Journal of Machine Learning Research 22 (2021) 1-49 Submitted 2/20; Revised 12/20; Published 4/21

L-SVRG and L-Katyusha with Arbitrary Sampling

Xun Qian XUN.QIAN@KAUST.EDU.SA
Division of Computer, Electrical and Mathematical Sciences, and Engineering
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Zheng Qu ZHENGQU@HKU.HK
Department of Mathematics
The University of Hong Kong
Hong Kong

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
Division of Computer, Electrical and Mathematical Sciences, and Engineering
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Editor: Simon Lacoste-Julien

Abstract

We develop and analyze a new family of *nonaccelerated and accelerated loopless variance-reduced methods* for finite-sum optimization problems. Our convergence analysis relies on a novel expected smoothness condition which upper bounds the variance of the stochastic gradient estimation by a constant times a distance-like function. This allows us to handle with ease *arbitrary sampling schemes* as well as the nonconvex case. We perform an in-depth estimation of these expected smoothness parameters and propose new importance samplings which allow *linear speedup* when the expected minibatch size is in a certain range. Furthermore, a connection between these expected smoothness parameters and expected separable overapproximation (ESO) is established, which allows us to exploit data sparsity as well. Our general methods and results recover as special cases the loopless SVRG (Hofmann et al., 2015) and loopless Katyusha (Kovalev et al., 2019) methods.

Keywords: L-SVRG, L-Katyusha, Arbitrary sampling, Expected smoothness, ESO

1. Introduction

In this work we consider the composite finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d} P(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x), \tag{1}$$

where $f \coloneqq \frac{1}{n}\sum_i f_i$ is an average of a very large number of smooth functions $f_i : \mathbb{R}^d \to \mathbb{R}$, and $\psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a proper closed convex function. We assume that problem (1) has at least one global optimal solution x^* and we denote by $P^* = P(x^*)$ the optimal value of problem (1).

Variance reduction. Variance reduced methods for solving (1) have recently become immensely popular and efficient alternatives of SGD (Nemirovski et al., 2009; Robbins and

Monro, 1951). Among the first such methods proposed were SAG (Schmidt et al., 2017), SAGA (Defazio et al., 2014), and SVRG (Johnson and Zhang, 2013; Xiao and Zhang, 2014), all with essentially identical theoretical complexity rates, but different practical use cases and different analysis techniques. While the first approaches to this were indirect and dual in nature (Shalev-Shwartz and Zhang, 2013), it later transpired that variance reduced methods can be accelerated, in the sense of Nesterov, directly. The first such method, Katyusha (Allen-Zhu, 2017)—an accelerated variant of SVRG—has become very popular due to its optimal complexity rate, versatility, and practical behavior. Both SVRG and Katyusha have a two-loop structure. In order for SVRG to obtain the best convergence rate, the inner loop must be terminated after a number of iterations proportional to the condition number of the problem. However, this is often unknown, or hard to estimate, and this has led practitioners to devise various heuristic strategies instead, departing from theory.

Loopless methods. This problem was remedied by the so-called *loopless* SVRG (L-SVRG) (Hofmann et al., 2015; Kovalev et al., 2019) and *loopless* Katyusha (L-Katyusha) (Kovalev et al., 2019). These methods dispense off the outer loop, replacing it with a biased coin-flip to be performed in each step. This simple change makes the methods easier to understand, and easier to analyze. The worst-case complexity bounds remain the same. Moreover, for L-SVRG the optimal probability of exit to the outer loop can be made independent of the condition number, which resolves the problem mentioned above, and makes the method more robust and markedly faster in practice. L-SVRG was analyzed in (Hofmann et al., 2015) and (Kovalev et al., 2019) for the strongly convex and smooth case ($\psi \equiv 0$); rates in the non-strongly convex and nonconvex case are not known.

Arbitrary sampling. The *arbitrary sampling* paradigm to developing and analyzing stochastic algorithms allows for a simultaneous study of countless importance and minibatch sampling strategies, thus leading to a tight unification of two previously separate topics. It was first proposed in (Richtárik and Takáč, 2016) in the context of randomized coordinate descent methods. Since then, many stochastic methods were studied in this regime. Methods already endowed with arbitrary sampling variants and analysis include, among others, the primal-dual method Quartz (Qu et al., 2015), accelerated randomized coordinate descent (Qu and Richtárik, 2016a,b; Hanzely and Richtárik, 2019), stochastic primal-dual hybrid gradient method (Chambolle et al., 2017), SGD (Gower et al., 2019), and SAGA (Qian et al., 2019). All these methods were studied in a convex or strongly convex setting only. In the nonconvex case, an arbitrary sampling analysis was performed only recently in (Horváth and Richtárik, 2019), for the SAGA, SVRG and SARAH methods, where an optimal sampling was developed.

1.1 Contributions

In this paper, we study L-SVRG and L-Katyusha with arbitrary sampling for the composite problem (1) in the case where f is convex, and L-SVRG with arbitrary sampling for the smooth problem ($\psi \equiv 0$) when f is nonconvex. We define two *expected smoothness* constants $\mathcal{L}_1$ and $\mathcal{L}_2$ (see Assumptions 5 and 6), which essentially help to upper bound the variance of the stochastic gradient estimation through the Bregman distance associated with f. With the aid of these two constants, the proof of L-SVRG and L-Katyusha can be completely detached from the sampling strategy as well as the convex and smooth properties of each

sample function f_i. We then reduce the algorithm parameter setting and complexity bound analysis for L-SVRG (resp. L-Katyusha) to the computation of the constant $\mathcal{L}_1$ (resp. $\mathcal{L}_2$). In the same spirit, we define in Assumption 7 a third expected smoothness constant $\mathcal{L}_3$, which plays a central role in the analysis of L-SVRG when f is nonconvex but smooth and $\psi \equiv 0$.

Our approach allows us to deal with many different cases in a unified way and also clarifies how sampling influences the stepsize choice and the overall complexity bound. We give computable upper bounds of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ under various scenarios, including the sum-of-convex, and nonconvex cases, and for arbitrary sampling strategies, see Section 6. It should be noticed that we estimate $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ under the smoothness (and convexity) of each f_i which is commonly used in stochastic optimization. It is possible to estimate these expected smoothness parameters under weaker conditions. However, the main goal of this paper is to study the influence of arbitrary sampling strategies. Hence, we leave this as future research. We now summarize a few important special cases covered by our results.

Strongly convex case. For L-SVRG, the iteration complexity is at least as good as that of SAGA-AS (Qian et al., 2019) and Quartz (Qu et al., 2015). Assume f is L_f-smooth and f_i is L_i-smooth. For the importance sampling, we can obtain linear speed up with respect to the expected minibatch size τ until $\tau = n$ or until the iteration complexity becomes $\mathcal{O}\left((n/\tau + L_f/\mu)\log\frac{1}{\epsilon}\right)$, where μ is the strongly convexity constant of P. For L-Katyusha, the iteration complexity is essentially the same with that of Katyusha (Allen-Zhu, 2017), and has linear speed up with respect to the expected minibatch size τ until $\tau = n$ or until the iteration complexity becomes $\mathcal{O}((n/\tau + \sqrt{L_f/\mu})\log 1/\epsilon)$. While in minibatch setting, Katyusha (Allen-Zhu, 2017) is only studied for the sampling with replacement. The estimation of $\mathcal{L}_2$ also gives the convergence result of Katyusha with arbitrary sampling. Furthermore, L-Katyusha is simpler and faster considering the running time in practice.

Nonconvex and smooth case. The first arbitrary sampling analysis in a nonconvex setting was performed in (Horváth and Richtárik, 2019). Our iteration complexity of L-SVRG with the importance sampling is at least as good as that of SAGA and SVRG with the optimal sampling in (Horváth and Richtárik, 2019), and could be better if L_f is smaller than $\bar{L} \coloneqq \sum_{i\in[n]} L_i/n$. Moreover, we can obtain linear speed up with respect to τ until $\tau = n$ or until the iteration complexity becomes $\mathcal{O}\left(L_f/\epsilon\right)$, while the results in (Horváth and Richtárik, 2019) holds for $\tau \leq \mathcal{O}(n^{2/3})$ only.

Sparsity All our convergence results rely on some expected smoothness parameters such that we can analyze the algorithms with arbitrary sampling and sampling with replacement in a unified framework. We establish the connection between these expected smoothness parameters and ESO (Richtárik and Takáč, 2016), which allows us the explore the sparsity of data as well.

1.2 Organization

In Section 2, we introduce the concepts of sampling and related notions. In Sections 3 and 4, we study L-SVRG and L-Katyusha in the strongly convex case and the non-strongly convex case, respectively. In Section 5, we study L-SVRG in the smooth and nonconvex case. In Section 6, we give the estimations of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, and propose the importance sampling. The numerical experiments are given in Section 7. We conclude this paper in Section 8. All

the missing proofs and the efficient implementations for L-SVRG, L-Katyusha, and Katyusha for sparse data can be found in the Appendix.

2. Preliminaries

2.1 Sampling

Let $F : \mathbb{R}^d \to \mathbb{R}^n$ be defined by $F(x) \coloneqq (f_1(x), \ldots, f_n(x))^\top$ and let

$$\mathbf{G}(x) \coloneqq [\nabla f_1(x), \ldots, \nabla f_n(x)] \in \mathbb{R}^{d\times n}$$

be the transpose of the Jacobian of F at point x. At each step of L-SVRG and L-Katyusha, a multiset S taking values from $[n] \coloneqq \{1, 2, \ldots, n\}$ is randomly generated and the vector

$$g = \frac{1}{n}\left(\mathbf{G}(x) - \mathbf{G}(w)\right)\Theta_S \mathbf{I}_S e + \frac{1}{n}\mathbf{G}(w)e, \tag{2}$$

is computed as a stochastic gradient estimator of $\nabla f(x)$. Here x is the current iterate and w is the current reference point. $\mathbf{I}_S \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix whose i-th diagonal entry is the number of copies of i in the multiset S and $\Theta_S \in \mathbb{R}^{n\times n}$ is a diagonal matrix associated with S such that (e is vector of all ones in $\mathbb{R}^n$)

$$\mathbb{E}[\Theta_S \mathbf{I}_S]e = e, \tag{3}$$

which ensures that g in (2) is an unbiased estimator of $\nabla f(x)$.

