and $c_p = 0$ otherwise. Then the following relations hold:*

$$\ell_p(y; x) + \frac{c_p}{2}\|x - y\|^2 \leq p(x) \leq \ell_p(y; x) + \frac{L}{2}\|x - y\|^2, \forall x, y \in X.$$

Proof See Lemma 2 in [36]. □

We will use the following lemma many times. The proof is left to the "Appendix".

Lemma 4 (Oracle's variance decay under multiplicative noise) *Suppose* (1) *and Assumption* 2 *hold. Given an i.i.d. sample $\{\xi_j\}_{j=1}^N$ drawn from the distribution* **P** *and $x \in \mathbb{R}^d$, set*

$$\varepsilon(x) := \sum_{j=1}^{N} \frac{\nabla F(x, \xi_j) - \nabla f(x)}{N}.$$

Then $\|\|\varepsilon(x)\|\|_2 \leq \frac{\sigma(x^*) + \sigma_L \|x - x^*\|}{\sqrt{N}}$.

3 Smooth convex optimization with multiplicative noise and acceleration

We propose Algorithm 1 for the composite problem (5) assuming an stochastic oracle satisfying (1) and Assumption 2 of a multiplicative noise in the case the objective f is smooth convex satisfying (6). For $t \in \mathbb{N}$, we define the oracle error $\varepsilon^t := F_t'(y^t, \xi^t) - \nabla f(y^t)$ and the filtration $\mathscr{F}_t := \sigma(y^1, \xi^1 \ldots, \xi^t)$.

Algorithm 1 Stochastic approximated FISTA with dynamic mini-batching

1: INITIALIZATION: initial iterates $y^1 = z^0$, positive stepsize sequence $\{\alpha_t\}$, positive weights $\{\beta_t\}$ and sampling rate $\{N_t\}$.
2: ITERATIVE STEP: Given y^t and z^{t-1}, generate i.i.d. sample $\xi^t := \{\xi_j^t\}_{j=1}^{N_t}$ from **P** independently from previous samples. Compute

$$F_t'(y^t, \xi^t) := \frac{1}{N_t} \sum_{j=1}^{N_t} \nabla F\left(y^t, \xi_j^t\right).$$

Then set

$$\begin{aligned} z^t &:= P_{y^t}^{\alpha_t \varphi}\left[\alpha_t F'(y^t, \xi^t)\right] \qquad (15)\\ &= \operatorname*{argmin}_{x \in X} \left\{ \ell_f\left(y^t, F'(y^t, \xi^t); x\right) + \frac{1}{2\alpha_t}\|x - y^t\|^2 + \varphi(x) \right\}, \\ y^{t+1} &:= \frac{(\beta_t - 1)}{\beta_{t+1}}(z^t - z^{t-1}) + z^t, \qquad (16) \end{aligned}$$

Before presenting the convergence analysis of Algorithm 1 in detail, it is instructive to comment on its proof strategy. In Sect. 3.1, we first derive a recursive error bound for the optimality gap based on the assumptions of the stepsize and weight sequences. This error bound depends on the squared norm and on a linear function of the oracle's error ε^τ accumulated during the iterations [see definitions (18)–(19) in Proposition 1]. Due to the i.i.d. sampling assumption, $\{\varepsilon^\tau\}$ is a vector-valued martingale.

Section 3.2 is devoted to the L^2-boundedness of the iterates. This is achieved by using the error bound derived in Proposition 1 together with a judicious stopping-time argument. This is a crucial step in the proof since a priori boundedness of the iterates is not guaranteed when the oracle has multiplicative noise, the feasible set is unbounded and the objective is not strongly convex. Proposition 2 is also important to obtain bounds in terms of local variances by using the multiplicative noise assumption. The proof and statement of Proposition 2 reveals a non-trivial relationship between the use of acceleration, multiplicative noise and variance reduction. For instance, we crucially use that the extrapolation (16) depends explicitly on the two previous iterates and that $\beta_t \geq 1$. These two facts are important to derive a judicious lower bound in the stopping-time argument [see (40)]. Moreover, the upper bound stated in Proposition 2

are given in terms of t_0 initial iterates. The number t_0 depends on the acceleration weights and on the sampling rate (see further comments in Remark 2).

Section 3.3 concludes with Theorem 1 and Corollary 1 presenting the non-asymptotic convergence rate of the optimality gap and a bound on the oracle complexity. To derive such statements, Propositions 1 and 2 are applied together with specific policies for the acceleration weights, the stepsizes and the sampling rate (see Theorem 1). Roughly, $\beta_t = \frac{t+1}{2}$, $\alpha_t \sim [L + \mathscr{O}(1)LN_0^{-1/2}]^{-1}$ and $N_t \sim N_0 t^3 (\ln t)^2$. The weights β_t are the same used in the deterministic version of FISTA. The term $\mathscr{O}(1)LN_0^{-1/2}$ in the stepsize α_t must be present so that variance reduction can be handled with constant stepsizes. Finally, we remark that the cubic term t^3 in the sampling rate is related with the use of Nesterov's acceleration.[11]

3.1 Derivation of an error bound

In the following, it will be convenient to define, for any $t \geq 2$,

$$s^t := \beta_t z^t - (\beta_t - 1) z^{t-1}. \tag{17}$$

Proposition 1 *Suppose Assumptions 1 and 2 hold for the problem (5) satisfying (1) and (6). Suppose $\{\alpha_t\}$ is non-increasing and*

$$\alpha_t \in \left(0, \frac{1}{L}\right), \quad \beta_t \geq 1, \quad \beta_t^2 = \beta_{t+1}^2 - \beta_{t+1}, \quad \forall t \in \mathbb{N}.$$

Then the sequence generated by Algorithm 1 satisfies: for all $1 \leq t \leq T$ and $x^* \in X^*$,

$$2\alpha_{T+1}\beta_{T+1}^2 \left[g(z^{T+1}) - g^*\right] + \|s^{T+1} - x^*\|^2 \leq 2\alpha_t \beta_t^2 \left[g(z^t) - g^*\right] + \|s^t - x^*\|^2 + \sum_{\tau=t}^{T} \Delta A_{\tau+1} + \sum_{\tau=t}^{T} \Delta M_{\tau+1}(x^*),$$

where for any $\tau \in \mathbb{N}$ and $x^* \in X^*$, we have defined

$$\Delta A_{\tau+1} := \frac{\alpha_{\tau+1}^2 \beta_{\tau+1}^2}{(1 - L\alpha_{\tau+1})} \|\varepsilon^{\tau+1}\|^2, \tag{18}$$

$$\Delta M_{\tau+1}(x^*) := 2\alpha_{\tau+1}\beta_{\tau+1} \left\langle \varepsilon^{\tau+1}, x^* - s^\tau \right\rangle. \tag{19}$$

Moreover, $\mathbb{E}[\Delta M_{\tau+1}(x^*)] = 0$ for all $\tau \in \mathbb{N}$.

Proof For convenience of notation, we will use the notation $v_t := g(z^t) - g^*$. We start by deriving the following fundamental inequality. We have, for any $t \in \mathbb{N}_0$ and $x \in X$,

[11] In the non-accelerated version of the method, $N_t \sim t(\ln t)^2$ would be sufficient.

$$
\begin{aligned}
g(z^{t+1}) &\le \ell_f\left(y^{t+1}; z^{t+1}\right) + \varphi\left(z^{t+1}\right) + \frac{L}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 \\
&= \ell_f\left(y^{t+1}, F'\left(y^{t+1}, \xi^{t+1}\right); z^{t+1}\right) + \varphi\left(z^{t+1}\right) + \frac{L}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 \\
&\quad - \left\langle \varepsilon^{t+1}, z^{t+1} - y^{t+1}\right\rangle \\
&\le \ell_f\left(y^{t+1}, F'\left(y^{t+1}, \xi^{t+1}\right); x\right) + \varphi(x) + \frac{\alpha_{t+1}^{-1}}{2}\left\|x - y^{t+1}\right\|^2 \\
&\quad - \frac{\alpha_{t+1}^{-1}}{2}\left\|x - z^{t+1}\right\|^2 + \frac{\left(L - \alpha_{t+1}^{-1}\right)}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 \\
&\quad - \left\langle \varepsilon^{t+1}, z^{t+1} - y^{t+1}\right\rangle \\
&= \ell_f\left(y^{t+1}; x\right) + \varphi(x) + \frac{\alpha_{t+1}^{-1}}{2}\left\|x - y^{t+1}\right\|^2 - \frac{\alpha_{t+1}^{-1}}{2}\left\|x - z^{t+1}\right\|^2 \\
&\quad + \frac{\left(L - \alpha_{t+1}^{-1}\right)}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 \\
&\quad + \left\langle \varepsilon^{t+1}, x - y^{t+1}\right\rangle - \left\langle \varepsilon^{t+1}, z^{t+1} - y^{t+1}\right\rangle \\
&\le g(x) + \frac{\alpha_{t+1}^{-1}}{2}\left\|x - y^{t+1}\right\|^2 - \frac{\alpha_{t+1}^{-1}}{2}\left\|x - z^{t+1}\right\|^2 \\
&\quad + \frac{\left(L - \alpha_{t+1}^{-1}\right)}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 + \left\langle \varepsilon^{t+1}, x - z^{t+1}\right\rangle, \qquad (20)
\end{aligned}
$$

where we used $g(z^{t+1}) = f(z^{t+1}) + \varphi(z^{t+1})$ and the upper inequality of Lemma 3 with $p := f$ in first inequality (by smoothness of f), definitions (13)–(14) and $\nabla f(y^{t+1}) = F'(y^{t+1}, \xi^{t+1}) - \varepsilon^{t+1}$ in the first equality, the expression (15) and Lemma 2 with the convex function $p := \ell_f(y^{t+1}, F'(y^{t+1}, \xi^{t+1}); \cdot) + \varphi$, $\alpha := \alpha_{t+1}$, $y := y^{t+1}$ and $z := z^{t+1}$ in second inequality, definitions (13)–(14) and $F'(y^{t+1}, \xi^{t+1}) = \nabla f(y^{t+1}) + \varepsilon^{t+1}$ in the second equality as well as $g(x) = f(x) + \varphi(x)$ and Lemma 3 with $p := f$ in the last inequality (by convexity of f).

We now set $x := z^t$ and $x := x^* \in X^*$ in (20) obtaining

$$
\begin{aligned}
\left[g\left(z^{t+1}\right) - g^*\right] - \left[g\left(z^t\right) - g^*\right] &= g\left(z^{t+1}\right) - g\left(z^t\right) \\
&\le \frac{\alpha_{t+1}^{-1}}{2}\left\|z^t - y^{t+1}\right\|^2 - \frac{\alpha_{t+1}^{-1}}{2}\left\|z^t - z^{t+1}\right\|^2 \\
&\quad + \frac{\left(L - \alpha_{t+1}^{-1}\right)}{2}\left\|z^{t+1} - y^{t+1}\right\|^2 \\
&\quad + \left\langle \varepsilon^{t+1}, z^t - z^{t+1}\right\rangle, \qquad (21)
\end{aligned}
$$

$$g\left(z^{t+1}\right)-g^{*} \leq \frac{\alpha_{t+1}^{-1}}{2}\left\|x^{*}-y^{t+1}\right\|^{2}-\frac{\alpha_{t+1}^{-1}}{2}\left\|x^{*}-z^{t+1}\right\|^{2} + \frac{\left(L-\alpha_{t+1}^{-1}\right)}{2}\left\|z^{t+1}-y^{t+1}\right\|^{2} + \left\langle \varepsilon^{t+1}, x^{*}-z^{t+1}\right\rangle. \tag{22}$$

We multiply by $\beta_{t+1}-1$ relation (21), add the result to (22) and then further multiply by $2\alpha_{t+1}$ obtaining

$$\begin{aligned}
&2\alpha_{t+1}\beta_{t+1}v_{t+1}-2\alpha_{t+1}\left(\beta_{t+1}-1\right)v_{t}\\
&\leq\left(\beta_{t+1}-1\right)\left\|z^{t}-y^{t+1}\right\|^{2}-\left(\beta_{t+1}-1\right)\left\|z^{t}-z^{t+1}\right\|^{2}\\
&\quad+\left\|x^{*}-y^{t+1}\right\|^{2}-\left\|x^{*}-z^{t+1}\right\|^{2}\\
&\quad+\alpha_{t+1}\left(L-\alpha_{t+1}^{-1}\right)\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^{2}\\
&\quad+2\alpha_{t+1}\left\langle\varepsilon^{t+1},\left(\beta_{t+1}-1\right)z^{t}+x^{*}-\beta_{t+1}z^{t+1}\right\rangle\\
&=\beta_{t+1}\left\|z^{t}-y^{t+1}\right\|^{2}-\beta_{t+1}\left\|z^{t}-z^{t+1}\right\|^{2}\\
&\quad+\left\|z^{t}-z^{t+1}\right\|^{2}-\left\|z^{t}-y^{t+1}\right\|^{2}\\
&\quad+\left\|x^{*}-y^{t+1}\right\|^{2}-\left\|x^{*}-z^{t+1}\right\|^{2}\\
&\quad+\alpha_{t+1}\left(L-\alpha_{t+1}^{-1}\right)\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^{2}\\
&\quad+2\alpha_{t+1}\left\langle\varepsilon^{t+1},\left(\beta_{t+1}-1\right)z^{t}+x^{*}-\beta_{t+1}z^{t+1}\right\rangle.
\end{aligned}\tag{23}$$

We will use repeatedly the following Pythagorean relation:

$$\|a-c\|^{2}-\|b-c\|^{2}=-\|b-a\|^{2}+2\langle b-a, c-a\rangle. \tag{24}$$

Corresponding to the first line of (23), we multiply β_{t+1} in (24) with $a:=y^{t+1}$, $b:=z^{t+1}$ and $c:=z^{t}$, obtaining

$$\beta_{t+1}\left\|y^{t+1}-z^{t}\right\|^{2}-\beta_{t+1}\left\|z^{t+1}-z^{t}\right\|^{2}=-\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^{2} + 2\beta_{t+1}\left\langle z^{t+1}-y^{t+1}, z^{t}-y^{t+1}\right\rangle. \tag{25}$$

Corresponding to the second line of (23) we get

$$\left\|z^{t+1}-z^{t}\right\|^{2}-\left\|y^{t+1}-z^{t}\right\|^{2}=\left\|z^{t+1}-y^{t+1}\right\|^{2}+2\left\langle z^{t+1}-y^{t+1}, y^{t+1}-z^{t}\right\rangle, \tag{26}$$

by multiplying (25) with $-\beta_{t+1}^{-1}$. Finally, corresponding to the third line of (23) we get

$$\left\|y^{t+1}-x^*\right\|^2-\left\|z^{t+1}-x^*\right\|^2=-\left\|z^{t+1}-y^{t+1}\right\|^2+2\left\langle z^{t+1}-y^{t+1},x^*-y^{t+1}\right\rangle, \tag{27}$$

using (24) with $a:=y^{t+1}$, $b:=z^{t+1}$ and $c:=x^*$.

Now we sum the identities (25)–(27) and use the result in the right hand side of (23) obtaining

$$\begin{aligned}
&2\alpha_{t+1}\beta_{t+1}v_{k+1}-2\alpha_{t+1}(\beta_{t+1}-1)v_t\\
&\quad\le-\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^2\\
&\qquad+2\left\langle z^{t+1}-y^{t+1},\beta_{t+1}(z^t-y^{t+1})+x^*-z^t\right\rangle\\
&\qquad+\alpha_{t+1}(L-\alpha_{t+1}^{-1})\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^2\\
&\qquad+2\alpha_{t+1}\left\langle\varepsilon^{t+1},(\beta_{t+1}-1)z^t+x^*-\beta_{t+1}z^{t+1}\right\rangle\\
&\quad=-\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^2\\
&\qquad+2\left\langle z^{t+1}-y^{t+1},(\beta_{t+1}-1)z^t+x^*-\beta_{t+1}y^{t+1}\right\rangle\\
&\qquad+\alpha_{t+1}(L-\alpha_{t+1}^{-1})\beta_{t+1}\left\|z^{t+1}-y^{t+1}\right\|^2\\
&\qquad+2\alpha_{t+1}\left\langle\varepsilon^{t+1},(\beta_{t+1}-1)z^t+x^*-\beta_{t+1}z^{t+1}\right\rangle.
\end{aligned}$$

We multiply the above inequality by β_{t+1} and use $\beta_t^2=\beta_{t+1}^2-\beta_{t+1}$ as assumed in the proposition. We then obtain

$$\begin{aligned}
2\alpha_{t+1}\beta_{t+1}^2v_{t+1}-2\alpha_{t+1}\beta_t^2v_t\le&-\left\|\beta_{t+1}z^{t+1}-\beta_{t+1}y^{t+1}\right\|^2\\
&+2\left\langle\beta_{t+1}z^{t+1}-\beta_{t+1}y^{t+1},(\beta_{t+1}-1)z^t\right.\\
&\left.+x^*-\beta_{t+1}y^{t+1}\right\rangle\\
&+\alpha_{t+1}(L-\alpha_{t+1}^{-1})\beta_{t+1}^2\left\|z^{t+1}-y^{t+1}\right\|^2\\
&+2\alpha_{t+1}\beta_{t+1}\left\langle\varepsilon^{t+1},(\beta_{t+1}-1)z^t+x^*-\beta_{t+1}z^{t+1}\right\rangle.
\end{aligned}$$

Corresponding to the first two lines in the right hand side of the previous inequality, we invoke again the Pythagorean relation (24) with $a:=\beta_{t+1}y^{t+1}$, $b:=\beta_{t+1}z^{t+1}$ and $c:=(\beta_{t+1}-1)z^t+x^*$, obtaining

$$2\alpha_{t+1}\beta_{t+1}^2 v_{t+1} - 2\alpha_{t+1}\beta_t^2 v_t \le \left\|\beta_{t+1}y^{t+1} - (\beta_{t+1}-1)z^t - x^*\right\|^2 - \left\|\beta_{t+1}z^{t+1} - (\beta_{t+1}-1)z^t - x^*\right\|^2 + \alpha_{t+1}(L-\alpha_{t+1}^{-1})\beta_{t+1}^2\left\|z^{t+1}-y^{t+1}\right\|^2 - 2\alpha_{t+1}\beta_{t+1}\left\langle \varepsilon^{t+1}, \beta_{t+1}z^{t+1} - (\beta_{t+1}-1)z^t - x^*\right\rangle. \tag{28}$$

Concerning the last line of (28), we will rewrite it as

$$\left\langle \varepsilon^{t+1}, \beta_{t+1}z^{t+1} - (\beta_{t+1}-1)z^t - x^*\right\rangle = \left\langle \varepsilon^{t+1}, \beta_{t+1}z^{t+1} - \beta_{t+1}y^{t+1}\right\rangle + \left\langle \varepsilon^{t+1}, \beta_{t+1}y^{t+1} - (\beta_{t+1}-1)z^t - x^*\right\rangle. \tag{29}$$

Now, by the extrapolation (16) and definition (17), we have

$$\beta_{t+1}y^{t+1} - (\beta_{t+1}-1)z^t = \beta_t z^t - (\beta_t - 1)z^{t-1} = s^t. \tag{30}$$

Using definition (17) for index $t+1$ and (29)–(30) in (28) we obtain

$$2\alpha_{t+1}\beta_{t+1}^2 v_{t+1} - 2\alpha_{t+1}\beta_t^2 v_t \le \left\|s^t - x^*\right\|^2 - \left\|s^{t+1} - x^*\right\|^2 + \alpha_{t+1}\beta_{t+1}^2\left(L - \alpha_{t+1}^{-1}\right)\left\|y^{t+1} - z^{t+1}\right\|^2 + 2\alpha_{t+1}\beta_{t+1}^2\left\langle \varepsilon^{t+1}, y^{t+1} - z^{t+1}\right\rangle + 2\alpha_{t+1}\beta_{t+1}\left\langle \varepsilon^{t+1}, x^* - s^t\right\rangle. \tag{31}$$

We now bound the second line in the above inequality as

$$-\left(\alpha_{t+1}^{-1} - L\right)\left\|y^{t+1} - z^{t+1}\right\|^2 + 2\left\langle \varepsilon^{t+1}, y^{t+1} - z^{t+1}\right\rangle \le \frac{\left\|\varepsilon^{t+1}\right\|^2}{\left(\alpha_{t+1}^{-1} - L\right)},$$

using Young's inequality $2\langle a, b\rangle \le \frac{\|a\|^2}{\lambda} + \lambda\|b\|^2$ with $a := \varepsilon^{t+1}$, $b := y^{t+1} - z^{t+1}$ and $\lambda := \alpha_{t+1}^{-1} - L > 0$ by the assumption on $\{\alpha_t\}$. We now use the above relation and the hypothesis that $\alpha_{t+1} \le \alpha_t$ in relation (31). We then obtain the following recursion: for all $t \ge 1$,

$$2\alpha_{t+1}\beta_{t+1}^2 v_{t+1} - 2\alpha_t\beta_t^2 v_t \le 2\alpha_{t+1}\beta_{t+1}^2 v_{t+1} - 2\alpha_{t+1}\beta_t^2 v_t \le \left\|s^t - x^*\right\|^2 - \left\|s^{t+1} - x^*\right\|^2$$

$$+\frac{\alpha_{t+1}\beta_{t+1}^2}{\left(\alpha_{t+1}^{-1}-L\right)}\left\|\varepsilon^{t+1}\right\|^2 + 2\alpha_{t+1}\beta_{t+1}\left\langle\varepsilon^{t+1}, x^*-s^t\right\rangle. \tag{32}$$

To finally obtain the recursion stated in the proposition, given $1 \le t \le T$ and $x^* \in X^*$, we simply sum recursively the inequality (32) from $\tau := t$ to $\tau := T$ and use definitions of $\Delta A_{\tau+1}$ and $\Delta M_{\tau+1}(x^*)$.

We now prove the second claim by showing that $\{\Delta M_{t+1}(x^*), \mathscr{F}_t\}$ defines a martingale difference, i.e., $\mathbb{E}[\Delta M_{t+1}(x^*)|\mathscr{F}_t] = 0$ for all $t \in \mathbb{N}$. To show this, note that $y^{t+1} \in \mathscr{F}_t$ since $z^{t-1}, z^t \in \mathscr{F}_t$. Moreover $\xi^{t+1} \perp\!\!\!\perp \mathscr{F}_t$. Since $\varepsilon^{t+1} = \nabla F(y^{t+1}, \xi^{t+1}) - \nabla f(y^{t+1})$, the previous statements imply that $\mathbb{E}[\varepsilon^{t+1}|\mathscr{F}_t] = 0$ and

$$\mathbb{E}\left[\Delta M_{t+1}(x^*)\,|\mathscr{F}_t\right] = 2\alpha_{t+1}\beta_{t+1}\left\langle\mathbb{E}\left[\varepsilon^{t+1}\,|\mathscr{F}_t\right], x^*-s^t\right\rangle = 0,$$

where we also used that $s^t \in \mathscr{F}_t$ since $z^{t-1}, z^t \in \mathscr{F}_t$. Using $\mathbb{E}[\mathbb{E}[\cdot|\mathscr{F}_t]] = \mathbb{E}[\cdot]$, we further conclude that $\mathbb{E}[\Delta M_{t+1}(x^*)] = 0$ as required. □

3.2 L^2-boundedness of the iterates

In the case X is unbounded and the oracle has multiplicative noise, it is not possible to infer boundedness of $\{\|z^t\|_2\}_{t=1}^\infty$ a priori (i.e., L^2-boundedness of the iterates). In this section we obtain such L^2-boundedness when using stochastic approximation with dynamic mini-batches. This is essential to obtain complexity estimates in the following section.