We next introduce some special samplings that we shall consider later and give examples of $\{\Theta_S\}$ satisfying (3).

Definition 1 (Sampling without replacement, i.e., set sampling) *A set sampling S is a random set-valued mapping with values being the subsets of $[n]$.*

A set sampling is uniquely characterized by the choice of probabilities $p_C \coloneqq \mathbb{P}[S = C]$ associated with every subset C of $[n]$. An example of set sampling for $n = 4$ can be generated by choosing $p_{\{1,2\}} = 0.3$, $p_{\{2,3,4\}} = 0.5$, $p_{\{1,4\}} = 0.2$. Given a sampling S, we let $p_i \coloneqq \mathbb{P}[i \in S] = \sum_{C:i\in C} p_C$. We say that S is *proper* if $p_i > 0$ for all i. We consider proper sampling only. S is a *uniform sampling* if $p_i = p_j$ for all $i, j \in [n]$. A *serial sampling* refers to the case when $|S| = 1$ with probability one. For an integer $\tau \in [n]$, a *τ-nice sampling* refers to the case when $|S| = \tau$ with probability one and each subset of size τ is selected with equal probability, see (Richtárik and Takáč, 2016).

Let Θ_S^i be the ith diagonal element of the matrix Θ_S. If S is a proper set sampling and

$$\Theta_S^i = p_i^{-1}, \ \ \forall i \in [n], \ \ S \subseteq [n],$$

then (3) holds.

We now introduce a new type of set sampling, called *group sampling*. As we shall see, group sampling will be useful to construct a set sampling S with prescribed $(p_1, \ldots, p_n) \in (0, 1]^n$.

Definition 2 (Group sampling) *Given $(p_1, \ldots, p_n) \in (0, 1]^n$ and a partition $\{C_j : j = 1, \ldots, t\}$ of $[n]$ such that $\sum_{i\in C_j} p_i \le 1$ for all $j \in \{1, \ldots, t\}$, a group sampling S is formed as follows.*

1. *For each $j = 1, \dots, t$, let $S_j = \{i\}$ with probability p_i for all $i \in C_j$ and let $S_j = \emptyset$ with probability $1 - \sum_{i \in C_j} p_i$.*
2. *$S = \cup_j S_j$.*

It is easy to see that $\mathbb{P}[i \in S] = (p_i / \sum_{s \in C_j} p_s) \cdot \sum_{s \in C_j} p_s = p_i$.

For example, for $n = 4$, given $(p_1, p_2, p_3, p_4) = (0.3, 0.6, 0.4, 0.6)$ and a partition $\{\{1, 2\}, \{3, 4\}\}$, we can get a group sampling with $p_{\{3\}} = 0.04$, $p_{\{4\}} = 0.06$, $p_{\{1,3\}} = 0.12$, $p_{\{1,4\}} = 0.18$, $p_{\{2,3\}} = 0.24$, and $p_{\{2,4\}} = 0.36$. Group sampling contains *independent sampling* as a special case. The latter is a goup sampling with $t = n$ groups so that each C_j contains only one element. It was studied in the arbitrary sampling paradigm (Hanzely and Richtárik, 2019; Horváth and Richtárik, 2019). A severe drawback of independent sampling is that the cost for each sample is $\mathcal{O}(n)$, whereas group sampling has the following nice property. The following property shows that when the expected cardinality $\mathbb{E}[|S|] = \sum_{C \subseteq [n]} |C| p_C = \sum_{C \subseteq [n]} \sum_{i \in C} p_C = \sum_{i \in [n]} \sum_{C : i \in C} p_C = \sum_{i \in [n]} p_i = \tau \geq 1$, there exists a group sampling such that the cost for generating each sample is $\mathcal{O}(\tau \log n)$.

Lemma 3 *For any set $\{p_i\}_{i=1}^n \subset (0, 1]$ such that $1 \leq \tau = \sum_{i \in [n]} p_i \leq n$, there exists a group sampling S such that $\mathbb{P}[i \in S] = p_i$ and the number of groups $t < 2\tau + 1$. If τ is an integer, then the number of groups can be reduced to $t \leq 2\tau - 1$.*

Apart from set sampling, we also consider a special multiset sampling where S is consisted of τ independent copies of a random integer in $[n]$, a.k.a. *sampling with replacement*.

Definition 4 (Sampling with replacement) *Let $\{\tilde{p}_i\}_{i=1}^n \subset (0, 1]$ satisfy $\sum_{i \in [n]} \tilde{p}_i = 1$. Let $\tau \in [n]$ and $s_1, \dots, s_\tau \in [n]$ be τ independent random integers with identical distribution so that s_1 equals to i with probability $\tilde{p}_i > 0$ for all $i \in [n]$. Then the multiset $S := \{s_1, \dots, s_\tau\}$ is a sampling with replacement of size τ with respect to the distribution vector $(\tilde{p}_1, \dots, \tilde{p}_n)$.*

A sampling with replacement S may contain multiple copies of a same index. For instance, for $n = 3$, given $(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3) = (0.2, 0.3, 0.5)$ and $\tau = 2$, we can get a multiset sampling with $p_{\{1,1\}} = 0.04$, $p_{\{2,2\}} = 0.09$, $p_{\{3,3\}} = 0.25$, $p_{\{1,2\}} = 0.12$, $p_{\{1,3\}} = 0.2$, and $p_{\{2,3\}} = 0.3$. In particular the diagonal matrix $\mathbf{I}_S$ may contain elements larger than 1. If S is a sampling with replacement of size τ with respect to the distribution vector $(\tilde{p}_1, \dots, \tilde{p}_n)$, then (3) holds for

$$\Theta_S^i = (\tau \tilde{p}_i)^{-1}, \ \ \forall i \in [n].$$

2.2 Assumptions

Throughout the paper, we always assume that (3) holds and make the following assumptions on f and ψ.

Assumption 1 *There are $L_f > 0$ and $\mu_f \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}^d$*

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu_f}{2} \|x - y\|^2 \leq f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_f}{2} \|x - y\|^2.$$

Assumption 2 *There is $\mu_\psi \geq 0$ such that for all $x, y \in \mathbb{R}^d$ and $v \in \partial \psi(y)$*

$$\psi(x) \geq \psi(y) + \langle v, x - y \rangle + \frac{\mu_\psi}{2} \|x - y\|^2.$$

We shall also need to assume that the following standard proximal operator of ψ

$$\text{prox}_{\eta\psi}(x) \coloneqq \arg\min_{y\in\mathbb{R}^d}\left\{\frac{1}{2}\|x-y\|^2+\eta\psi(y)\right\},\ \forall \eta>0, x\in\mathbb{R}^d,$$

is easily computable. Note that Assumption 1 implies

$$\|\nabla f(x)-\nabla f(y)\|\le \max\left(L_f,|\mu_f|\right)\|x-y\|,\ \ \forall x,y\in\mathbb{R}^d. \tag{4}$$

The proof of (4) can be found in Appendix B. If in addition f is convex, it's well known that (Nesterov, 2004, Theorem 2.1.5)

$$\|\nabla f(x)-\nabla f(y)\|^2\le 2L_f(f(x)-f(y)-\langle\nabla f(y),x-y\rangle),\ \ \forall x,y\in\mathbb{R}^d. \tag{5}$$

3. Strongly Convex Case

In this section, we develop loopless SVRG and loopless Katyusha for the composite problem (1). Throughout this section, we make the following assumptions on the functions f and ψ.

Assumption 3 *f is convex, i.e., $\mu_f\ge 0$.*

Assumption 4 *Either f or ψ is strongly convex, i.e., $\mu\coloneqq\mu_f+\mu_\psi>0$.*

It should be noticed that the results in this section do not require the convexity of each individual function f_i. Instead, we provide convergence guarantees under some expected smoothness assumptions.

3.1 Loopless SVRG (L-SVRG)

The loopless SVRG algorithm with arbitrary sampling is described in Algorithm 1. Loopless SVRG was first proposed in (Hofmann et al., 2015) for smooth problems ($\psi\equiv 0$) and in the serial and uniform sampling case. Algorithm 1 extends the work of (Hofmann et al., 2015) to the composite case and the arbitrary sampling regime. The same as in (Hofmann et al., 2015), we use a factor $p\in(0,1]$ to control the frequency of updating the reference point w^k. In contrast with (Hofmann et al., 2015) which picks up one sample each iteration, Algorithm 1 generates a sequence of i.i.d. random multisets $S_0,S_1,\ldots,$ and the stochastic gradient estimator g^k at step k is computed from $\{\nabla f_i: i\in S_k\}$. The common distribution of the random multisets is required as an input $\mathcal{S}$ of the algorithm. At each iteration, the expected number of gradient evaluations is bounded by $\mathcal{O}(\mathbb{E}_{S\sim\mathcal{S}}[|S|]+np)$. The parameter η determines the stepsize and is related to the expected smoothness constant $\mathcal{L}_1$ defined as follows.

Assumption 5 (Expected smoothness) *There is a constant $\mathcal{L}_1>0$ such that for any $x\in\mathbb{R}^d$ and $S\sim\mathcal{S}$*

$$\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x)-\mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2\right]\le 2\mathcal{L}_1\big(f(x)-f(x^*)-\langle\nabla f(x^*),x-x^*\rangle\big).$$

Algorithm 1 Loopless SVRG (L-SVRG)

Require: stepsize $\eta > 0$; probability $p \in (0, 1]$; multiset sampling distribution $\mathcal{S}$
Ensure: $x^0 = w^0 \in \mathbb{R}^d$
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Sample $S_k \sim \mathcal{S}$ independently from each other
3: $g^k = \frac{1}{n}\left(\mathbf{G}(x^k) - \mathbf{G}(w^k)\right)\Theta_{S_k}\mathbf{I}_{S_k}e + \frac{1}{n}\mathbf{G}(w^k)e$
4: $x^{k+1} = \mathrm{prox}_{\eta\psi}\left(x^k - \eta g^k\right)$
5: $w^{k+1} = \begin{cases} x^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}$
6: **end for**

When $\psi \equiv 0$, we have $\nabla f(x^*) = 0$ and Assumption 5 reduces to Assumption 2.1 in (Gower et al., 2019), which was an essential condition to obtain a general analysis of stochastic gradient descent for the nonconvex and smooth problem. Following (Gower et al., 2019), we say that the finite-sum function f is $\mathcal{L}_1$-smooth with respect to the sampling S if Assumption 5 holds. A closely related notion is the *expected separable overapproximation* (ESO) property, developed in the context of parallel coordinate descent methods with arbitrary sampling (Richtárik and Takáč, 2016; Qu and Richtárik, 2016a), see Section 6.4.