Proposition 2 *Suppose Assumptions 1 and 2 hold for the problem (5) satisfying (1) and (6), $\{\alpha_t\}$ is non-increasing such that $\alpha_t \in \left(0, \frac{1}{L}\right)$ and $\beta_t \ge 1$, $\beta_t^2 = \beta_{t+1}^2 - \beta_{t+1}$ for all $t \in \mathbb{N}$. Suppose further that $\sum_{t=1}^\infty \frac{\alpha_{t+1}^2\beta_{t+1}^2}{(1-L\alpha_{t+1})N_{t+1}} < \infty$. Choose $t_0 \in \mathbb{N}$ and $\gamma > 0$ such that*

$$\sum_{t\ge t_0}^{\infty}\frac{\alpha_{t+1}^2\beta_{t+1}^2}{(1-L\alpha_{t+1})\,N_{t+1}} < \gamma < \frac{1}{15\sigma_L^2}. \tag{33}$$

Then for all $x^* \in X^*$,

$$\sup_{t\ge0}\left|\left\|z^t-x^*\right\|\right|_2^2 \le \frac{\max_{t\in[t_0]}\left\{2\alpha_{t_0}\beta_{t_0}^2\mathbb{E}\left[g(z^{t_0})-g^*\right]+\mathbb{E}\left[\left\|s^{t_0}-x^*\right\|^2\right]\right\}+\frac{\sigma(x^*)^2}{3\sigma_L^2}}{1-15\gamma\sigma_L^2}. \tag{34}$$

Proof For clarity of exposition we use the notation $v^t := \mathbb{E}[g(z^t) - g^*]$. Given $t \geq t_0$, we take total expectation in the inequality of Proposition 1 and get, using $v^{t+1} \geq 0$,

$$\mathbb{E}\left[\left\|s^{t+1} - x^*\right\|^2\right] \leq 2\alpha_{t+1}\beta_{k+1}^2 v^{t+1} + \mathbb{E}\left[\left\|s^{t+1} - x^*\right\|^2\right]$$
$$\leq 2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + \sum_{i=t_0}^{t} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \mathbb{E}\left[\left\|\varepsilon^{i+1}\right\|^2\right]. \tag{35}$$

Let $i \in \mathbb{N}$. From definition (16) we have

$$y^{i+1} - x^* = \left(\frac{\beta_i - 1}{\beta_{i+1}} + 1\right)\left(z^i - x^*\right) - \left(\frac{\beta_i - 1}{\beta_{i+1}}\right)\left(z^{i-1} - x^*\right). \tag{36}$$

We now use the above expression in Lemma 4. Precisely, we use Assumptions 1 and 2, $\varepsilon^{i+1} = F'(y^{i+1}, \xi^{i+1}) - \nabla f(y^{i+1})$, definition of $F'(y^{i+1}, \xi^{i+1})$ in Algorithm 1, $\xi^{i+1} \perp\!\!\!\perp \mathscr{F}_{i+1}$ and $y^{i+1} \in \mathscr{F}_{i+1}$. Then we get

$$\sqrt{N_{i+1}} \cdot \left|\left\|\varepsilon^{i+1}\right\| \middle| \mathscr{F}_{i+1}\right|_2 \leq \sigma(x^*) + \sigma_L \left\|y^{i+1} - x^*\right\|$$
$$\leq \sigma(x^*) + \sigma_L \left(\frac{\beta_i - 1}{\beta_{i+1}} + 1\right)\left\|z^i - x^*\right\| + \sigma_L \left(\frac{\beta_i - 1}{\beta_{i+1}}\right)\left\|z^{i-1} - x^*\right\|.$$

From the above, we use $\|\cdot|\mathscr{F}_{i+1}|_2|_2 = |\cdot|_2$ and take squares to get

$$N_{i+1} \cdot \left|\left\|\varepsilon^{i+1}\right\|\right|_2^2 \leq 3\sigma(x^*)^2 + 3\sigma_L^2 \left(\frac{\beta_i - 1}{\beta_{i+1}} + 1\right)^2 \left|\left\|z^i - x^*\right\|\right|_2^2 + 3\sigma_L^2 \left(\frac{\beta_i - 1}{\beta_{i+1}}\right)^2 \left|\left\|z^{i-1} - x^*\right\|\right|_2^2, \tag{37}$$

where we used the relation $(\sum_{i=1}^3 a_i)^2 \leq 3\sum_{i=1}^3 a_i^2$.

For simplicity, we define $q_i := \frac{\beta_i - 1}{\beta_{i+1}}$ and $d_i := \left|\|z^i - x^*\|\right|_2$. From (35), (37) and $0 \leq q_i \leq 1$ for all $i \in \mathbb{N}$, we finally get the recursion for any $t \geq t_0$,

$$\mathbb{E}\left[\left\|s^{t+1} - x^*\right\|^2\right] \leq 2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + \sum_{i=t_0}^{t} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{3\sigma(x^*)^2}{N_{i+1}},$$
$$+ \sum_{i=t_0}^{t} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{12\sigma_L^2}{N_{i+1}} d_i^2 + \sum_{i=t_0}^{t} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{3\sigma_L^2}{N_{i+1}} d_{i-1}^2. \tag{38}$$

For any $a > 0$, we define the stopping time

$$\tau_a := \inf \{t \geq t_0 : d_t > a\}, \tag{39}$$

where t_0 and γ are as defined in the statement of the proposition. Note that for any $t \in \mathbb{N}$,

$$s^{t+1} - x^* = \beta_{t+1}\left(z^{t+1} - x^*\right) - (\beta_{t+1} - 1)\left(z^t - x^*\right).$$

From the above equality, $\beta_{t+1} \geq 1$, the triangle inequality for $\|\cdot\|$ and Minkowski's inequality for $|\cdot|_2$, we get

$$\left|\left\|s^{t+1} - x^*\right\|\right|_2 \geq \beta_{t+1}\left|\left\|z^{t+1} - x^*\right\|\right|_2 - (\beta_{t+1} - 1)\left|\left\|z^t - x^*\right\|\right|_2.$$

The above relation implies that for any $a > 0$ such that $\tau_a < \infty$,

$$\left\|s^{\tau_a} - x^*\right\|_2 \geq \beta_{\tau_a} d_{\tau_a} - \left(\beta_{\tau_a} - 1\right) d_{\tau_a - 1} > \beta_{\tau_a} a - \left(\beta_{\tau_a} - 1\right) a = a, \tag{40}$$

since $d_{\tau_a} > a$ and $d_{\tau_a - 1} \leq a$ by definition (39).[12]

From (33) and (38)–(40), we have that for any $a > 0$ such that $\tau_a < \infty$,

$$\begin{aligned} a^2 < \mathbb{E}\left[\left\|s^{\tau_a} - x^*\right\|^2\right] &\leq 2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + \sum_{i=t_0}^{\tau_a - 1} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{3\sigma(x^*)^2}{N_{i+1}}, \\ &+ \sum_{i=t_0}^{\tau_a - 1} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{12\sigma_L^2}{N_{i+1}} d_i^2 + \sum_{i=t_0}^{\tau_a - 1} \frac{\alpha_{i+1}\beta_{i+1}^2}{\left(\alpha_{i+1}^{-1} - L\right)} \cdot \frac{3\sigma_L^2}{N_{i+1}} d_{i-1}^2 \\ &\leq 2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + 3\gamma\left[\sigma(x^*)^2 + 5\sigma_L^2 a^2\right], \end{aligned}$$

and hence,

$$a^2 < \frac{2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + 3\gamma\sigma(x^*)^2}{1 - 15\gamma\sigma_L^2}, \tag{41}$$

where we used that $0 < \gamma < \frac{1}{15\sigma_L^2}$. By definition of τ_a for any $a > 0$, the argument above shows that any threshold a^2 which the sequence $\{d_t^2\}_{t \geq t_0}$ eventually exceeds is bounded above by (41). Hence $\{d_t^2\}_{t \geq t_0}$ is bounded and

$$\sup_{t \geq t_0} d_t^2 \leq \frac{2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0} - x^*\right\|^2\right] + 3\gamma\sigma(x^*)^2}{1 - 15\gamma\sigma_L^2}.$$

[12] We note here the importance of assuming $\beta_t \geq 1$ and the specific form of the extrapolation in (16) in terms of previous iterates.

Since the denominator above is less then 1, the bound above implies further that

$$\sup_{t\geq 0} d_t^2 \leq \frac{\max_{t\in[t_0]}\left\{2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0}-x^*\right\|^2\right]\right\} + 3\gamma\sigma(x^*)^2}{1-15\gamma\sigma_L^2}$$
$$\leq \frac{\max_{t\in[t_0]}\left\{2\alpha_{t_0}\beta_{t_0}^2 v^{t_0} + \mathbb{E}\left[\left\|s^{t_0}-x^*\right\|^2\right]\right\} + \frac{\sigma(x^*)^2}{5\sigma_L^2}}{1-15\gamma\sigma_L^2},$$

where we used $0<\gamma<\frac{1}{15\sigma_L^2}$. This concludes the proof of the claim. □

3.3 Convergence rate and oracle complexity

We now derive a convergence rate and estimate the oracle complexity of Algorithm 1.

Theorem 1 *Suppose Assumptions 1 and 2 hold for the problem (5) satisfying (1) and (6). Let $\{y^t, z^t\}$ be the sequence generated by Algorithm 1. Let $\mu\in(0,1)$, $a,b,\delta>0$, $N_0\in\mathbb{N}$ and set for all $t\in\mathbb{N}$,*

$$\beta_t := \frac{1+t}{2},\quad \alpha_t := \frac{\mu}{L+\frac{a}{\sqrt{N_0}}},\quad N_t := N_0\left\lfloor (t+2+\delta)^3\left[\ln(t+2+\delta)\right]^{1+2b}\right\rfloor.$$

Choose $\phi\in(0,1)$ and let $t_0 := t_0(\alpha_1\sigma_L, N_0, b, \delta)\in\mathbb{N}$ be given by

$$t_0 := \left\lceil \exp\left\{\sqrt[2b]{\frac{15(\alpha_1\sigma_L)^2}{8\phi N_0 b}}\right\} - 1 - \delta\right\rceil \bigvee 1. \tag{42}$$

Then Proposition 2 holds. Moreover, given $x^*\in X^*$ and $\mathsf{J} := \mathsf{J}(x^*,t_0) > 0$ such that

$$\sup_{\tau\in\mathbb{N}}\left\|\left\|z^\tau - x^*\right\|\right\|_2^2 \leq \mathsf{J},$$

the following bound holds for all $t\in\mathbb{N}$,

$$\mathbb{E}\left[g(z^{t+1})-g^*\right] \leq \frac{4\mathbb{E}\left[g(z^1)-g^*\right]}{(t+2)^2} + 2\left[\frac{L}{\mu}+\frac{a}{\mu\sqrt{N_0}}\right]\frac{\mathbb{E}\left[\left\|s^1-x^*\right\|^2\right]}{(t+2)^2}$$
$$+\left[\frac{L}{\mu}+\frac{a}{\mu\sqrt{N_0}}\right]\cdot\frac{3\mu^2}{4(1-\mu)a^2b\left[\ln(2+\delta)\right]^{2b}}\cdot\frac{\sigma(x^*)^2}{(t+2)^2}$$
$$+\left[\frac{L}{\mu}+\frac{a}{\mu\sqrt{N_0}}\right]\cdot\frac{15\mu^2}{4(1-\mu)N_0b\left[\ln(2+\delta)\right]^{2b}}\cdot\frac{(\sigma_L/L)^2\mathsf{J}}{(t+2)^2}.$$

Proof We first show that $\{\beta_t\}$, $\{\alpha_t\}$ and $\{N_t\}$ satisfy the conditions of Propositions 1 and 2. We have that $\{\alpha_t\}$ is a constant (an hence non-increasing) sequence satisfying

$0 < L\alpha_t \leq \mu$. By inspection, it is easy to check that $\beta_t \geq 1$ and $\beta_t^2 = \beta_{t+1}^2 - \beta_t^2$ for all $t \geq 1$. Also,

$$\sum_{t=1}^{\infty} \frac{\alpha_{t+1}\beta_{t+1}^2}{\left(\alpha_{t+1}^{-1} - L\right) N_{t+1}} \leq \frac{\alpha_1^2}{4(1-L\alpha_1)N_0} \sum_{t=1}^{\infty} \frac{1}{(t+2+\delta)[\ln(t+2+\delta)]^{1+2b}} < \infty.$$

Hence, we have have shown that the policies $\{\beta_t\}$, $\{\alpha_t\}$ and $\{N_t\}$ satisfy the conditions of Propositions 1 and 2.

For $\phi \in (0, 1)$, we want t_0 to satisfy

$$\sum_{t \geq t_0}^{\infty} \frac{\alpha_{t+1}\beta_{t+1}^2}{\left(\alpha_{t+1}^{-1} - L\right) N_{t+1}} \leq \frac{\phi}{15\sigma_L^2}, \tag{43}$$

as prescribed in Proposition 2. We have

$$\begin{aligned}
\sum_{t \geq t_0}^{\infty} \frac{\alpha_{t+1}\beta_{t+1}^2}{\left(\alpha_{t+1}^{-1} - L\right) N_{t+1}} &\leq \frac{\alpha_1^2}{4(1-L\alpha_1)N_0} \sum_{t \geq t_0}^{\infty} \frac{1}{(t+2+\delta)[\ln(t+2+\delta)]^{1+2b}} \\
&\leq \frac{\alpha_1^2}{4(1-L\alpha_1)N_0} \int_{t_0-1}^{\infty} \frac{\mathrm{d}\, t}{(t+2+\delta)[\ln(t+2+\delta)]^{1+2b}} \\
&= \frac{\alpha_1^2}{4(1-L\alpha_1)N_0} \cdot \frac{1}{2b\ln(t_0+1+\delta)^{2b}} \\
&\leq \frac{\alpha_1^2}{8(1-\mu)N_0 b \ln(t_0+1+\delta)^{2b}}.
\end{aligned} \tag{44}$$

From the above relation, it is sufficient to choose t_0 as the minimum natural number such that the right hand side of (44) is less than $\phi/15\sigma_L^2$. This is satisfied by t_0 in (42).