For each iteration k, denote by $\mathbb{E}_k[\cdot]$ the conditional expectation given w^k and x^k. By (3) it is clear that g^k in Algorithm 1 is an unbiased estimator of $\nabla f(x^k)$, i.e.,

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \ \ \forall k \geq 0. \tag{6}$$

For each iteration k, denote

$$\mathcal{D}^k \coloneqq \mathbb{E}_k\left[\left\|\frac{1}{n}\left(\mathbf{G}(w^k) - \mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2\right], \tag{7}$$

where $S \sim \mathcal{S}$ and S is independent of w^k and x^k.

Lemma 5 *Under Assumption 5, we have for all $k \geq 0$*

$$\mathbb{E}_k[\mathcal{D}^{k+1}] \leq (1-p)\mathcal{D}^k + 2p\mathcal{L}_1\left(f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle\right),$$
$$\mathbb{E}_k[\|g^k - \nabla f(x^*)\|^2] \leq 4\mathcal{L}_1(f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle) + 2\mathcal{D}^k.$$

To analyze the convergence of Algorithm 1, we shall consider the following stochastic Lyapunov function

$$\Psi^k \coloneqq \|x^k - x^*\|^2 + \frac{4\eta^2}{p(1+\eta\mu_\psi)}\mathcal{D}^k.$$

Theorem 6 (Linear convergence of L-SVRG) *Consider Algorithm 1. Under Assumptions 3, 4 and 5, if the stepsize η satisfies $\eta \leq \frac{1}{6\mathcal{L}_1}$, then*

$$\mathbb{E}_k\left[\Psi^{k+1}\right] \leq \left(1 - \frac{\eta\mu}{1+\eta\mu_\psi}\right)\|x^k - x^*\|^2 + \left(1 - \frac{p}{2}\right)\frac{4\eta^2}{p(1+\eta\mu_\psi)}\mathcal{D}^k, \ \ \forall k \geq 0.$$

In particular, if we choose $\eta = \frac{1}{6\mathcal{L}_1}$, *then*

$$\mathbb{E}\Big[\Psi^k\Big] \leq \left(1 - \min\left(\frac{\mu}{6\mathcal{L}_1 + \mu_\psi}, \frac{p}{2}\right)\right)^k \Psi^0, \ \ \forall k \geq 0. \tag{8}$$

Proof Since x^* is the solution of problem (1), we have $x^* = \text{prox}_{\eta\psi}(x^* - \eta\nabla f(x^*))$. Then

$$\begin{aligned}
\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &= \mathbb{E}_k\left[\|\text{prox}_{\eta\psi}(x^k - \eta g^k) - \text{prox}_{\eta\psi}(x^* - \eta\nabla f(x^*))\|^2\right] \\
&\leq \frac{1}{1+\eta\mu_\psi}\mathbb{E}_k\left[\|x^k - \eta g^k - (x^* - \eta\nabla f(x^*))\|^2\right] \\
&= \frac{1}{1+\eta\mu_\psi}\Big(\|x^k - x^*\|^2 - 2\eta\langle\nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle\Big) \\
&\quad + \frac{\eta^2}{1+\eta\mu_\psi}\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \\
&\leq \frac{(1-\eta\mu_f)}{1+\eta\mu_\psi}\|x^k - x^*\|^2 - \frac{2\eta}{1+\eta\mu_\psi}(f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle) \\
&\quad + \frac{\eta^2}{1+\eta\mu_\psi}\mathbb{E}_k[\|g^k - \nabla f(x^*)\|^2].
\end{aligned}$$

Here the first inequality is due to the contraction property of the proximal operator. The second equality follows from (6) and the last inequality uses Assumption 4. Hence, by Lemma 5 we have

$$\begin{aligned}
\mathbb{E}_k\left[\Psi^{k+1}\right] &\leq \frac{(1-\eta\mu_f)}{1+\eta\mu_\psi}\|x^k - x^*\|^2 + \frac{4\eta^2(1-p)}{p(1+\eta\mu_\psi)}\mathbb{E}_k[\mathcal{D}^k] + \frac{\eta^2}{1+\eta\mu_\psi}\mathbb{E}_k[\|g^k - \nabla f(x^*)\|^2] \\
&\quad - \frac{2\eta}{1+\eta\mu_\psi}(1 - 4\eta\mathcal{L}_1)(f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle) \\
&\leq \frac{(1-\eta\mu_f)}{1+\eta\mu_\psi}\|x^k - x^*\|^2 + \frac{4\eta^2(1-p/2)}{p(1+\eta\mu_\psi)}\mathbb{E}_k[\mathcal{D}^k] \\
&\quad - \frac{2\eta}{1+\eta\mu_\psi}(1 - 6\eta\mathcal{L}_1)(f(x^k) - f(x^*) - \langle\nabla f(x^*), x^k - x^*\rangle).
\end{aligned}$$

Now if the step size $\eta \leq \frac{1}{6\mathcal{L}_1}$, then by $\mu = \mu_f + \mu_\psi$, we obtain the desired inequality. ■

From (8), in order to gurantee $\mathbb{E}[\Psi^k] \leq \epsilon \cdot \Psi^0$, it suffices to let $\eta = \frac{1}{6\mathcal{L}_1}$ and

$$k \geq \mathcal{O}\left(\left(\frac{1}{p} + \frac{\mathcal{L}_1}{\mu}\right)\log\frac{1}{\epsilon}\right).$$

If the sampling $S \sim \mathcal{S}$ has expected size τ, then the expected iteration cost is $O(\tau + np)$ and the expected batch complexity is $\mathcal{O}\left(\left(n + \frac{\tau}{p} + \frac{\mathcal{L}_1(\tau+np)}{\mu}\right)\log\frac{1}{\epsilon}\right)$, which is $\mathcal{O}\left(\left(n + \frac{\mathcal{L}_1\tau}{\mu}\right)\log\frac{1}{\epsilon}\right)$ for any p between τ/n and $\mu/\mathcal{L}_1$. In the serial and uniform sampling case, i.e., when $\tau = 1$ and $p_i = 1/n$, Algorithm 1 and Theorem 6 recovers the loopless SVRG algorithm and convergence result given in (Hofmann et al., 2015) and (Kovalev et al., 2019), where can be found a detailed comparison with the original SVRG method.

3.2 Loopless Katyusha (L-Katyusha)

In this section, we present loopless Katyusha with arbitrary sampling in Algorithm 2, which covers the work of (Kovalev et al., 2019) when $\psi \equiv 0$ and $\mathcal{S}$ is a serial and uniform distribution. When $S \sim \mathcal{S}$ is sampling with replacement, Algorithm 2 is similar to the original Katyusha method (Allen-Zhu, 2017) but can be more efficient in practice. We defer a detailed comparison to Remark 14. As in many accelerated methods, Algorithm 2 operates

Algorithm 2 Loopless Katyusha (L-Katyusha)

Require: stepsize parameters $\eta > 0$, $L > 0$, $\sigma_1 \geq 0$, $\theta_1, \theta_2 \in (0,1)$; probability $p \in (0,1]$; multiset sampling distribution $\mathcal{S}$

Ensure: $y^0 = z^0 = w^0 \in \mathbb{R}^d$

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $x^k = \theta_1 z^k + \theta_2 w^k + (1 - \theta_1 - \theta_2) y^k$
3: Sample $S_k \sim \mathcal{S}$ independently from each other
4: $g^k = \frac{1}{n}\left(\mathbf{G}(x^k) - \mathbf{G}(w^k)\right)\Theta_{S_k}\mathbf{I}_{S_k} e + \frac{1}{n}\mathbf{G}(w^k)e$
5: $z^{k+1} = \operatorname{prox}_{\frac{\eta}{(1+\eta\sigma_1)L}\psi}\left(\frac{1}{1+\eta\sigma_1}(\eta\sigma_1 x^k + z^k - \frac{\eta}{L} g^k)\right)$
6: $y^{k+1} = x^k + \theta_1(z^{k+1} - z^k)$
7: $w^{k+1} = \begin{cases} y^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}$
8: **end for**

on three sequences $\{x^k\}_k, \{y^k\}_k, \{z^k\}_k$. At each iteration k, the reference point w^{k+1} is updated to y^k with probability p. The convergence of Algorithm 2 relies on the following assumption on the expected smoothness of f with respect to the input distribution $\mathcal{S}$.

Assumption 6 *There is a constant $\mathcal{L}_2 > 0$ such that for all $x, y \in \mathbb{R}^d$ and $S \sim \mathcal{S}$*

$$\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x) - \mathbf{G}(y)\right)\Theta_S \mathbf{I}_S e - \frac{1}{n}\left(\mathbf{G}(x) - \mathbf{G}(y)\right)e\right\|^2\right] \leq 2\mathcal{L}_2(f(x) - f(y) - \langle \nabla f(y), x - y\rangle).$$

In this section, $\mathbb{E}_k[\cdot]$ denotes the conditional expectation given (x^k, y^k, z^k, w^k). By (3) it is immediate that g^k in Algorithm 2 is an unbiased estimator of $\nabla f(x^k)$, i.e.,

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \quad \forall k \geq 0. \tag{9}$$

The following lemma is a direct consequence of Assumption 6.

Lemma 7 *Under Assumption 6, we have*

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] \leq 2\mathcal{L}_2(f(w^k) - f(x^k) - \langle \nabla f(x^k), w^k - x^k\rangle). \tag{10}$$

We shall also need the next Lemma which can be proved in the same way as Lemma 5.3 in (Kovalev et al., 2019).