Let $x^* \in X^*$. From Proposition 2 and (43) we know that $\{\|\|z^\tau - x^*\|\|_2\}_{\tau \geq 1}$ is bounded, say $\sup_{\tau \in \mathbb{N}} \|\|z^\tau - x^*\|\|_2^2 \leq \mathsf{J}$. From this, (37) and $\sup_{\tau \in \mathbb{N}} \frac{\beta_\tau - 1}{\beta_{\tau+1}} \leq 1$ we get

$$\mathbb{E}\left[\left\|\varepsilon^{\tau+1}\right\|^2\right] \leq \frac{3\sigma(x^*)^2 + 15\sigma_L^2\mathsf{J}}{N_{\tau+1}}. \tag{45}$$

We now bound the expectation of the sum $\sum_t \Delta A_t$ in the inequality of Proposition 1. Precisely, for any $t \geq 1$,

$$\begin{aligned}
\sum_{\tau=1}^{t} \mathbb{E}\left[\Delta A_\tau\right] &= \sum_{\tau=1}^{t} \frac{\alpha_{\tau+1}^2\beta_{\tau+1}^2\mathbb{E}\left[\left\|\varepsilon^{\tau+1}\right\|^2\right]}{(1-L\alpha_{\tau+1})} \\
&\leq \frac{\alpha_1^2}{4(1-\mu)} \sum_{\tau=1}^{t} (\tau+2)^2 \mathbb{E}\left[\left\|\varepsilon^{\tau+1}\right\|^2\right]
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\alpha_1^2\left[3\sigma(x^*)^2+15\sigma_L^2\mathsf{J}\right]}{4(1-\mu)N_0}\sum_{\tau=1}^{t}\frac{(\tau+2)^2}{(\tau+2+\delta)^3\left[\ln(\tau+2+\delta)\right]^{1+2b}}\\
&\leq \frac{\alpha_1^2\left[3\sigma(x^*)^2+15\sigma_L^2\mathsf{J}\right]}{4(1-\mu)N_0}\sum_{\tau=1}^{t}\frac{1}{(\tau+2+\delta)\left[\ln(\tau+2+\delta)\right]^{1+2b}}\\
&\leq \frac{\alpha_1^2\left[3\sigma(x^*)^2+15\sigma_L^2\mathsf{J}\right]}{4(1-\mu)N_0}\int_{\tau=0}^{t}\frac{\mathrm{d}\,\tau}{(\tau+2+\delta)\left[\ln(\tau+2+\delta)\right]^{1+2b}}\\
&\leq \frac{\alpha_1^2\left[3\sigma(x^*)^2+15\sigma_L^2\mathsf{J}\right]}{8(1-\mu)N_0 b\left[\ln(2+\delta)\right]^{2b}}=\frac{3\left[\alpha_1\sigma(x^*)\right]^2+15\left(\alpha_1\sigma_L\right)^2\mathsf{J}}{8(1-\mu)N_0 b\left[\ln(2+\delta)\right]^{2b}},
\end{aligned}$$

where we used (45) and definition of N_τ in third inequality.

From the above inequality and the bounds

$$\left[\alpha_1\sigma(x^*)\right]^2\leq\left[\frac{\mu\sqrt{N_0}}{a}\sigma(x^*)\right]^2=\frac{\mu^2N_0\sigma(x^*)^2}{a^2},\qquad(\alpha_1\sigma_L)^2\leq\left(\frac{\mu}{L}\sigma_L\right)^2=\frac{\mu^2\sigma_L^2}{L^2},$$

we finally get

$$\sum_{\tau=1}^{t}\mathbb{E}\left[\Delta A_\tau\right]\leq\frac{3\mu^2\sigma(x^*)^2}{8(1-\mu)a^2b\left[\ln(2+\delta)\right]^{2b}}+\frac{15\mu^2(\sigma_L/L)^2\mathsf{J}}{8(1-\mu)N_0b\left[\ln(2+\delta)\right]^{2b}}.\tag{46}$$

We also have for any $t\geq 1$,

$$\alpha_t^{-1}=\frac{L}{\mu}+\frac{a}{\mu\sqrt{N_0}},\quad\frac{1}{\beta_t^2}=\frac{4}{(t+1)^2}.\tag{47}$$

Finally, we take total expectation in the inequality of Proposition 1 for $t:=1$ and $T:=t$ and use $\mathbb{E}[\Delta M_t(x^*)]=0$ for all $t\in\mathbb{N}$, $\beta_1=1$ and (46)–(47) to derive the required claim. □

Remark 1 Regarding the constant J in Theorem 1, from Proposition 2 we have the upper bound

$$\mathsf{J}\leq\frac{\max_{t\in[t_0]}\left\{\frac{1}{2}\alpha_1(t_0+1)^2\mathbb{E}\left[g(z^{t_0})-g^*\right]+\mathbb{E}\left[\left\|s^{t_0}-x^*\right\|^2\right]\right\}+\frac{\sigma(x^*)^2}{5\sigma_L^2}}{1-\phi}.\tag{48}$$

Corollary 1 *Let the assumptions of Theorem 1 hold. Given $\varepsilon>0$, Algorithm 1 achieves the tolerance $\mathbb{E}\left[g(z^T)-g^*\right]\leq\varepsilon$ after $T=\mathscr{O}(\varepsilon^{-\frac{1}{2}})$ iterations using an oracle complexity of*

$$\sum_{\tau=1}^{T}N_\tau\leq\mathscr{O}\left(\varepsilon^{-2}\right)\left[\ln\left(\varepsilon^{-\frac{1}{2}}\right)\right]^2.$$

Proof In Theorem 1, we set $b = \frac{1}{2}$. For every $t \in \mathbb{N}$, let B_{t+1} be the right hand side expression in the bound of the optimality gap $\mathbb{E}[g(z^{t+1}) - g^*]$ stated in Theorem 1. Up to a constant $B > 0$, for every $t \in \mathbb{N}$, we have

$$\mathbb{E}\left[g(z^t) - g^*\right] \leq B_t \leq \frac{B}{t^2}.$$

Given $\varepsilon > 0$, let T be the least natural number such that $BT^{-2} \leq \varepsilon$. Then $T = \mathcal{O}(\varepsilon^{-\frac{1}{2}})$, $\mathbb{E}\left[g(z^T) - g^*\right] \leq \varepsilon$ and

$$\sum_{\tau=1}^{T} N_\tau \lesssim \sum_{\tau=1}^{T} \tau^3 (\ln \tau)^2 \lesssim T^4 (\ln T)^2 \lesssim \varepsilon^{-2} \left[\ln\left(\varepsilon^{-\frac{1}{2}}\right)\right]^2.$$

We have thus proved the required claims. □

Remark 2 (Constants for the smooth non-strongly convex case) We discuss the constants in the bounds of Theorem 1 and Corollary 1 and compare it to the bounds in [22] under (8). The optimality gap rate in Theorem 1 depends on t_0 *initial iterates* (with possibly $t_0 > 1$) given in (42) and (48). This requirement is needed in Proposition 2 since *no boundedness of the oracle's variance is assumed* a priori. Another distinctive feature is the presence of the factor $(t_0 + 1)^2$ in (48) as a consequence of *acceleration* under Assumption 2. These observations require showing that t_0 in (42) is not too large. In that respect, we note that t_0 *does not depend on* $x^* \in X^*$, but only on $(\alpha_1 \sigma_L)^2$ and the exogenous parameters ϕ, N_0, b and δ. If we assume the standard Lipschitz continuity (12) in Lemma 1, then $\sigma_L = 2L$ for $L := |\mathsf{L}(\xi)|_2$. Also, the stepsize satisfies $\alpha_1 \leq \frac{\mu}{L}$ so that $(\alpha_1 \sigma_L)^2 \leq 4\mu^2$.

We thus conclude: the iteration t_0 is dictated solely by the multiplicative per unit distance variance σ_L^2, independently of the variances $\{\sigma(x)^2\}_{x \in X}$ at the points of the feasible set X. Moreover, assuming L is known for the stepsize policy, there exists an upper bound on t_0 which is also independent of σ_L^2 and only depends on the exogenous parameters μ, N_0, b and δ chosen on the stepsize and sampling rate policies.

For instance, if we set $\phi := \mu := b := \frac{1}{2}$, $N_0 := 2$, then $t_0 \sim \lceil 44.52 - \delta \rceil$. We now set $\delta := 44$ so that $t_0 = 1$ and $(t_0 + 1)^2 = 4$. We further choose $a \sim L$. Then using (48), $\alpha_1 \leq \frac{\mu}{L}$, the bound in Theorem 1 becomes of the form

$$\mathbb{E}\left[g(z^{t+1}) - g^*\right] \lesssim \frac{1}{t^2}\left\{\mathbb{E}\left[g(z^1) - g^* + L\left\|s^1 - x^*\right\|^2\right] + \frac{\sigma(x^*)^2}{L}\right\}. \quad (49)$$

We note that we may further obtain

$$\mathbb{E}\left[g(z^1) - g^* + L\left\|s^1 - x^*\right\|^2\right] \lesssim L\left\|z^0 - x^*\right\|^2 + \frac{\sigma(x^*)^2}{L} \quad (50)$$

by using inequality (20) in the proof of Proposition 1. For completeness, we derive this statement.

Set $t := 0$ and $x := x^* \in X^*$ in (20). We then obtain

$$\begin{aligned} g(z^1) - g^* + \frac{\alpha_1^{-1}}{2}\left\|z^1 - x^*\right\|^2 &\leq \frac{\alpha_1^{-1}}{2}\left\|y^1 - x^*\right\|^2 - \frac{\left(\alpha_1^{-1} - L\right)}{2}\left\|z^1 - y^1\right\|^2 \\ &\quad + \left\langle \varepsilon^1, y^1 - z^1\right\rangle + \left\langle \varepsilon^1, x^* - y^1\right\rangle \\ &\leq \frac{\alpha_1^{-1}}{2}\left\|y^1 - x^*\right\|^2 + \frac{1}{2\left(\alpha_1^{-1} - L\right)}\left\|\varepsilon^1\right\|^2 \\ &\quad + \left\langle \varepsilon^1, x^* - y^1\right\rangle, \end{aligned}$$

where we used Young's inequality $\langle \varepsilon^1, y^1 - z^1\rangle \leq \frac{1}{2\lambda}\|\varepsilon^1\|^2 + \frac{\lambda}{2}\|y^1 - z^1\|^2$ with $\lambda := \alpha_1^{-1} - L > 0$. From the above inequality, $\mathbb{E}[\langle \varepsilon^1, x^* - y^1\rangle] = 0$ and $\mathbb{E}[\|\varepsilon^1\|^2] \leq \frac{2\sigma(x^*)^2 + 2\sigma_L^2\|y^1 - x^*\|^2}{N_1}$, which follows from Lemma 4, we get that

$$\begin{aligned} \mathbb{E}\left[g(z^1) - g^*\right] + \frac{\alpha_1^{-1}}{2}\mathbb{E}\left[\left\|z^1 - x^*\right\|^2\right] &\leq \left[\frac{\alpha_1^{-1}}{2} + \frac{\sigma_L^2}{N_1\left(\alpha_1^{-1} - L\right)}\right]\left\|y^1 - x^*\right\|^2 \\ &\quad + \frac{\sigma(x^*)^2}{N_1\left(\alpha_1^{-1} - L\right)}. \end{aligned}$$

From the previous inequality, $y^1 = z^0$, $s^1 = z^1$ and (47), we obtain

$$\begin{aligned} \mathbb{E}\left[g(z^1) - g^*\right] &\leq \left[\frac{\frac{L}{\mu} + \frac{a}{\mu\sqrt{N_0}}}{2} + \frac{\sigma_L^2}{N_1\left[\frac{L(1-\mu)}{\mu} + \frac{a}{\mu\sqrt{N_0}}\right]}\right]\left\|z^0 - x^*\right\|^2 \\ &\quad + \frac{\sigma(x^*)^2}{N_1\left[\frac{L(1-\mu)}{\mu} + \frac{a}{\mu\sqrt{N_0}}\right]}, \end{aligned}$$

and

$$\begin{aligned} &\left(\frac{L}{\mu} + \frac{a}{\mu\sqrt{N_0}}\right)\mathbb{E}\left[\left\|s^1 - x^*\right\|^2\right] \\ &\leq \left[\frac{L}{\mu} + \frac{a}{\mu\sqrt{N_0}} + \frac{2\sigma_L^2}{N_1\left[\frac{L(1-\mu)}{\mu} + \frac{a}{\mu\sqrt{N_0}}\right]}\right]\left\|z^0 - x^*\right\|^2 + \frac{2\sigma(x^*)^2}{N_1\left[\frac{L(1-\mu)}{\mu} + \frac{a}{\mu\sqrt{N_0}}\right]}. \end{aligned}$$

From the two previous inequalities and previous choice of parameters (recall that $\sigma_L \sim L$ and $a \sim L$), we obtain (50). This and (49) imply that the bound of Theorem 1 becomes of the form

$$\mathbb{E}\left[g(z^{t+1}) - g^*\right] \lesssim \frac{1}{t^2}\left\{L\left\|z^0 - x^*\right\|^2 + \frac{\sigma(x^*)^2}{L}\right\}.$$

The above inequality *resembles, up to absolute constants*, bounds obtained in [22] if it was supposed that (8) holds *but replacing*[13] σ by $\sigma(x^*)$. In this sense, we improve upon the mentioned bounds by showing that under the more aggressive setting of Assumption 2, our bounds depends on *local variances* $\sigma(x^*)^2$ at points $x^* \in X^*$. Moreover, in case (8) indeed holds, our bounds are sharper than in [22] since typically $\sigma(x^*)^2 \ll \sigma^2$ for large $\sqrt{\mathscr{D}(X)}$ (see Example 1). This may be seen as a *localization property* of Algorithm 1 in terms of the oracle's variance.