Lemma 8 *[see Kovalev et al., 2019, Lemma 5.3] If $L \geq L_f$, then*

$$\frac{1}{\theta_1}(f(y^{k+1}) - f(x^k)) - \frac{1}{4L\theta_1}\|g^k - \nabla f(x^k)\|^2 \leq \frac{L}{2\eta}\|z^{k+1} - z^k\|^2 + \langle g^k, z^{k+1} - z^k\rangle. \tag{11}$$

For any $q \in (0,1)$, we define the stochastic Lyapunov function

$$\Phi^k := \mathcal{Z}^k + \mathcal{Y}^k + \mathcal{W}^k,$$

where

$$\mathcal{Z}^k = \frac{L+\eta\mu}{2\eta}\|z^k - x^*\|^2, \qquad \mathcal{Y}^k = \frac{1}{\theta_1}(P(y^k) - P^*), \qquad \mathcal{W}^k = \frac{\theta_2}{pq\theta_1}(P(w^k) - P^*).$$

It should be noticed that the definition of $\mathcal{Z}^k$ is the same as that of (Kovalev et al., 2019), but that of $\mathcal{Y}^k$ and $\mathcal{W}^k$ are different. It is easy to check that

$$\mathbb{E}_k[\mathcal{W}^{k+1}] = (1-p)\mathcal{W}^k + \frac{\theta_2}{q}\mathcal{Y}^k, \ \ \forall k \geq 0. \tag{12}$$

Lemma 9 *If $\sigma_1 = \mu_f/L$, then we have*

$$\langle g^k, x^* - z^{k+1}\rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 \geq \frac{L}{2\eta}\|z^k - z^{k+1}\|^2 + \mathcal{Z}^{k+1} - \frac{L\mathcal{Z}^k}{L+\eta\mu} + \psi(z^{k+1}) - \psi(x^*). \tag{13}$$

Theorem 10 (Accelerated linear convergence of L-Katyusha) *Consider Algorithm 2. Under Assumptions 3, 4 and 6, if the stepsize parameters are set as follows*

$$L = \max\left(\mathcal{L}_2, L_f\right), \ \sigma_1 = \frac{\mu_f}{L}, \ \theta_2 = \frac{\mathcal{L}_2}{2L}, \ \theta_1 = \begin{cases} \min\left(\sqrt{\frac{\mu}{\mathcal{L}_{2p}}}\theta_2, \theta_2\right) & \text{if } L_f \leq \frac{\mathcal{L}_2}{p} \\ \min\left(\sqrt{\frac{\mu}{L_f}}, \frac{p}{2}\right) & \text{otherwise} \end{cases}, \ \eta = \frac{1}{3\theta_1}, \tag{14}$$

then we have

$$\mathbb{E}\left[\Phi^k\right] \leq \left(1 - \min\left(\frac{\mu}{\mu + 3\theta_1 L}, \theta_1 + \theta_2 - \frac{\theta_2}{q}, p(1-q)\right)\right)^k \Phi^0, \ \ \forall k \geq 0. \tag{15}$$

Proof It is easy to check that we always have $\theta_1 + \theta_2 \leq 1$. By the μ_f-strong convexity of f, we have

$$\begin{aligned}
f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k\rangle + \frac{\mu_f}{2}\|x^k - x^*\|^2 \\
&= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k + z^k - x^k\rangle \\
&= f(x^k) + \frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle \nabla f(x^k), x^* - z^k\rangle + \frac{\theta_2}{\theta_1}\langle \nabla f(x^k), x^k - w^k\rangle \\
&\quad + \frac{1-\theta_1-\theta_2}{\theta_1}\langle \nabla f(x^k), x^k - y^k\rangle \\
&\geq f(x^k) + \frac{\theta_2}{\theta_1}\langle \nabla f(x^k), x^k - w^k\rangle + \frac{1-\theta_1-\theta_2}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathbb{E}_k\left[\frac{\mu_f}{2}\|x^k - x^*\|^2 + \langle g^k, x^* - z^{k+1}\rangle + \langle g^k, z^{k+1} - z^k\rangle\right],
\end{aligned}$$

where the last inequality follows from the convexity of f and (9). For the last term in the above inequality, we have

$$\begin{aligned}
& \mathbb{E}_k\left[\frac{\mu_f}{2}\|x^k-x^*\|^2+\langle g^k,x^*-z^{k+1}\rangle+\langle g^k,z^{k+1}-z^k\rangle-\psi(z^{k+1})+\psi(x^*)-\mathcal{Z}^{k+1}\right] \\
\overset{(13)}{\geq} & -\frac{L\mathcal{Z}^k}{L+\eta\mu}+\mathbb{E}_k\left[\langle g^k,z^{k+1}-z^k\rangle+\frac{L}{2\eta}\|z^k-z^{k+1}\|^2\right] \\
\overset{(11)}{\geq} & -\frac{L\mathcal{Z}^k}{L+\eta\mu}+\mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1})-f(x^k))-\frac{1}{4L\theta_1}\|g^k-\nabla f(x^k)\|^2\right] \\
\overset{(10)}{\geq} & -\frac{L\mathcal{Z}^k}{L+\eta\mu}+\mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1})-f(x^k))-\frac{\mathcal{L}_2}{2L\theta_1}(f(w^k)-f(x^k)-\langle\nabla f(x^k),w^k-x^k\rangle)\right] \\
= & -\frac{L\mathcal{Z}^k}{L+\eta\mu}+\mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1})-f(x^k))-\frac{\theta_2}{\theta_1}(f(w^k)-f(x^k)-\langle\nabla f(x^k),w^k-x^k\rangle)\right].
\end{aligned}$$

Therefore

$$\begin{aligned}
& \mathbb{E}_k\left[f(x^*)-\psi(z^{k+1})+\psi(x^*)-\mathcal{Z}^{k+1}\right] \\
\geq & f(x^k)+\frac{1-\theta_1-\theta_2}{\theta_1}(f(x^k)-f(y^k))-\frac{L\mathcal{Z}^k}{L+\eta\mu} \\
& +\mathbb{E}_k\left[\frac{1}{\theta_1}(f(y^{k+1})-f(x^k))\right]-\frac{\theta_2}{\theta_1}(f(w^k)-f(x^k)) \\
= & -\frac{L\mathcal{Z}^k}{L+\eta\mu}-\frac{1-\theta_1-\theta_2}{\theta_1}f(y^k)+\frac{1}{\theta_1}\mathbb{E}_k\left[f(y^{k+1})\right]-\frac{\theta_2}{\theta_1}f(w^k).
\end{aligned}$$

Moreover, since ψ is convex and

$$y^{k+1}=x^k+\theta_1(z^{k+1}-z^k)=\theta_1 z^{k+1}+\theta_2 w^k+(1-\theta_1-\theta_2)y^k,$$

we have

$$\psi(z^{k+1})\geq\frac{1}{\theta_1}\psi(y^{k+1})-\frac{\theta_2}{\theta_1}\psi(w^k)-\frac{1-\theta_1-\theta_2}{\theta_1}\psi(y^k).$$

Hence, we arrive at

$$f(x^*)\geq\mathbb{E}_k[\mathcal{Z}^{k+1}]-\frac{L\mathcal{Z}^k}{L+\eta\mu}-\frac{1-\theta_1-\theta_2}{\theta_1}P(y^k)+\frac{1}{\theta_1}\mathbb{E}_k\left[P(y^{k+1})\right]-\frac{\theta_2}{\theta_1}P(w^k)-\psi(x^*).$$

After rearranging, we get $\mathbb{E}_k\left[\mathcal{Z}^{k+1}+\mathcal{Y}^{k+1}\right]\leq\frac{L\mathcal{Z}^k}{L+\eta\mu}+(1-\theta_1-\theta_2)\mathcal{Y}^k+pq\mathcal{W}^k$. In view of (12), we deduce that

$$\begin{aligned}
& \mathbb{E}_k\left[\mathcal{Z}^{k+1}+\mathcal{Y}^{k+1}+\mathcal{W}^{k+1}\right]\leq\frac{L\mathcal{Z}^k}{L+\eta\mu}+(1-\theta_1-\theta_2)\mathcal{Y}^k+pq\mathcal{W}^k+(1-p)\mathcal{W}^k+\frac{\theta_2}{q}\mathcal{Y}^k \\
= & \left(1-\frac{\eta\mu}{L+\eta\mu}\right)\mathcal{Z}^k+\left(1-\left(\theta_1+\theta_2-\frac{\theta_2}{q}\right)\right)\mathcal{Y}^k+(1-p(1-q))\,\mathcal{W}^k.
\end{aligned}$$

Finally we note that with $\eta=1/(3\theta_1)$ we have $\eta\mu/(L+\eta\mu)=\mu/(\mu+3\theta_1 L)$. ■

It should be noticed that $q\in(0,1)$ has no impact on the parameter choice of Algorithm 2.

However, it defines the Lyapunov function Φ^k and thus appears in the linear convergence rate in (15). We further study the rate in four different cases with specific values of q and obtain the following results.

Corollary 11 *Under the premise of Theorem 10, we have*

$$\mathbb{E}\Big[\Phi^k\Big] \leq \begin{cases} \left(1-\frac{p}{4}\right)^k \Phi^0 & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} \geq p \text{ and } q = \frac{2}{3} \\ \left(1-\sqrt{\frac{\mu p}{16\mathcal{L}_2}}\right)^k \Phi^0 & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} < p \text{ and } q = 1-\sqrt{\frac{\mu}{9\mathcal{L}_2 p}} \\ \left(1-\frac{p}{7}\right)^k \Phi^0 & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} \geq \frac{p^2}{4} \text{ and } q = \frac{2}{3} \\ \left(1-\sqrt{\frac{\mu}{16L_f}}\right)^k \Phi^0 & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} < \frac{p^2}{4} \text{ and } q = 1-\sqrt{\frac{4\mu}{9\mathcal{L}_2 p^2}}. \end{cases} \tag{16}$$

Therefore, with some $q \in [\frac{2}{3}, 1)$, $\mathbb{E}[\Phi^k] \leq \epsilon\Phi^0$ *for* $k \geq \mathcal{O}\left(\left(\frac{1}{p}+\sqrt{\frac{L_f}{\mu}}+\sqrt{\frac{\mathcal{L}_2}{\mu p}}\right)\log\frac{1}{\epsilon}\right)$.