4 Smooth strongly convex optimization with multiplicative noise

We propose Algorithm 2 for the problem (5) assuming an stochastic oracle satisfying (1) and Assumption 2 (multiplicative noise) in the case the objective f is smooth strongly convex satisfying (6) and (7). For $t \in \mathbb{N}$, we define the stochastic error $\varepsilon^t := F_t'(x^t, \xi^t) - \nabla f(x^t)$ and the filtration $\mathscr{F}_t := \sigma(x^1, \xi^1 \ldots, \xi^{t-1})$.

Algorithm 2 Stochastic proximal gradient method with dynamic mini-batching

1: INITIALIZATION: initial iterate x^1, positive stepsize sequence $\{\alpha_t\}$ and sampling rate $\{N_t\}$.
2: ITERATIVE STEP: Given x^t, generate i.i.d. sample $\xi^t := \{\xi_j^t\}_{j=1}^{N_t}$ from $\mathbf{P}$ independently from previous samples. Compute

$$F_t'(x^t, \xi^t) := \frac{1}{N_t}\sum_{j=1}^{N_t} \nabla F(x^t, \xi_j^t).$$

Then set

$$\begin{aligned} x^{t+1} &:= P_{x^t}^{\alpha_t \varphi}\left[\alpha_t F'(x^t, \xi^t)\right] \\ &= \operatorname*{argmin}_{x \in X}\left\{\ell_f\left(x^t, F'(x^t, \xi^t); x\right) + \frac{1}{2\alpha_t}\left\|x - x^t\right\|^2 + \varphi(x)\right\}. \end{aligned} \tag{51}$$

We comment on the proof strategy used in the convergence analysis of Algorithm 2. In Sect. 4.1, we derive recursive error bounds for the squared norm $\|x^t - x^*\|^2$ and the optimality gap based on the assumptions of the stepsize. In this quest we use Lemmas 2–3, the i.i.d. sampling assumption and Lemma 4 several times. The error

[13] We refer to [22, Corollary 5], equations (3.38) and (3.40). For the comparison mentioned, set $L_f = 0$ and $L_\Psi \tilde{D}^2 \sim \frac{\sigma^2}{L_\Psi}$ following the notation in [22]. The scaling $L_\Psi \tilde{D}^2 \sim \frac{\sigma^2}{L_\Psi}$ is the one used in our setting where the sampling rate m_k in (3.38) is independent of L_Ψ and σ^2.

bound is stated in Proposition 3. It is interesting to remark that, differently to the non-strongly convex case (which needed Proposition 2 subsequently after Proposition 1), the L^2-boundedness of the iterates in case the objective is strongly convex is a direct consequence of the recursion derived in Proposition 3. Moreover, such recursion also gives upper bounds in terms of the local variances $\sigma(x^*)^2$ and σ_L^2. This important difference is a result of the coercivity property of a strongly convex objective (Lemma 3) and the variance decay established in Lemma 4.

Section 4.2 concludes with Theorem 2 and Corollary 2 presenting the non-asymptotic convergence rate of $\|x^t - x^*\|^2$ and the optimality gap and a bound on the oracle complexity. To derive such statements, Proposition 3 is applied together with specific policies for the stepsizes and the sampling rate. These are of the form $\alpha_t = \mu/L$ and $N_t \sim N_0 \zeta^{-t}$ with $\mu, \zeta \in (0, 1)$ (see Theorem 2). Typically, if $\sigma_L \sim L$, the choice $\mu := \mathscr{O}(1) \in (0, 1)$, $N_0 := \mathscr{O}(\kappa)$ and $\zeta := 1 - \mathscr{O}(\kappa^{-1})$ suffice to produce the rate stated in Sect. 1.2.

4.1 Derivation of error bounds

Proposition 3 *Suppose Assumption 1 and 2 hold for the problem (5) satisfying (1) as well as (6)–(7). Let x^* be the unique solution of (5). Suppose $0 < \alpha_t < \frac{1}{L}$ for all $t \in \mathbb{N}$.*

Then the sequence generated by Algorithm 2 satisfies for all $t \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2 | \mathscr{F}_t\right] \leq \left[1 - c\alpha_t + \frac{2(\alpha_t \sigma_L)^2}{(1 - L\alpha_t)N_t}\right]\left\|x^t - x^*\right\|^2 + \frac{2\alpha_t^2 \sigma(x^*)^2}{(1 - L\alpha_t)N_t},$$

$$\mathbb{E}\left[g(x^{t+1}) - g^*\right] \leq \left[\frac{\left(\alpha_t^{-1} - c\right)}{2} + \frac{2\alpha_t \sigma_L^2}{(1 - L\alpha_t)N_t}\right]\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] + \frac{2\alpha_t \sigma(x^*)^2}{(1 - L\alpha_t)N_t}.$$

Proof We use (51) and Lemma 2 with the convex function $p := \ell_f\left(x^t, F'(x^t, \xi^t); \cdot\right) + \varphi$ and $\alpha := \alpha_t$, $y := x^t$, $z := x^{t+1}$ and $x := x^*$ obtaining

$$\begin{aligned}
\frac{1}{2\alpha_t}\left\|x^* - x^{t+1}\right\|^2 &\leq \frac{1}{2\alpha_t}\left\|x^* - x^t\right\|^2 + \ell_f\left(x^t, F_t'\left(x^t, \xi^t\right); x^*\right) + \varphi(x^*) \\
&\quad - \ell_f\left(x^t, F_t'\left(x^t, \xi^t\right); x^{t+1}\right) - \varphi\left(x^{t+1}\right) - \frac{1}{2\alpha_t}\left\|x^{t+1} - x^t\right\|^2 \\
&= \frac{1}{2\alpha_t}\left\|x^* - x^t\right\|^2 + \ell_f\left(x^t, \nabla f\left(x^t\right); x^*\right) + \varphi(x^*) \\
&\quad - \ell_f\left(x^t, \nabla f(x^t); x^{t+1}\right) - \varphi\left(x^{t+1}\right) \\
&\quad - \frac{1}{2\alpha_t}\left\|x^{t+1} - x^t\right\|^2 + \left\langle \varepsilon^t, x^* - x^t\right\rangle + \left\langle \varepsilon^t, x^t - x^{t+1}\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{2\alpha_t}\left\|x^*-x^t\right\|^2+\left[f(x^*)+\varphi(x^*)\right]-\frac{c}{2}\left\|x^*-x^t\right\|^2\\
&\quad-\left[f(x^{t+1})+\varphi(x^{t+1})\right]+\frac{L}{2}\left\|x^{t+1}-x^t\right\|^2\\
&\quad-\frac{1}{2\alpha_t}\left\|x^{t+1}-x^t\right\|^2+\left\langle\varepsilon^t,x^*-x^t\right\rangle+\left\langle\varepsilon^t,x^t-x^{t+1}\right\rangle\\
&=\frac{(1-c\alpha_t)}{2\alpha_t}\left\|x^*-x^t\right\|^2+g^*-g(x^{t+1})\\
&\quad-\frac{(1-L\alpha_t)}{2\alpha_t}\left\|x^{t+1}-x^t\right\|^2+\left\langle\varepsilon^t,x^t-x^{t+1}\right\rangle\\
&\quad+\left\langle\varepsilon^t,x^*-x^t\right\rangle\\
&\leq\frac{(1-c\alpha_t)}{2\alpha_t}\left\|x^*-x^t\right\|^2+\frac{\alpha_t}{2(1-L\alpha_t)}\left\|\varepsilon^t\right\|^2+\left\langle\varepsilon^t,x^*-x^t\right\rangle,
\end{aligned}
\tag{52}
$$

where we used (13)–(14) and definition of ε^t in the first equality, the lower and upper inequalities of Lemma 3 for $p := f$ (by strong convexity and smoothness of f) in second inequality while in the last inequality we used $g^*-g(x^{t+1})\leq 0$ and Young's inequality $2\langle\varepsilon^t,x^t-x^{t+1}\rangle\leq\lambda^{-1}\|\varepsilon^t\|^2+\lambda\|x^{t+1}-x^t\|^2$ with $\lambda:=\frac{1-L\alpha_t}{\alpha_t}>0$ (since $0<L\alpha_t<1$ by assumption).

We now observe that $x^t\in\mathscr{F}_t,\xi^t\perp\!\!\!\perp\mathscr{F}_t,\varepsilon^t=F'(x^t,\xi^t)-\nabla f(x^t)$, Assumption 1 and (1) imply that $\mathbb{E}[\langle\varepsilon^t,x^*-x^t\rangle|\mathscr{F}_t]=0$. Using this observation and $x^t\in\mathscr{F}_t$, we have that

$$
\mathbb{E}\left[\frac{\alpha_t}{2(1-L\alpha_t)}\left\|\varepsilon^t\right\|^2+\left\langle\varepsilon^t,x^*-x^t\right\rangle\middle|\mathscr{F}_t\right]\leq\frac{2\alpha_t}{(1-L\alpha_t)N_t}\left[\sigma(x^*)^2+\sigma_L^2\left\|x^t-x^*\right\|^2\right],
\tag{53}
$$

where we have used the relation $(a+b)^2\leq 2a^2+2b^2$ and $\|\|\varepsilon^t\||\mathscr{F}_t|_2\leq\frac{\sigma(x^*)+\sigma_L\|x^t-x^*\|}{\sqrt{N_t}}$, which follows from Assumptions 1 and 2, $\varepsilon^t=\sum_{j=1}^{N_t}\frac{\nabla F(x^t,\xi_j^t)-\nabla f(x^t)}{N_t}$, relation (1) and Lemma 4.

We take $\mathbb{E}[\cdot|\mathscr{F}_t]$, use $x^t\in\mathscr{F}_t$ and multiply by $2\alpha_t$ in (52) and use (53) to get

$$
\begin{aligned}
\mathbb{E}\left[\left\|x^{t+1}-x^*\right\|^2\middle|\mathscr{F}_t\right]&\leq(1-c\alpha_t)\left\|x^*-x^t\right\|^2\\
&\quad+\frac{2\alpha_t^2}{(1-L\alpha_t)N_t}\left[\sigma(x^*)^2+\sigma_L^2\left\|x^t-x^*\right\|^2\right]\\
&=\left[1-c\alpha_t+\frac{2\alpha_t^2\sigma_L^2}{(1-L\alpha_t)N_t}\right]\left\|x^t-x^*\right\|^2\\
&\quad+\frac{2\alpha_t^2\sigma(x^*)^2}{(1-L\alpha_t)N_t}.
\end{aligned}
\tag{54}
$$

To conclude, we take total expectation above and use the hereditary property $\mathbb{E}[\mathbb{E}[\cdot|\mathscr{F}_t]] = \mathbb{E}[\cdot]$.