Corollary 12 *Under the premise of Theorem 10, we have*

$$\mathbb{E}\Big[P(w^k)-P^*\Big] \leq \begin{cases} 6\left(1-\frac{p}{4}\right)^k \left(P(x^0)-P^*\right) & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} \geq p \\ 6\left(1-\sqrt{\frac{\mu p}{16\mathcal{L}_2}}\right)^k \left(P(x^0)-P^*\right) & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} < p \\ \frac{12pL_f}{\mathcal{L}_2}\left(1-\frac{p}{7}\right)^k \left(P(x^0)-P^*\right) & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} \geq \frac{p^2}{4} \\ \frac{12pL_f}{\mathcal{L}_2}\left(1-\sqrt{\frac{\mu}{16L_f}}\right)^k \left(P(x^0)-P^*\right) & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} < \frac{p^2}{4}. \end{cases} \tag{17}$$

Corollary 13 *Define*

$$T(\mu, \mathcal{L}_2, L_f, p) := \begin{cases} \frac{13}{p} & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} \geq p \\ 13\sqrt{\frac{\mathcal{L}_2}{\mu p}} & \text{if } L_f \leq \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_2} < p \\ \frac{7}{p}\ln(\frac{48pL_f}{\mathcal{L}_2}) & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} \geq \frac{p^2}{4} \\ 4\ln(\frac{48pL_f}{\mathcal{L}_2})\sqrt{\frac{L_f}{\mu}} & \text{if } L_f > \frac{\mathcal{L}_2}{p},\ \ \frac{\mu}{\mathcal{L}_f} < \frac{p^2}{4}. \end{cases} \tag{18}$$

Under the premise of Theorem 10, we have

$$\mathbb{E}\Big[P(w^k)-P^*\Big] \leq \frac{1}{4}\left(P(x^0)-P^*\right),\ \ \forall k \geq T(\mu, \mathcal{L}_2, L_f, p).$$

Remark 14 *There are two major differences between L-Katyusha (Algorithm 2) and the original Katyusha algorithm (Allen-Zhu, 2017).*

1. We here consider both arbitrary sampling and sampling with replacement while Katyusha (Allen-Zhu, 2017) only considered the second case. Note however that our Assumption 6 allows us to easily extend the original Katyusha method (Allen-Zhu, 2017) into arbitrary sampling schemes as well, by simply replacing everywhere the $\bar{L}/b$ *in their proof by the constant* $\mathcal{L}_2$*. This also yields a direct extension of Katyusha when each* f_i *is not necessarily convex.*

2. Our method is loopless and the reference point w^k *is set to be* y^{k-1} *with probability* p*. Recall that in the original Katyusha method, the reference point* $\tilde{x}^s$ *for each outer loop*

s is set to be a weighted average of past iterates of y^k. Not only does this difference bring a simplified algorithm and proof, but also a non-negligible practical convergence speed up. Indeed, the number of epochs of the two methods are essentially the same (see Section 6), but the computation overhead caused by the calculation of $\tilde{x}^s$ makes Katyusha slower than our loopless variant, especially in the case when a sparse implementation is needed. We provide in Appendix A further details. In Section 7 we show through numerical evidence the better convergence speed of our loopless variant.

4. Non-Strongly Convex Case

In this section, we consider the non-strongly convex case. In the first subsection, we show an ergodic sublinear convergence rate $\mathcal{O}(1/k)$ of L-SVRG. In the second subsection, we show how to combine L-Katyusha with the black-box reduction technique proposed in (Allen-Zhu and Hazan, 2016) to obtain an improved sublinear convergence rate.

4.1 Sublinear Convergence of L-SVRG

Consider the Lyapunov function

$$\Xi^k := \frac{1}{2\eta}\|x^k - x^*\|^2 + \frac{6\eta}{5p}\mathcal{D}^k,$$

with $\mathcal{D}^k$ defined in (7). The following Lemma can be proved similarly as the second formula in Lemma 5.

Lemma 15 *Under Assumption 5, we have for all $k \geq 0$*

$$\mathbb{E}_k[\|g^k - \nabla f(x^k)\|^2] \leq 4\mathcal{L}_1(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) + 2\mathcal{D}^k.$$

We shall also need the following well-known inequality.

Lemma 16 *[see e.g. Tseng, 2008] We have*

$$\langle g^k, x^* - x^{k+1}\rangle \geq \psi(x^{k+1}) - \psi(x^*) + \frac{1}{2\eta}\|x^k - x^{k+1}\|^2 + \frac{1}{2\eta}\|x^{k+1} - x^*\|^2 - \frac{1}{2\eta}\|x^k - x^*\|^2.$$

Since x^* is an optimal solution, we have $-\nabla f(x^*) \in \partial\psi(x^*)$, which along with the convexity of ψ implies that

$$f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle \leq P(x^k) - P^*. \tag{19}$$

Theorem 17 (Sublinear convergence of L-SVRG) *Consider Algorithm 1. Under Assumptions 3 and 5, if $\eta \leq \min\left(\frac{1}{8\mathcal{L}_1}, \frac{1}{6L_f}\right)$, then*

$$\mathbb{E}\left[P(x^{k+1}) - P^*\right] - \frac{3}{5}\mathbb{E}\left[P(x^k) - P^*\right] \leq \mathbb{E}\left[\Xi^k\right] - \mathbb{E}\left[\Xi^{k+1}\right], \ \forall k \geq 0. \tag{20}$$

Let $\tilde{x}^k = (x^0 + \cdots + x^k)/(k+1)$. Then

$$\mathbb{E}\left[P(\tilde{x}^k) - P^*\right] \leq \frac{1}{k+1}\left(\frac{5}{4\eta}\|x^0 - x^*\|^2 + \left(\frac{5}{2} + \frac{6\eta\mathcal{L}_1}{p}\right)\left(P(x^0) - P^*\right)\right).$$

This implies that $\mathbb{E}[P(\tilde{x}^k) - P^*] \leq \epsilon$ as long as $k \geq \mathcal{O}\left(\left(\frac{1}{p} + \mathcal{L}_1 + L_f\right)\frac{1}{\epsilon}\right)$.

Proof Let $\beta = \frac{5}{6\eta}$. Since f is convex and $\mathbb{E}_k[g^k] = \nabla f(x^k)$, we have

$$\begin{aligned} f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \\ &= f(x^k) + \mathbb{E}_k\left[\langle g^k, x^* - x^{k+1}\rangle\right] + \mathbb{E}_k[\langle g^k - \nabla f(x^k), x^{k+1} - x^k\rangle] \\ &\quad + \mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k\rangle] \\ &\geq \mathbb{E}_k\left[f(x^{k+1})\right] - \frac{L_f}{2}\mathbb{E}_k\left[\|x^{k+1} - x^k\|^2\right] + \mathbb{E}_k\left[\langle g^k, x^* - x^{k+1}\rangle\right] \\ &\quad + \mathbb{E}_k[\langle g^k - \nabla f(x^k), x^{k+1} - x^k\rangle] \\ &\geq \mathbb{E}_k\left[f(x^{k+1})\right] - \frac{L_f}{2}\mathbb{E}_k\left[\|x^{k+1} - x^k\|^2\right] + \mathbb{E}_k\left[\langle g^k, x^* - x^{k+1}\rangle\right] \\ &\quad - \frac{1}{2\beta}\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] - \frac{\beta}{2}\mathbb{E}_k\left[x^{k+1} - x^k\|^2\right], \end{aligned}$$

where the second inequality comes from that f is L_f-smooth and the third inequality comes from Young's inequality. Moreover, from Lemmas 15 and 16, we can obtain

$$\begin{aligned} P^* \geq\ & \mathbb{E}_k\left[P(x^{k+1})\right] + \left(\frac{1}{2\eta} - \frac{\beta}{2} - \frac{L_f}{2}\right)\mathbb{E}_k\left[\|x^{k+1} - x^k\|^2\right] + \frac{1}{2\eta}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] \\ & - \frac{1}{2\eta}\|x^k - x^*\|^2 - \frac{2\mathcal{L}_1}{\beta}(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) - \frac{1}{\beta}\mathcal{D}^k. \end{aligned}$$

Since $\beta = \frac{5}{6\eta}$ and $\eta \leq \frac{1}{6L_f}$, we have $\frac{1}{2\eta} - \frac{\beta}{2} - \frac{L_f}{2} \geq 0$. Therefore

$$\begin{aligned} & \frac{1}{2\eta}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] + \mathbb{E}_k\left[P(x^{k+1}) - P^*\right] \\ \leq\ & \frac{1}{2\eta}\|x^k - x^*\|^2 + \frac{2\mathcal{L}_1}{\beta}(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) + \frac{1}{\beta}\mathcal{D}^k. \end{aligned}$$

Let $\alpha = \frac{6\eta}{5p}$. In view of Lemma 5, we have

$$\begin{aligned} & \mathbb{E}_k\left[\Xi^{k+1}\right] + \mathbb{E}_k\left[P(x^{k+1}) - P^*\right] \\ \leq\ & \frac{1}{2\eta}\|x^k - x^*\|^2 + \left(\frac{2\mathcal{L}_1}{\beta} + 2\alpha p\mathcal{L}_1\right)(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) \\ & + \frac{1}{\beta}\mathcal{D}^k + \alpha(1-p)\mathcal{D}^k \\ =\ & \Xi^k + \frac{4\mathcal{L}_1}{\beta}(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) \\ \leq\ & \Xi^k + \frac{3}{5}(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle) \overset{(19)}{\leq} \Xi^k + \frac{3}{5}\left(P(x^k) - P^*\right). \end{aligned}$$

Taking expectation on both sides we obtain (20). Summing up (20) from iteration 0 to iteration k we obtain

$$\mathbb{E}\Big[P(x^k) - P^*\Big] + \frac{2}{5}\sum_{i=0}^{k-1}\mathbb{E}\big[P(x^i) - P^*\big] \leq P(x^0) - P^* + \Xi^0 - \mathbb{E}\Big[\Xi^k\Big], \tag{21}$$

By Assumption 5, we have

$$\Xi^0 \quad \leq \quad \frac{1}{2\eta}\|x^0 - x^*\|^2 + 2\alpha\mathcal{L}_1(f(x^0) - f(x^*) - \langle \nabla f(x^*), x^0 - x^*\rangle),$$

which with (21) yields

$$\frac{2}{5}\sum_{i=0}^{k}\mathbb{E}[P(x^i) - P^*] \leq \frac{1}{2\eta}\|x^0 - x^*\|^2 + \left(1 + \frac{12\eta\mathcal{L}_1}{5p}\right)\left(P(x^0) - P^*\right).$$

This along with the convexity of P implies the result. ■

Theorem 17 can be compared with Theorem 3 in (Shang et al., 2018). However, note that the reference point in our loopless SVRG is simply chosen to be x^k without the need for an extra averaging step. As discussed in Remark 14, the loopless variant has both simpler implementation and faster convergence speed.