We now prove the second claim. We use the upper inequality of Lemma 3 with $p := f$ (by smoothness of f) and obtain

$$\begin{aligned}
g(x^{t+1}) &\leq \ell_f(x^t; x^{t+1}) + \varphi(x^{t+1}) + \frac{L}{2}\left\|x^{t+1} - x^t\right\|^2 \\
&= \ell_f(x^t, F'(x^t, \xi^t); x^{t+1}) + \varphi(x^{t+1}) + \frac{L}{2}\left\|x^{t+1} - x^t\right\|^2 + \left\langle -\varepsilon^t, x^{t+1} - x^t\right\rangle \\
&\leq \ell_f(x^t, F'(x^t, \xi^t); x^*) + \varphi(x^*) + \frac{\alpha_k^{-1}}{2}\left\|x^* - x^t\right\|^2 \\
&\quad - \frac{\left(\alpha_t^{-1} - L\right)}{2}\left\|x^{t+1} - x^t\right\|^2 - \frac{\alpha_t^{-1}}{2}\left\|x^* - x^{t+1}\right\|^2 + \left\langle \varepsilon^t, x^t - x^{t+1}\right\rangle \\
&= \ell_f(x^t; x^*) + \varphi(x^*) + \frac{\alpha_t^{-1}}{2}\left\|x^* - x^t\right\|^2 - \frac{\left(\alpha_t^{-1} - L\right)}{2}\left\|x^{t+1} - x^t\right\|^2 \\
&\quad - \frac{\alpha_t^{-1}}{2}\left\|x^* - x^{t+1}\right\|^2 + \left\langle \varepsilon^t, x^* - x^t\right\rangle + \left\langle \varepsilon^t, x^t - x^{t+1}\right\rangle \\
&\leq g(x^*) - \frac{c}{2}\left\|x^* - x^t\right\|^2 + \frac{\alpha_t^{-1}}{2}\left\|x^* - x^t\right\|^2 + \left\langle \varepsilon^t, x^* - x^t\right\rangle \\
&\quad - \frac{(1 - L\alpha_t)}{2\alpha_t}\left\|x^{t+1} - x^t\right\|^2 + \left\langle \varepsilon^t, x^t - x^{t+1}\right\rangle \\
&\leq g(x^*) + \frac{\left(\alpha_t^{-1} - c\right)}{2}\left\|x^* - x^t\right\|^2 + \left\langle \varepsilon^t, x^* - x^t\right\rangle + \frac{\alpha_t\left\|\varepsilon^t\right\|^2}{2(1 - L\alpha_t)}, \qquad (55)
\end{aligned}$$

where we used (13)–(14), definition of the prox-mapping and definition of ε^t in the equalities, the expression (51) and Lemma 2 with $\alpha := \alpha_t, y := x^t, z := x^{t+1}, x := x^*$ and the convex function $p := \ell_f\left(x^t, F'(x^t, \xi^t); \cdot\right) + \varphi$ in the second inequality, the lower inequality of Lemma 3 with $p := f$ (by strong convexity of f) in third inequality while in last inequality we used Young's inequality $2\langle \varepsilon^t, x^t - x^{t+1}\rangle \leq \lambda^{-1}\|\varepsilon^t\|^2 + \lambda\|x^{t+1} - x^t\|^2$ with $\lambda := \frac{1 - L\alpha_t}{\alpha_t} > 0$ (since $0 < L\alpha_t < 1$ by assumption).

We take $\mathbb{E}[\cdot|\mathscr{F}_t]$ and use $x^t \in \mathscr{F}_t$ in (55) and then further use (53) to finally obtain

$$\begin{aligned}
\mathbb{E}\left[g(x^{t+1}) - g^* \middle| \mathscr{F}_t\right] &\leq \frac{\left(\alpha_t^{-1} - c\right)}{2}\left\|x^* - x^t\right\|^2 \\
&\quad + \frac{2\alpha_t}{(1 - L\alpha_t)N_t}\left[\sigma(x^*)^2 + \sigma_L^2\left\|x^t - x^*\right\|^2\right] \\
&= \left[\frac{\left(\alpha_t^{-1} - c\right)}{2} + \frac{2\alpha_t\sigma_L^2}{(1 - L\alpha_t)N_t}\right]\left\|x^t - x^*\right\|^2 + \frac{2\alpha_t\sigma(x^*)^2}{(1 - L\alpha_t)N_t}.
\end{aligned}$$

Finally, we take $\mathbb{E}[\cdot|\mathscr{F}_t]$ again and use $\mathbb{E}[\mathbb{E}[\cdot|\mathscr{F}_t]] = \mathbb{E}[\cdot]$ to finish the proof. □

4.2 Convergence rate and oracle complexity

We now derive a convergence rate and estimate the oracle complexity of Algorithm 2.

Theorem 2 *Suppose Assumption 1–2 hold for the problem (5) satisfying (1) as well as (6)–(7). Let x^* be the unique solution of (5). Let $\mu \in (0,1)$ and ϕ such that $0 < \phi < \mu\frac{c}{2L} < 1$. Choose $\zeta \in (0,1)$ and $N_0 \in \mathbb{N}$ and set the stepsize and sampling rate sequences as*

$$\alpha_t \equiv \alpha := \frac{\mu}{L}, \quad N_t = N_0 \lfloor \zeta^{-t} \rfloor, \quad (t \in \mathbb{N}).$$

We define $\rho := \left(1 - \mu\frac{c}{2L} + \phi\right) \bigvee \zeta < 1$ and

$$t_0 := \left\lceil \log_{\frac{1}{\zeta}} \left(\frac{2\mu^2}{(1-\mu)\phi N_0} \cdot \frac{\sigma_L^2}{L^2} \right) \right\rceil \bigvee 1. \tag{56}$$

Then there exists constant $\mathsf{C} > 0$ such that

$$\mathbb{E}\left[\left\| x^{t+1} - x^* \right\|^2 \right] \le \mathsf{C}\rho^{t+1}, \quad \forall t \ge 1 \tag{57}$$

$$\mathbb{E}\left[g(x^{t+1}) - g^* \right] \le \left[\frac{\left(L\mu^{-1} - c\right)}{2} + \frac{2\mu}{(1-\mu)N_0} \cdot \frac{\sigma_L^2}{L} \zeta^t \right] \mathsf{C}\rho^t + \frac{2\mu}{(1-\mu)N_0} \cdot \frac{\sigma (x^*)^2}{L} \zeta^t, \quad \forall t \ge 2 \tag{58}$$

Proof For simplicity of notation, in the following we set $v_t := \mathbb{E}\left[\|x^t - x^*\|^2 \right]$, and define

$$\beta := \frac{2\alpha^2 \sigma (x^*)^2}{(1 - L\alpha)N_0}, \quad \delta := \frac{2(\alpha\sigma_L)^2}{(1 - L\alpha)N_0}, \quad \lambda := 1 - c\alpha. \tag{59}$$

Note that from the definitions of λ and $\alpha = \frac{\mu}{L}$, we get $\lambda = 1 - \mu\frac{c}{L} \in (0,1)$.

From the first recursion in Proposition 3, $N_t^{-1} \le N_0^{-1}\zeta^t$ and the above definitions, we have for all $t \in \mathbb{N}$,

$$v_{t+1} \le \left(\lambda + \delta\zeta^t\right) v_t + \beta\zeta^t. \tag{60}$$

Moreover, by definitions of t_0, δ, λ, ρ and $\alpha = \mu/L$ we have for all $t \ge t_0$,

$$\lambda < \lambda + \delta\zeta^t \le \lambda + \delta\zeta^{t_0} \le \lambda + \phi < \rho. \tag{61}$$

We now claim that for all $t \ge t_0$,

$$v_{t+1} \le (\lambda + \phi)^{t+1-t_0} v_{t_0} + \beta \sum_{\tau=0}^{t-t_0} (\lambda + \phi)^\tau \zeta^{t-\tau}. \tag{62}$$

We prove the claim by induction. Indeed, for $t := t_0$ the claim (62) follows from (60). Supposing (62) holds for $t \geq t_0$, then again by (60),

$$\begin{aligned} v_{t+2} &\leq \left(\lambda + \delta\zeta^{t+1}\right) v_{t+1} + \beta\zeta^{t+1} \\ &\leq (\lambda+\phi)\left[(\lambda+\phi)^{t+1-t_0} v_{t_0} + \beta \sum_{\tau=0}^{t-t_0} (\lambda+\phi)^{\tau}\zeta^{t-\tau}\right] + \beta\zeta^{t+1} \\ &= (\lambda+\phi)^{t+2-t_0} v_{t_0} + \beta \sum_{\tau=1}^{t+1-t_0} (\lambda+\phi)^{\tau}\zeta^{t+1-\tau} + \beta\zeta^{t+1} \\ &= (\lambda+\phi)^{t+2-t_0} v_{t_0} + \beta \sum_{\tau=0}^{t+1-t_0} (\lambda+\phi)^{\tau}\zeta^{t+1-\tau}, \end{aligned} \tag{63}$$

where we used (61) in second inequality. This shows (62) holds for $t+1$. The claim (62) is hence proved.

We will now bound the sum in (62). First, we note that $\rho \geq 1 - \mu\frac{c}{2L} + \phi$ and $\lambda + \phi < \rho$ which imply

$$\rho \geq \lambda + \phi + \mu\frac{c}{2L} \Rightarrow \frac{1}{1 - \frac{\lambda+\phi}{\rho}} \leq \frac{2L\rho}{c\mu}. \tag{64}$$

For $t \geq t_0$, we have

$$\begin{aligned} \beta \sum_{\tau=0}^{t-t_0} (\lambda+\phi)^{\tau}\zeta^{t-\tau} &\leq \frac{2\alpha^2\sigma(x^*)^2}{(1-L\alpha)N_0} \sum_{\tau=0}^{t-t_0} (\lambda+\phi)^{\tau}\rho^{t-\tau} \\ &= \frac{2\alpha^2\sigma(x^*)^2}{(1-L\alpha)N_0}\rho^t \sum_{\tau=0}^{t-t_0} \left(\frac{\lambda+\phi}{\rho}\right)^{\tau} \\ &\leq \frac{2\alpha^2\sigma(x^*)^2}{(1-L\alpha)N_0} \cdot \frac{\rho^t}{1-\frac{\lambda+\phi}{\rho}} \\ &\leq \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}\rho^{t+1}, \end{aligned} \tag{65}$$

where we used definition of β and $\zeta \leq \rho$ in first inequality, (61) in the second inequality and $\alpha = \mu/L$ and (64) in the last one.

From recursion (62) and the bound (65), we get for $t \geq t_0$,

$$\begin{aligned} v_{t+1} &\leq (\lambda+\phi)^{t+1-t_0} v_{t_0} + \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}\rho^{t+1} \\ &\leq \rho^{t+1}(\lambda+\phi)^{-t_0} v_{t_0} + \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}\rho^{t+1} = \mathsf{C}_0\rho^{t+1}, \end{aligned} \tag{66}$$

using (61) in the second inequality and the definitions of λ in (59) and C_0 in (77) in the equality.

The above relation implies that the sequence $\{v_t\}$ has linear convergence once the iteration t_0+1 is achieved. By changing the constants C_0 and ρ properly, this implies that the whole sequence $\{v_t\}$ has linear convergence. For our purposes (see Remark 3), we next derive a refined bound in terms of constants showing that linear convergence is obtained from the initial iteration when *only* C_0 is changed in (66) in a prescribed way.

From (60), we also have for $1 \le t < t_0$,

$$v_{t+1} \le (\lambda + \delta\zeta^t)v_t + \beta\zeta^t \le \lambda v_t + \left(\delta \max_{\tau\in[t_0-1]} v_\tau + \beta\right)\zeta^t.$$

We may use the above relation and proceed by induction analogously to (62)–(63) to get for $1 \le t < t_0$,

$$v_{t+1} \le \lambda^t v_1 + \left(\delta \max_{\tau\in[t_0-1]} v_\tau + \beta\right)\sum_{\tau=0}^{t-1}\lambda^\tau\zeta^{t-\tau} \tag{67}$$

The sum above is bounded by

$$\sum_{\tau=0}^{t-1}\lambda^\tau\zeta^{t-\tau} \le \sum_{\tau=0}^{t-1}\lambda^\tau\rho^{t-\tau} = \rho^t\sum_{\tau=0}^{t-1}\left(\frac{\lambda}{\rho}\right)^\tau \le \frac{\rho^t}{1-\frac{\lambda}{\rho}} < \frac{2L\rho^{t+1}}{c\mu}, \tag{68}$$

where we used $\zeta \le \rho$ in first inequality, (61) in second inequality and (64) in the last one. From (67)–(68), we get for $1 \le t < t_0$,

$$\begin{aligned} v_{t+1} &\le \rho^{t+1}\lambda^{-1}v_1 + \frac{2L}{c\mu}\left(\delta \max_{\tau\in[t_0-1]} v_\tau + \beta\right)\rho^{t+1} \\ &= \rho^{t+1}\left[\lambda^{-1}v_1 + \frac{4\left(\frac{L}{c}\right)(\alpha\sigma_L)^2}{\mu(1-\mu)N_0}\max_{\tau\in[t_0-1]} v_\tau + \frac{4\left(\frac{L}{c}\right)\alpha^2\sigma(x^*)^2}{\mu(1-\mu)N_0}\right] \\ &= \mathsf{C}_1\rho^{t+1}, \end{aligned} \tag{69}$$

where we used (61) in the inequality, definitions of δ, β, $\alpha = \mu/L$ and

$$\mathsf{C}_1 := \lambda^{-1}v_1 + \frac{4\mu\sigma_L^2}{(1-\mu)N_0Lc}\cdot\max_{\tau\in[t_0-1]} v_\tau + \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}, \tag{70}$$

in the equalities.