4.2 Accelerated Sublinear Convergence

We propose to apply the black-box oracle in (Allen-Zhu and Hazan, 2016) in order to obtain an accelerated sublinear rate in the non-strongly convex case. We recall this oracle in Algorithm 3. The idea is to adaptively add a strongly convex term on the non-strongly

Algorithm 3 AdaptReg($\mathcal{A}$) (Allen-Zhu and Hazan, 2016)

Require: $\mu_0 > 0$; $x^0 \in \mathbb{R}^d$
1: $\hat{x}^0 = x^0$
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Define $P^{(\mu_t)} := P(x) + \frac{\mu_t}{2}\|x - x^0\|^2$
4: $\hat{x}^{t+1} = \mathcal{A}(P^{(\mu_t)}, \hat{x}_t, \frac{1}{4})$
5: $\mu_{t+1} = \mu_t/2$
6: **end for**

convex function P and minimize approximately the strongly convex function $P^{(\mu_t)}$:

$$\min_x \left[P^{(\mu_t)}(x) \equiv P(x) + \frac{\mu_t}{2}\|x - x^0\|^2\right]. \tag{22}$$

The fourth lines $\hat{x}^{t+1} = \mathcal{A}(P^{(\mu_t)}, \hat{x}_t, \frac{1}{4})$ in Algorithm 3 refers to applying algorithm $\mathcal{A}$ to solve (22) so that

$$\mathbb{E}\Big[P^{(\mu_t)}(\hat{x}^{t+1}) \;\Big|\; \hat{x}^t\Big] - \min_x P^{(\mu_t)}(x) \leq \frac{1}{4}\left(P^{(\mu_t)}(\hat{x}^t) - \min_x P^{(\mu_t)}(x)\right). \tag{23}$$

The following property can be found in (Allen-Zhu and Hazan, 2016).

Proposition 18 *[Allen-Zhu and Hazan, 2016] Consider Algorithm 3. If P is convex, then*

$$\mathbb{E}\big[P(\hat{x}^t) - P^*\big] \leq \frac{1}{4^t}\left(P(x^0) - P^*\right) + \frac{9\mu_0}{2^{t+1}}\|x^0 - x^*\|^2, \;\; \forall t \geq 0.$$

According to Corollary 13, if $\mathcal{A}$ is L-Katyusha, then (23) can be achieved within

$$k_t = T(\mu_t, \mathcal{L}_2, L_f, p)$$

iterations where $T : \mathbb{R}^4 \to \mathbb{R}_{>0}$ is the function defined as in (18) and p is the reference point updating probability in Algorithm 2. For completeness we give in Algorithm 4 the full description of Algorithm 3 with $\mathcal{A}$ being L-Katyusha (Algorithm 2).

Algorithm 4 AdaptReg-L-Katyusha

Require: $\mu_0 > 0$; $\hat{x}^0 = x^0 \in \mathbb{R}^d$; $\mathcal{L}_2 > 0$; probability $p \in (0, 1]$; multiset sampling distribution $\mathcal{S}$

$L = \max(\mathcal{L}_2, L_f)$, $\theta_2 = \frac{\mathcal{L}_2}{2L}$
for $t = 0, 1, 2, \ldots$ **do**
  if $L_f p \le \mathcal{L}_2$ **then**
    $\theta_1 = \min\left(\sqrt{\mu_t/(\mathcal{L}_2 p)}\theta_2, \theta_2\right)$
  else
    $\theta_1 = \min\left(\sqrt{\mu_t/L_f}, p/2\right)$
  end if
  $\eta_t = 1/(3\theta_1)$, $k_t = T(\mu_t, \mathcal{L}_2, L_f, p)$
  $y_t^0 = z_t^0 = w_t^0 = \hat{x}^t$
  for $k = 0, 1, \ldots, k_t - 1$ **do**
    $x_t^k = \theta_1 z_t^k + \theta_2 w_t^k + (1 - \theta_1 - \theta_2) y_t^k$
    Sample $S_k \sim \mathcal{S}$ independently from each other
    $g_t^k = \frac{1}{n}\left(\mathbf{G}(x_t^k) - \mathbf{G}(w_t^k)\right)\Theta_{S_k}\mathbf{I}_{S_k} e + \frac{1}{n}\mathbf{G}(w_t^k)e$
    $z_t^{k+1} = \mathrm{prox}_{\frac{\eta_t}{L+\eta_t\mu_t}\psi}\left(\frac{1}{L+\eta_t\mu_t}(\eta_t\mu_t x^0 + L z_t^k - \eta_t g_t^k)\right)$
    $y_t^{k+1} = x_t^k + \theta_1(z_t^{k+1} - z_t^k)$
    $w_t^{k+1} = \begin{cases} y_t^k & \text{with probability } p \\ w_t^k & \text{with probability } 1-p \end{cases}$
  end for
  $\hat{x}^{t+1} = w_t^{k_t}$, $\mu_{t+1} = \mu_t/2$
  $s = s + k_t$, $\bar{x}^s = \hat{x}^{t+1}$
end for

Theorem 19 (Accelerated sublinear convergence) *Consider Algorithm 4. Under Assumptions 3 and 6, we have*

$$\mathbb{E}[P(\bar{x}^s) - P^*] \le \frac{C_0^4}{(s - T_0)^4}\left(P(x^0) - P^*\right) + \frac{9\mu_0 C_0^2}{2(s - T_0)^2}\|x^0 - x^*\|^2.$$

Here C_0 and T_0 are two positive constants defined as follows.

$$T_0 = \begin{cases} \frac{13}{p}\lceil \max(\log_2 \frac{\mu_0}{\mathcal{L}_2 p}, 0)\rceil & \text{if } L_f \le \frac{\mathcal{L}_2}{p} \\ \frac{7}{p}\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right)\lceil \max(\log_2 \frac{4\mu_0}{\mathcal{L}_f p^2}, 0)\rceil & \text{otherwise.} \end{cases} \tag{24}$$

$$C_0 = \begin{cases} 32\sqrt{\frac{\mathcal{L}_2}{\mu_0 p}} & \text{if } L_f \le \frac{\mathcal{L}_2}{p} \\ 10\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right)\sqrt{\frac{\mathcal{L}_f}{\mu_0}} & \text{otherwise.} \end{cases} \tag{25}$$

Proof Case 1. Suppose $L_f \leq \frac{\mathcal{L}_2}{p}$. Let $\bar{t} \coloneqq \lceil \max(\log_2 \frac{\mu_0}{\mathcal{L}_2 p}, 0) \rceil \geq 0$. Then

$$\sum_{i=0}^{t} k_i = \sum_{i=0}^{\bar{t}-1} \frac{13}{p} + \sum_{i=\bar{t}}^{t} 13\sqrt{\frac{\mathcal{L}_2}{\mu_i p}} \leq \frac{13}{p}\left\lceil \max\left(\log_2 \frac{\mu_0}{\mathcal{L}_2 p}, 0\right)\right\rceil + 32\sqrt{\frac{2^{t+1}\mathcal{L}_2}{\mu_0 p}}, \ \ \forall t \geq 0.$$

Case 2. Suppose $L_f > \frac{\mathcal{L}_2}{p}$. Let $\bar{t} \coloneqq \lceil \max(\log_2 \frac{4\mu_0}{\mathcal{L}_f p^2}, 0) \rceil \geq 0$. Then

$$\begin{aligned}\sum_{i=0}^{t} k_i &= \sum_{i=0}^{\bar{t}-1} \frac{7}{p}\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right) + \sum_{i=\bar{t}}^{t} 4\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right)\sqrt{\frac{L_f}{\mu_i}} \\ &\leq \frac{7}{p}\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right)\left\lceil \max\left(\log_2 \frac{4\mu_0}{\mathcal{L}_f p^2}, 0\right)\right\rceil + 10\ln\left(\frac{48pL_f}{\mathcal{L}_2}\right)\sqrt{\frac{2^{t+1}\mathcal{L}_f}{\mu_0}}, \ \ \forall t \geq 0.\end{aligned}$$

Hence

$$\frac{1}{2^{t+1}} \leq \frac{C_0^2}{(\sum_{i=0}^{t} k_i - T_0)^2},$$

with T_0 and C_0 defined in (24) and (25). According to Proposition 18, we know that

$$\begin{aligned}\mathbb{E}\left[P(\hat{x}^{t+1}) - P^*\right] &\leq \frac{1}{4^{t+1}}\left(P(x^0) - P^*\right) + \frac{9\mu_0}{2^{t+2}}\|x^0 - x^*\|^2 \\ &\leq \frac{C_0^4}{(\sum_{i=0}^{t} k_i - T_0)^4}\left(P(x^0) - P^*\right) + \frac{9\mu_0 C_0^2}{2(\sum_{i=0}^{t} k_i - T_0)^2}\|x^0 - x^*\|^2.\end{aligned}$$

Then it suffices to note that $\bar{x}^s = \hat{x}^{t+1}$ where $s = \sum_{i=0}^{t} k_i$ is the number of L-Katyusha iterations. ■

Note that $\bar{x}^s$ is obtained after s L-Katyusha iterations. Hence, in view of Theorem 19, to obtain $\mathbb{E}[P(\bar{x}^s) - P^*] \leq \epsilon$, the total number of iterations is bounded by

$$\begin{aligned}&\mathcal{O}\left(T_0 + \frac{C_0}{\epsilon^{1/4}} + \frac{\sqrt{\mu_0}C_0}{\sqrt{\epsilon}}\right) \qquad (26)\\ &= \begin{cases} \mathcal{O}\left(\frac{1}{p\epsilon^{1/4}} + \sqrt{\frac{\mathcal{L}_2}{p\epsilon}}\right) & \text{if } L_f \leq \frac{\mathcal{L}_2}{p} \text{ and } \mu_0 = \mathcal{L}_2 p \\ \mathcal{O}\left(\left(\frac{1}{p\epsilon^{1/4}} + \sqrt{\frac{\mathcal{L}_f}{\epsilon}}\right)\ln\frac{pL_f}{\mathcal{L}_2}\right) & \text{if } L_f > \frac{\mathcal{L}_2}{p} \text{ and } \mu_0 = L_f p^2/4. \end{cases}\end{aligned}$$

5. L-SVRG in the Nonconvex and Smooth Case

In this section, we consider L-SVRG (Algorithm 1) with $\psi \equiv 0$ and f being possibly nonconvex.