From (69), we have in particular $v_{t_0} \leq \mathsf{C}_1 \rho^{t_0}$. Using this and (66), we get for $t \geq t_0$,

$$v_{t+1} \leq (\lambda+\phi)^{t+1-t_0}\mathsf{C}_1\rho^{t_0} + \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}\rho^{t+1}$$
$$\leq \rho^{t+1-t_0}\mathsf{C}_1\rho^{t_0} + \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}\rho^{t+1} = \mathsf{C}\rho^{t+1}, \tag{71}$$

using (61) in the second inequality and definitions of C_1 in (70) and of C in (74) in the equality.

From relation (71) established for $t \geq t_0$, relation (69) established for $1 \leq t < t_0$ and $\mathsf{C}_1 = \mathsf{C} - \frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc} < \mathsf{C}$, we finally prove (57).

To prove the second claim in (58), we use $L\alpha_t = \mu$, $N_t = N_0\lfloor\zeta^{-t}\rfloor$ and the derived relation (57) for $\mathbb{E}[\|x^t - x^*\|^2]$ in the second recursion of Proposition 3 which bounds $\mathbb{E}[g(x^{t+1}) - g^*]$ in terms of $\mathbb{E}[\|x^t - x^*\|^2]$. □

Corollary 2 *Let assumptions of Theorem 2 hold. Then there exists constant* $\mathsf{A} > 0$ *such that for given* $\varepsilon > 0$*, Algorithm 2 achieves the tolerance* $\mathbb{E}\left[g(x^T) - g^*\right] \leq \varepsilon$ *after* $T = \mathscr{O}\left(\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})\right)$ *iterations using an oracle complexity of* $\sum_{\tau=1}^{T} N_\tau \leq \frac{\mathsf{A}N_0\cdot\mathscr{O}(\varepsilon^{-1})}{\rho(1-\zeta)}\cdot\left(\frac{\rho}{\zeta}\right)^{\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})+1}$.

In particular, if the stepsize parameter μ *and sampling rate parameter* ζ *satisfy* $\zeta = 1 - \mathsf{a}\mu\frac{c}{2L} + \phi$ *for some* $\mathsf{a} \in (0,1]$ *and* $0 < \phi < \mathsf{a}\mu\frac{c}{2L}$*, then the oracle complexity is*

$$\sum_{\tau=1}^{T} N_\tau \leq \frac{\mathsf{A}N_0}{\rho\left(\mathsf{a}\mu\frac{c}{2L} - \phi\right)}\mathscr{O}\left(\varepsilon^{-1}\right).$$

Up to constants, the same oracle complexity is achieved for the iteration error $\mathbb{E}[\|x^T - x^*\|^2]$.

Proof We only prove the result to $\mathbb{E}[g(x^T) - g^*]$ as the proof for $\mathbb{E}[\|x^T - x^*\|^2]$ is similar. Set $\varepsilon > 0$ and let T be minimum number of iterations for $\mathbb{E}[g(x^T) - g^*] \leq \varepsilon$ to hold. We have

$$\sum_{\tau=1}^{T} N_\tau \lesssim N_0\sum_{\tau=1}^{T}(\zeta^{-1})^\tau = N_0\frac{(\zeta^{-1})^T - 1}{1-\zeta} \leq N_0\frac{(\zeta^{-1})^T}{1-\zeta}. \tag{72}$$

From Theorem 2, for some constant $\mathsf{A} > 0$, $T \leq \log_{\frac{1}{\rho}}\left(\mathsf{A}\varepsilon^{-1}\right) + 1$. Using this we get

$$(\zeta^{-1})^T \leq (\zeta^{-1})^{\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})+1} = \left(\frac{\rho}{\zeta}\right)^{\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})+1}\cdot\left(\frac{1}{\rho}\right)^{\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})+1}$$
$$= \mathsf{A}\rho\varepsilon^{-1}\left(\frac{\rho}{\zeta}\right)^{\log_{\frac{1}{\rho}}(\mathsf{A}\varepsilon^{-1})+1}. \tag{73}$$

From (72)–(73), we prove the first claim of the corollary. If μ, ζ and $\mathsf{a} \in (0, 1]$ are chosen as stated in the corollary, we have $\rho = \zeta$ by definition of ρ in Theorem 2. From this, $\zeta = 1 - \mathsf{a}\mu\frac{c}{2L} + \phi$ and (72)–(73), we prove the second statement of the corollary. □

Remark 3 (Constants for the strongly convex case) As in Remark 2, we compare the bounds of Theorem 2 with the bounds given in [9] under (8). By the proof of Theorem 2, the constant C in Theorem 2 is

$$\mathsf{C} := \left(1 - \mu\frac{c}{L}\right)^{-1} \left\|x^1 - x^*\right\|^2 + \frac{4\mu}{(1-\mu)N_0 Lc}\left\{\sigma_L^2 \cdot \max_{\tau\in[t_0-1]} \mathbb{E}\left[\left\|x^\tau - x^*\right\|^2\right] + 2\sigma(x^*)^2\right\}, \tag{74}$$

where t_0 is estimated by (56). As in the case of ill-conditioned smooth convex functions, we may have $t_0 > 1$. However, such dependence is milder since we do not have the factor $(t_0+1)^2$ and t_0 is *logarithmic* with the endogenous and exogenous parameters. As a result, larger values of t_0 are acceptable. An interesting property in the case of strongly convex functions under Assumption 2 is that the L^2-boundedness and linear convergence of the generated sequence are obtained in the same proof. Next, we show that t_0 in (56) is not too large in comparison to problems satisfying (8). We note that t_0 does not depend on any $x \in X$, but only on $\left(\frac{\sigma_L}{L}\right)^2$, $\phi := \phi(\mu, \kappa)$ and the exogenous parameters μ, ζ and N_0. Let us assume the standard Lipschitz continuity (12) of Lemma 1 so that $\sigma_L = 2L$ for $L := |\mathsf{L}(\xi)|_2$, and $\alpha \le \frac{\mu}{L}$. Without loss on generality, we may set $\phi := \mu/(4\kappa)$ in Theorem 2 obtaining

$$t_0 = \left\lceil \log_{\frac{1}{\zeta}}\left(\frac{16\mu\kappa}{(1-\mu)N_0}\right)\right\rceil \bigvee 1, \tag{75}$$

from (56). Thus $t_0 = \mathscr{O}(1)\log_{\zeta^{-1}}(\kappa/N_0)$ for a given set of exogenous parameters.

We thus conclude: the iteration t_0 is dictated solely by the multiplicative per unit distance variance σ_L^2 and the condition number κ, independently of the variances $\{\sigma(x)^2\}_{x\in X}$ at the points of the feasible set X. Moreover, assuming L is known for the stepsize policy, there exists an upper bound on t_0 which is also independent of σ_L^2 and that only depends on $\log_{\zeta^{-1}}(\kappa/N_0)$ and the exogenous parameters μ, N_0, ζ defined in the stepsize and sampling rate policies.

The proof of Theorem 2 also says that for all $t \ge t_0$, $\mathbb{E}\left[\|x^{t+1} - x^*\|^2\right] \le \mathsf{C}_0\rho^{t+1}$ and for all $t \ge t_0 + 1$,

$$\mathbb{E}\left[g(x^{t+1}) - g^*\right] \le \left[\frac{(L\mu^{-1} - c)}{2} + \frac{2\mu}{(1-\mu)N_0}\cdot\frac{\sigma_L^2}{L}\zeta^t\right]\mathsf{C}_0\rho^t + \frac{2\mu}{(1-\mu)N_0}\cdot\frac{\sigma(x^*)^2}{L}\zeta^t, \tag{76}$$

where

$$\mathsf{C}_0 := \left(1-\mu\frac{c}{L}+\phi\right)^{-t_0}\mathbb{E}\left[\left\|x^{t_0}-x^*\right\|^2\right]+\frac{4\mu\sigma(x^*)^2}{(1-\mu)N_0Lc}. \tag{77}$$

If we compare the constant C in (74) with the constant C_0 in (77), we note that if $t_0>1$ then C_0 has a larger factor $\left(1-\mu\frac{c}{L}+\phi\right)^{-t_0}$ when compared to $\left(1-\mu\frac{c}{L}+\phi\right)^{-1}$ but C_0 does not have the additional term $\frac{\sigma_L^2}{Lc}\cdot\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]$ found in C. This last term is of the order of $\kappa\cdot\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]$ and may be larger than $\frac{\sigma(x^*)^2}{Lc}$.

Based on (75)–(77), we can obtain $t_0=1$ by setting, $\mu:=1/2$ and $N_0:=16\kappa$. Since $t_0=1$, the linear convergence stated by (76)–(77) is an improvement when compared to a policy in which $t_0>1$ since for $t_0=1$ there is no dependence on $\kappa\cdot\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]$. For $t_0=1$, the bound given by (76)–(77) states that for all $t\geq 2$,

$$\mathbb{E}\left[g(x^{t+1})-g^*\right]\lesssim\left[L\left\|x^1-x^*\right\|^2+\frac{\sigma(x^*)^2}{c}\right]\rho^{t-1},$$

where we only considered the dominant terms [ignoring $-c$ and the decaying terms of $\mathscr{O}(\zeta^t)$]. The above bound has the following property (P): it *resembles, up to absolute constants,* the bound obtained in [9] for the strongly convex case if it was supposed that (8) *holds but replacing* $\frac{\sigma^2}{c}$ *with* $\frac{\sigma(x^*)^2}{c}$ (see[14] [9, Theorem 4.2]). In this sense, we improve on the mentioned results by showing that under the more aggressive setting of Assumption 2, our bounds depends on the *local variance* $\sigma(x^*)^2$. Moreover, in case (8) indeed holds, our bounds are sharper than the ones in [9] since typically $\sigma(x^*)^2\ll\sigma^2$ for large $\sqrt{\mathscr{D}(X)}$ (see Example 1). This may be seen as a *localization property* of Algorithm 2 in terms of the oracle's variance.

We also observe the mild dependence $t_0\sim\mathscr{O}(1)\ln_{\zeta^{-1}}(\mu\kappa/(1-\mu)N_0)$ with κ. Hence, $t_0\to 1$ as either μ decreases, N_0 increases or κ decreases. We thus conclude that property (P) tends to be satisfied for better conditioned problems or for bigger initial batch sizes. Also, the bounds given by (58) and (74) depend on the local variance estimation $\sigma(x^*)^2+\sigma_L^2\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]$. Moreover, if (8) indeed holds, our rate statements given by (58) and (74) may be sharper than the ones in [9] since typically $\sigma(x^*)^2+\sigma_L^2\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]\ll\sigma^2$ if $\max_{\tau\in[t_0-1]}\mathbb{E}[\|x^\tau-x^*\|^2]\ll\mathscr{D}(X)^2$. See Example 1. This is again a localization property of Algorithm 2 in terms of the oracle's variance.

Remark 4 (Robust sampling) From Theorems 1 and 2, $\{\alpha_t\}$ and $\{N_t\}$ do not require knowledge of $\{\sigma^2(x)\}_{x\in X}$. Precisely, if a and N_0 are not tuned to $\sigma(x^*)^2$ for $x^*\in X^*$,

[14] The result in [9, Theorem 4.2] is for $\varphi\equiv 0$ and $X=\mathbb{R}^d$ so the bound $f(x^1)-f^*\leq\frac{L}{2}\|x^1-x^*\|^2$ holds by Lemma 3. For $\varphi\equiv 0$ and $X=\mathbb{R}^d$, we could obtain $\mathbb{E}\left[f(x^{t+1})-f^*\right]\lesssim\left[f(x^1)-f^*+\frac{\sigma(x^*)^2}{c}\right]\rho^{t-1}$. Also, their multiplicative constant is $\mathsf{C}=\max\{f(x^1)-f^*,\frac{2\sigma^2}{c}\}$ [see (4.23)]. Up to absolute constants, $f(x^1)-f^*+\frac{\sigma(x^*)^2}{c}\lesssim\mathsf{C}\lesssim f(x^1)-f^*+\frac{\sigma(x^*)^2}{c}$.

then the algorithm keeps running with proportional scaling in the convergence rate and oracle complexity. In this sense, the dynamic mini-batch scheme we propose is robust (see [27,42] for comments on robust methods).