Assumption 7 *There is a constant $\mathcal{L}_3 > 0$ such that*

$$\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x) - \mathbf{G}(y)\right)\Theta_S \mathbf{I}_S e - \frac{1}{n}\left(\mathbf{G}(x) - \mathbf{G}(y)\right)e\right\|^2\right] \leq \mathcal{L}_3\|x - y\|^2, \ \ \forall x, y \in \mathbb{R}^d.$$

5.1 Two Lemmas

We first state and prove two auxiliary results.

Lemma 20 *Under Assumption 7, we have*

$$\mathbb{E}_k\left[\|g^k\|^2\right] \leq 2\|\nabla f(x^k)\|^2 + 2\mathcal{L}_3\|x^k - w^k\|^2.$$

Lemma 21 *For any $\beta > 0$, we have*

$$\mathbb{E}_k\left[\|x^{k+1} - w^{k+1}\|^2\right] \leq \eta^2\mathbb{E}_k\left[\|g^k\|^2\right] + (1-p)(1+\eta\beta)\|x^k - w^k\|^2 + (1-p)\frac{\eta}{\beta}\|\nabla f(x^k)\|^2.$$

5.2 The Main Result

We are now ready to state the main result of this section.

Theorem 22 (Sublinear convergence of L-SVRG in nonconvex and smooth case) *Consider Algorithm 1. Under Assumption 7, consider the Lyapunov function $\Upsilon^k := f(x^k) + \alpha\|x^k - w^k\|^2$, where $\alpha = 3\eta^2 L_f\mathcal{L}_3/p$. Let $\beta = p/3\eta$. If stepsize η satisfies*

$$\eta \leq \min\left\{\frac{1}{4L_f}, \frac{p^{\frac{2}{3}}}{36^{\frac{1}{3}}(L_f\mathcal{L}_3)^{\frac{1}{3}}}, \frac{\sqrt{p}}{\sqrt{6\mathcal{L}_3}}\right\}, \tag{27}$$

then $\mathbb{E}_k\left[\Upsilon^{k+1}\right] \leq \Upsilon^k - \frac{\eta}{4}\|\nabla f(x^k)\|^2$.

Proof Since f is L_f-smooth, we have $f(x^{k+1}) \leq f(x^k) + \langle\nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2$, which implies $\mathbb{E}_k\left[f(x^{k+1})\right] \leq f(x^k) - \eta\|\nabla f(x^k)\|^2 + \frac{L_f\eta^2}{2}\mathbb{E}_k\left[\|g^k\|^2\right]$. Hence, we have

$$\begin{aligned}
\mathbb{E}_k\left[\Upsilon^{k+1}\right] &= \mathbb{E}_k\left[f(x^{k+1}) + \alpha\|x^{k+1} - w^{k+1}\|^2\right] \\
&\leq f(x^k) - \eta\|\nabla f(x^k)\|^2 + \frac{L_f\eta^2}{2}\mathbb{E}_k\left[\|g^k\|^2\right] + \alpha\mathbb{E}_k\left[\|x^{k+1} - w^{k+1}\|^2\right] \\
&\overset{Lemma\ 21}{\leq} f(x^k) - \eta\|\nabla f(x^k)\|^2 + \eta^2\left(\frac{L_f}{2} + \alpha\right)\mathbb{E}_k\left[\|g^k\|^2\right] \\
&\quad + \alpha(1-p)(1+\eta\beta)\|x^k - w^k\|^2 + \alpha(1-p)\frac{\eta}{\beta}\|\nabla f(x^k)\|^2 \\
&\overset{Lemma\ 20}{\leq} f(x^k) - \eta\left(1 - \frac{\alpha(1-p)}{\beta}\right)\|\nabla f(x^k)\|^2 + \alpha(1-p)(1+\eta\beta)\|x^k - w^k\|^2 \\
&\quad + \eta^2\left(\frac{L_f}{2} + \alpha\right)\left(2\|\nabla f(x^k)\|^2 + 2\mathcal{L}_3\|x^k - w^k\|^2\right) \\
&= f(x^k) - \eta\left(1 - \frac{\alpha(1-p)}{\beta} - L_f\eta - 2\alpha\eta\right)\|\nabla f(x^k)\|^2 \\
&\quad + \alpha\left((1-p)(1+\eta\beta) + \eta^2\left(\frac{L_f}{\alpha} + 2\right)\mathcal{L}_3\right)\|x^k - w^k\|^2.
\end{aligned}$$

Since $\alpha = 3\eta^2 L_f \mathcal{L}_3/p$ and $\beta = p/3\eta$, we have

$$(1-p)(1+\eta\beta) + \eta^2\left(\frac{L_f}{\alpha}+2\right)\mathcal{L}_3 \leq 1 - \frac{p}{3} + 2\mathcal{L}_3\eta^2.$$

Let

$$2\mathcal{L}_3\eta^2 \leq \frac{p}{3}, \qquad \frac{\alpha}{\beta} = \frac{9\eta^3 L_f \mathcal{L}_3}{p^2} \leq \frac{1}{4}, \qquad L_f\eta \leq \frac{1}{4}, \qquad 2\alpha\eta = \frac{6\eta^3 L_f\mathcal{L}_3}{p} \leq \frac{1}{4},$$

which implies

$$\eta \leq \min\left\{\frac{1}{4L_f}, \frac{p^{\frac{2}{3}}}{36^{\frac{1}{3}}(L_f\mathcal{L}_3)^{\frac{1}{3}}}, \frac{\sqrt{p}}{\sqrt{6\mathcal{L}_3}}\right\}.$$

Then $(1-p)(1+\eta\beta)+\eta^2(\frac{L_f}{\alpha}+2)\mathcal{L}_3 \leq 1$ and $1 - \frac{\alpha(1-p)}{\beta} - L_f\eta - 2\alpha\eta \geq \frac{1}{4}$, which indicate that

$$\mathbb{E}_k\left[\Upsilon^{k+1}\right] \leq \Upsilon^k - \frac{\eta}{4}\|\nabla f(x^k)\|^2.$$

■

Corollary 23 *Let x^a be chosen uniformly at random from $\{x^i\}_{i=0}^k$ and the stepsize η satisfy (27). Then $\mathbb{E}\left[\|\nabla f(x^a)\|^2\right] \leq \frac{4}{\eta}\cdot\frac{f(x^0)-f(x^*)}{k+1}$. If the stepsize η is equal to the upper bound in (27), then $\mathbb{E}\left[\|\nabla f(x^a)\|^2\right] \leq \epsilon$ as long as*

$$k \geq \mathcal{O}\left(\left(L_f + \frac{(L_f\mathcal{L}_3)^{\frac{1}{3}}}{p^{\frac{2}{3}}} + \sqrt{\frac{\mathcal{L}_3}{p}}\right)\frac{f(x^0)-f(x^*)}{\epsilon}\right).$$

6. Estimations of Expected Smoothness Parameters

In the previous sections, we have shown that the stepsize control and convergence analysis of stochastic variance reduced methods rely on the constants $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ satisfying Assumption 5, Assumption 6 and Assumption 7 respectively. In this section, we study the constants $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ under the following assumption.

Assumption 8 *For each $i \in [n]$, there is $L_i > 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x-y\|.$$

Recall that under the above assumption, if in addition f_i is convex, then (Nesterov, 2004, Theorem 2.1.5)

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i(f_i(x) - f_i(y) - \langle\nabla f_i(y), x-y\rangle). \tag{28}$$

First we give the estimations of these expected smoothness parameters for arbitrary set sampling S. Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be defined by $\mathbf{P}_{ij} = \mathbb{P}\left[\{i,j\} \subset S\right]$. Note that

$$\mathbf{P}_{ij} = \sum_{C\subset[n]: i,j\in C} p_C. \tag{29}$$

Recall that Θ_C^i denotes the ith diagonal entry of Θ_C. Let $f_S \coloneqq \frac{1}{n} F \Theta_S \mathbf{I}_S e$, and the Lipschitz smoothness constant of f_S be L_S. Obviously $L_S \leq \frac{1}{n}\sum_{i\in S} L_i \Theta_S^i$. Let $\mathcal{L}_{\max} \coloneqq \max_{i\in[n]} \sum_{C:i\in C} p_C L_C \Theta_C^i$. Then we have the following lemma.