Acknowledgements The authors thank the referees for improving the presentation of the paper.

Concluding remarks This work proposes stochastic approximation methods with dynamic sampling for stochastic optimization problems where the objective is the sum of a smooth convex function with a convex regularizer and where proximal evaluations are easy to compute. We require a significantly milder oracle assumption, namely an oracle with *multiplicative noise*, and show that our proposed methods converge under this assumption. By allowing oracles with multiplicative noise, we may include the solution of a wider class of realistic problems where the sampling procedure is aggressively corrupted. We believe that our work could have an impact in the analysis of problems found in robust statistics and statistical learning as well as problems with model misspecification.

In the set-up of multiplicative noise, the convergence analysis presented shows that our proposed methods converge with optimal rates and (near) optimal oracle complexity (in the precise sense discussed in Introduction).

Additionally, we show that our methods have a variance *localization* property as long as the oracle can be called multiple times per iteration and a proper sampling rate is used. One interesting direction of future research is to exploit acceleration with variance reduction for a strongly convex objective function.

Appendix: Proofs of Lemmas 1 and 4

Proof of Lemma 1 Let $x, y \in X$. Jensen's inequality and (12) imply

$$\|\nabla f(x) - \nabla f(y)\| \leq \mathbb{E}\left[\|\nabla F(x,\xi) - \nabla F(y,\xi)\|\right] \leq \mathbb{E}\left[\mathsf{L}(\xi)\right] \|x - y\| .$$

The first claim is proved using the above and $\mathbb{E}[\mathsf{L}(\xi)] \leq |\mathsf{L}(\xi)|_2$ by Hölder's inequality. Using this, we get

$$\begin{aligned}\sigma(x) &\leq \|\!|\nabla F(x,\xi) - \nabla F(y,\xi)|\!\|_2 + \|\!|\nabla F(y,\xi) - \nabla f(y)|\!\|_2 \\ &\quad + \|\!|\nabla f(x) - \nabla f(y)|\!\|_2 \\ &\leq |\mathsf{L}(\xi)\,\|x - y\||_2 + \sigma(y) + L\,\|x - y\| = \sigma(y) + 2L\,\|x - y\| ,\end{aligned}$$

where we used the triangle inequality for $\|\cdot\|$ and Minkowski's inequality for $|\cdot|_2$. □

Proof of Lemma 4 Let $x \in \mathbb{R}^d$. Since $\{\xi_j\}_{j=1}^N$ is i.i.d. and (1) holds, the sequence $\{\varepsilon_j\}_{j=1}^N$ defined by $\varepsilon_j := \frac{\nabla F(x,\xi_j) - \nabla f(x)}{N}$ is an i.i.d. sequence of random vectors with zero mean. As a consequence,[15] we have $\mathbb{E}\left[\left\|\sum_{j=1}^N \varepsilon_j(x)\right\|^2\right] = \sum_{j=1}^N \mathbb{E}\left[\|\varepsilon_j(x)\|^2\right]$.

[15] To show this, let $\varepsilon_i[\ell]$ denote the ℓ-th coordinate of the vector ε_i. From the Pythagorean identity and linearity of the expectation, we get $\mathbb{E}\left[\left\|\sum_{j=1}^N \varepsilon_j\right\|^2\right] = \sum_{j=1}^N \mathbb{E}\left[\|\varepsilon_j\|^2\right] + 2\sum_{i<j} \mathbb{E}\left[\langle \varepsilon_i, \varepsilon_j\rangle\right] = \sum_{j=1}^N \mathbb{E}\left[\|\varepsilon_j\|^2\right] + 2\sum_{i<j}\sum_{\ell=1}^d \mathbb{E}\left\{\varepsilon_i[\ell]\varepsilon_j[\ell]\right\}$. The claim follows immediately from $\mathbb{E}\left\{\varepsilon_i[\ell]\varepsilon_j[\ell]\right\} = \mathbb{E}\left\{\varepsilon_i[\ell]\right\}\mathbb{E}\left\{\varepsilon_j[\ell]\right\} = 0$ for $i < j$, since for every $\ell \in [d]$, $\varepsilon_i[\ell]$ and $\varepsilon_j[\ell]$ are centered independent real-valued random variables.

Thus we get

$$|||\varepsilon(x)|||_2 = \sqrt{\sum_{j=1}^{N} \mathbb{E}\left[\left\|\varepsilon_j(x)\right\|^2\right]} = \sqrt{\sum_{j=1}^{N} \frac{\sigma(x)^2}{N^2}} = \frac{\sigma(x)}{\sqrt{N}},$$

where in second equality we used that $\{\xi_j\}$ is drawn from $\mathbf{P}$. The claim follows immediately from the above and Assumption 2. □

References

1. Allen-Zhu, Z.: Katyusha: the first direct acceleration of stochastic gradient methods (2016). Preprint at arXiv:1603.05953
2. Agarwal, A., Barlett, P., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory **58**(5), 3235–3249 (2012)
3. Atchadé, Y.F., Fort, G., Moulines, E.: On perturbed proximal gradient algorithms. J. Mach. Learn. Res. **18**, 1–33 (2017)
4. Bach, F.: Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. J. Mach. Learn. Res. **15**, 595–627 (2014)
5. Bach, F., Moulines, E.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Conference Paper, Advances in Neural Information Processing Systems (NIPS) (2011)
6. Balamurugan, P., Bach, F.: Stochastic variance reduction methods for saddle-point problems. In: Conference Paper, Advances in Neural Information Processing Systems (NIPS) (2016)
7. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
8. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning (2016). Preprint at arXiv:1606.04838
9. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. Ser. B **134**(1), 127–155 (2012)
10. Chung, K.L.: On a stochastic approximation method. Ann. Math. Stat. **25**, 463–483 (1954)
11. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 1646–1654. Curran Associates, Inc., Red Hook (2014)
12. Dieuleveut, A., Flammarion, N., Bach, F.: Harder, better, faster stronger convergence rates for least-squares regression (2016). Preprint at arXiv:1602.05419
13. Duchi, J., Singer, Y.: Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. **10**, 2899–2934 (2009)
14. Dvoretzky, A.: On stochastic approximation. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 39–55. University of California Press (1956)
15. Flammarion, N., Bach, F.: Stochastic composite least-squares regression with convergence rate $O(1/n)$ (2017). Preprint at arXiv:1702.06429
16. Friedlander, M., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Sci. Comput. **34**(3), 1380–1405 (2012)
17. Frostig, R., Ge, R., Kakade, S.M., Sidford, A.: Competing with the empirical risk minimizer in a single pass. In: COLT 2015 Proceedings (2015)
18. Fu, M.C. (ed.): Handbook of Simulation Optimization. Springer, New York (2015)
19. Gadat, S., Panloup, F.: Optimal non-asymptotic bound of the Ruppert–Polyak averaging without strong convexity (2017). Preprint at arXiv:1709.03342
20. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. SIAM J. Optim. **22**(4), 1469–1492 (2012)
21. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)

22. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. Ser. A **156**(1), 59–99 (2016)
23. Gürbüzbalaban, M., Ozdaglar, A., Parrilo, P.: Convergence rate of incremental aggregated gradient algorithms. SIAM J. Optim. Preprint at arXiv:1506.02081
24. Hazan, E., Kale, S.: Beyond the regret minimization barrier optimal algorithms for stochastic strongly-convex optimization. J. Mach. Learn. Res. **15**, 2489–2512 (2014)
25. Hu, C., Kwok, J.T., Pan, W.: Accelerated gradient methods for stochastic optimization and online learning. In: Advances in Neural Information Processing Systems (NIPS) (2009)
26. Iusem, A., Jofré, A., Thompson, P.: Incremental constraint projection methods for monotone stochastic variational inequalities. Math. Oper. Res. Preprint at arXiv:1703.00272
27. Iusem, A., Jofré, A., Oliveira, R.I., Thompson, P.: Extragradient methods with variance reduction for stochastic variational inequalities. SIAM J. Optim. **27**(2), 686–724 (2017)
28. Iusem, A., Jofré, A., Oliveira, R.I., Thompson, P.: Variance-based stochastic extragradient methods with line search for stochastic variational inequalities (2016). Preprint at arXiv:1703.00262
29. Jain, P., Kakade, S.M., Kidambi, R., Netrapalli, P., Sidford, A.: Parallelizing stochastic approximation through mini-batching and tail-averaging (2016). Preprint at arXiv:1610.03774
30. Jain, P., Kakade, S.M., Kidambi, R., Netrapalli, P., Sidford, A.: Accelerating stochastic gradient descent (2017). Preprint at arXiv:1704.08227
31. Jiang, H., Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem. IEEE Trans. Autom. Control **53**(6), 1462–1475 (2008)
32. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems (NIPS) (2013)
33. Juditsky, A.B., Nazin, A.V., Tsybakov, A.B., Vayatis, N.: Recursive aggregation of estimators via the mirror descent algorithm with averaging. Probl. Inf. Transm. **41**(4), 368–384 (2005)
34. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst. **1**(1), 17–58 (2011)
35. Juditsky, A., Rigollet, P., Tsybakov, A.B.: Learning by mirror averaging. Ann. Stat. **36**(5), 2183–2206 (2008)
36. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. Ser. A **133**(1), 365–397 (2012)
37. Le Roux, N., Schmidt, M., Bach, F.R.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Advances in Neural Information Processing Systems 25 (NIPS) (2012)
38. Lee, S., Wright, S.: Manifold identification in dual averaging for regularized stochastic online learning. J. Mach. Learn. Res. **13**, 1705–1744 (2012)
39. Lin, H., Mairal, J., Harchaoui, Z.: A universal catalyst for first-order optimization. In: Advances in Neural Information Processing Systems (NIPS) (2015)
40. Lin, Q., Chen, X., Peña, J.: A sparsity preserving stochastic gradient methods for sparse regression. Comput. Optim. Appl. **58**(2), 455–482 (2014)
41. 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)
42. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
43. Nemirovskii, A., Yudin, D.: On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. (in Russian)—Doklady Akademii Nauk SSSR, 239, 5 (1978) (English translation: Soviet Math. Dokl. 19, 2 (1978))
44. Nemirovski, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics. Wiley, New York (1983)
45. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Soviet Math. Doklady **27**, 372–376 (1983)
46. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Cambridge (2004)
47. Nesterov, Y.: Gradient methods for minimizing composite objective function. Math. Program. Ser. B **140**(1), 125–161 (2013)
48. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. Ser. B **120**(1), 221–259 (2009)
49. Nesterov, Y.V.: Confidence level solutions for stochastic programming. Automatica **44**(6), 1559–1568 (2008)

50. Polyak, B.T.: New method of stochastic approximation type. Autom. Remote Control **51**, 937–946 (1991)
51. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**, 838–855 (1992)
52. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
53. Rosasco, L., Villa, S., Vũ, B.C.: Convergence of a stochastic proximal gradient algorithm (2014). Preprint at arXiv:1403.5074
54. Ruppert, D.: Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Cornell University Operations Research and Industrial Engineering (1988). Preprint at https://ecommons.cornell.edu/handle/1813/8664
55. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. Ser. A **162**(1), 83–112 (2017)
56. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2009)
57. Sra, S., Nowozin, S., Wright, S.J. (eds.): Optimization for Machine Learning. The MIT Press, Cambridge, MA (2012)
58. Tseng, P.: On Accelerated Proximal Gradient Methods for Convex–Concave Optimization. University of Washington, Seattle (2008)
59. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **9**, 2543–2596 (2010)
60. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)
61. Woodworth, B.E., Srebro, N.: Tight complexity bounds for optimizing composite objectives. In: Advances in Neural Information Processing Systems 29 (NIPS) (2016)
62. Zhang, L., Yang, T., Jin, R., He, X.: $O(logT)$ projections for stochastic optimization of smooth and strongly convex functions. In: Proceedings of the International Conference on International Conference on Machine Learning (ICML), Vol. 28 (2013)