Lemma 24 *Let S be a proper set sampling. Suppose that each f_i is convex and Assumption 8 holds. Then Assumption 5 and Assumption 6 hold with*

$$\mathcal{L}_i \leq \mathcal{L}_{\max} \leq \frac{1}{n} \max_{i\in[n]} \left\{ \sum_{j\in[n]} \sum_{C:i,j\in C} p_C \Theta_C^i \Theta_C^j L_j \right\}, \quad i = 1, 2.$$

Specifically, if $\Theta_C^i = \frac{1}{p_i}$ for all i and C, then

$$\mathcal{L}_i \leq \mathcal{L}_{\max} \leq \frac{1}{n} \max_{i\in[n]} \left\{ \sum_{j\in[n]} \frac{\mathbf{P}_{ij}}{p_i p_j} L_j \right\}, \quad i = 1, 2.$$

Lemma 25 *Let S be a proper set sampling. Suppose that Assumption 8 holds. Then Assumption 7 holds with*

$$\mathcal{L}_3 \leq \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{C:i,j\in C} p_C \Theta_C^i \Theta_C^j L_i L_j.$$

Specifically, if $\Theta_C^i = \frac{1}{p_i}$ for all i and C, then

$$\mathcal{L}_3 \leq \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{\mathbf{P}_{ij}}{p_i p_j} L_i L_j.$$

Next, we consider the following assumption on the sampling distribution $\mathcal{S}$.

Assumption 9 *There exist constants $\mathcal{A}_i \geq 0$ for each $i \in [n]$ and $0 \leq \mathcal{B} \leq 1$ such that for any matrix $\mathbf{M} \in \mathbb{R}^{d\times n}$ and $S \sim \mathcal{S}$*

$$\mathbb{E}\left[\left\| \mathbf{M}\Theta_S \mathbf{I}_S e \right\|^2 \right] \leq \sum_{i=1}^{n} \mathcal{A}_i \left\| \mathbf{M}_{:i} \right\|^2 + \mathcal{B} \left\| \mathbf{M} e \right\|^2, \tag{30}$$

where $\mathbf{M}_{:i}$ denotes the ith column vector of $\mathbf{M}$.

Assumption 9 appeared in (Qian et al., 2019) for the convergence analysis of SAGA. As we shall see, the constants $\mathcal{A}_i$, B and L_i jointly determine the constants $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$.

Theorem 26 *Under Assumption 8 and Assumption 9, we have*

$$\mathcal{L}_3 \leq \frac{1}{n^2} \sum_{i=1}^{n} \mathcal{A}_i L_i^2 + \max(\mathcal{B} - 1, 0) \max\left(L_f, |\mu_f| \right).$$

If in addition f_i is convex for each $i \in [n]$, then

$$\mathcal{L}_1 \leq \frac{1}{n} \max_i \mathcal{A}_i L_i + \mathcal{B} L_f,$$
$$\mathcal{L}_2 \leq \frac{1}{n} \max_i \mathcal{A}_i L_i + \max\left(\mathcal{B} - 1, 0 \right) L_f.$$

Theorem 26 reduces the estimations of the constants $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ to the computation of constants $\mathcal{A}_i$ and $\mathcal{B}$ satisfying (30). We next provide computation formula of $\mathcal{A}_i$ and $\mathcal{B}$ for different types of sampling distribution $\mathcal{S}$.

6.1 Sampling Without Replacement

In this subsection, we consider the case where $S \sim \mathcal{S}$ is a set sampling.

Lemma 27 *[Qian et al., 2019] For an arbitrary set sampling S, Assumption 9 is satisfied for $\mathcal{B} = 0$ and*

$$\mathcal{A}_i = \sum_{C \subseteq [n]: i \in C} p_C |C| (\Theta_C^i)^2, \ \ \forall i \in [n].$$

Lemma 27 provides an estimation for arbitrary set sampling S and $\{\Theta_C : C \subset [n]\}$. Next we consider the special case where $\Theta_C^i = p_i^{-1}$ for all $i \in [n]$ and all subset $C \subset [n]$. In this case,

$$\mathbb{E}\left[\|\mathbf{M}\Theta_S \mathbf{I}_S e\|^2\right] = \sum_{i,j=1}^n \sum_{C \subset [n]: i,j \in C} p_C \left\langle \Theta_C^i \mathbf{M}_{:i}, \Theta_C^j \mathbf{M}_{:j} \right\rangle \overset{(29)}{=} \sum_{i,j=1}^n \frac{\mathbf{P}_{ij}}{p_i p_j} \left\langle \mathbf{M}_{:i}, \mathbf{M}_{:j} \right\rangle. \quad (31)$$

Based on (31), we now provide explicit bounds for two specific set samplings.

6.1.1 τ-Nice Sampling

Lemma 28 *[Qian et al., 2019] For τ-nice sampling S and $\Theta_S^i = p_i^{-1}$, Assumption 9 is satisfied for $\mathcal{B} = \frac{n(\tau-1)}{\tau(n-1)}$ and*

$$\mathcal{A}_i = \frac{n(n-\tau)}{\tau(n-1)}, \ \ \forall i \in [n].$$

6.1.2 Group Sampling

We consider the group sampling associated with a partition $\{C_1, \ldots, C_t\}$ of $[n]$. We have $\mathbf{P}_{ij} = 0$ if i, j are in the same group, and $\mathbf{P}_{ij} = p_i p_j$ if i, j are in different groups. Denote by $\mathcal{I}$ the subset of indices which are in a group of one single element.

Lemma 29 *For group sampling S and $\Theta_S^i = p_i^{-1}$, Assumption 9 is satisfied for $\mathcal{B} = 1$ and*

$$\mathcal{A}_i = \left(\frac{1}{p_i} - 1\right), \ \ \forall i \in \mathcal{I}, \qquad \mathcal{A}_i = \frac{1}{p_i}, \ \ \forall i \notin \mathcal{I}.$$

6.2 Sampling With Replacement

In this subsection, we consider the special case when $S \sim \mathcal{S}$ is sampling with replacement of size τ with respect to the distribution vector $(\tilde{p}_1, \ldots, \tilde{p}_n)$. We let $\Theta_S^i = (\tau \tilde{p}_i)^{-1}$. Then

$$\mathbb{E}\left[\|\mathbf{M}\Theta_S \mathbf{I}_S e\|^2\right] = \mathbb{E}\left[\left\|\sum_{k=1}^{\tau} \frac{1}{\tau \tilde{p}_{i_k}} \mathbf{M}_{:i_k}\right\|^2\right],$$

where $i_1, \dots i_\tau$ are i.i.d. random integers and equal to $i \in [n]$ with probability $\tilde{p}_i$. Hence

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{M}\Theta_S\mathbf{I}_S e\|^2\right] &= \mathbb{E}\left[\sum_{k=1}^{\tau}\frac{1}{\tau^2\tilde{p}_{i_k}^2}\|\mathbf{M}_{:i_k}\|^2\right] + \sum_{k\neq\ell}\frac{1}{\tau^2}\left\langle\mathbb{E}\left[\frac{1}{\tilde{p}_{i_k}}\mathbf{M}_{:i_k}\right], \mathbb{E}\left[\frac{1}{\tilde{p}_{i_\ell}}\mathbf{M}_{:i_\ell}\right]\right\rangle \\ &= \frac{1}{\tau}\sum_{i=1}^{n}\frac{1}{\tilde{p}_i}\|\mathbf{M}_{:i}\|^2 + \frac{\tau^2-\tau}{\tau^2}\|\mathbf{M}e\|^2.\end{aligned}$$

Lemma 30 *For sampling with replacement S of size τ with respect to the distribution vector $(\tilde{p}_1, \dots, \tilde{p}_n)$, Assumption 9 is satisfied for $\mathcal{B} = 1 - \frac{1}{\tau}$ and*

$$\mathcal{A}_i = \frac{1}{\tau\tilde{p}_i}, \ \ \forall i \in [n].$$

6.3 Constants Estimations

Combining Theorem 26 with the estimations of constants $\mathcal{A}_i$ and $\mathcal{B}$ in Lemma 27, Lemma 28, Lemma 29, and Lemma 30, we obtain the following bounds on $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$.

Corollary 31 *1. For arbitrary set sampling S,*

	Upper bounds	*Assumptions*
$\mathcal{L}_1$	$\frac{1}{n}\max_i \sum_{C\subseteq[n]:i\in C} p_C\|C\|(\Theta_C^i)^2 L_i$	*all f_i are convex, Assumption 8*
$\mathcal{L}_2$	$\frac{1}{n}\max_i \sum_{C\subseteq[n]:i\in C} p_C\|C\|(\Theta_C^i)^2 L_i$	*all f_i are convex, Assumption 8*
$\mathcal{L}_3$	$\frac{1}{n^2}\sum_{i=1}^n \sum_{C\subseteq[n]:i\in C} p_C\|C\|(\Theta_C^i)^2 L_i^2$	*Assumption 8*

2. For τ-nice sampling S and $\Theta_C^i \equiv p_i^{-1}$,

	Upper bounds	*Assumptions*
$\mathcal{L}_1$	$\frac{n-\tau}{\tau(n-1)}\max_i L_i + \frac{n(\tau-1)}{\tau(n-1)}L_f$	*all f_i are convex, Assumption 8*
$\mathcal{L}_2$	$\frac{n-\tau}{\tau(n-1)}\max_i L_i$	*all f_i are convex, Assumption 8*
$\mathcal{L}_3$	$\frac{n-\tau}{\tau n(n-1)}\sum_{i=1}^n L_i^2$	*Assumption 8*

3. For group sampling S and $\Theta_C^i \equiv p_i^{-1}$,

	Upper bounds	*Assumptions*
$\mathcal{L}_1$	$\frac{1}{n}\max\left\{\max_{i\in\mathcal{I}}(\frac{1}{p_i}-1)L_i, \ \max_{i\notin\mathcal{I}}\frac{L_i}{p_i}\right\} + L_f$	*all f_i are convex, Assumption 8*
$\mathcal{L}_2$	$\frac{1}{n}\max\left\{\max_{i\in\mathcal{I}}(\frac{1}{p_i}-1)L_i, \ \max_{i\notin\mathcal{I}}\frac{L_i}{p_i}\right\}$	*all f_i are convex, Assumption 8*
$\mathcal{L}_3$	$\frac{1}{n^2}\sum_{i\in\mathcal{I}}\left(\frac{1}{p_i}-1\right)L_i^2 + \frac{1}{n^2}\sum_{i\notin\mathcal{I}}\frac{L_i^2}{p_i}$	*Assumption 8*

4. For sampling with replacement S and $\Theta_C^i \equiv (\tau\tilde{p}_i)^{-1}$,

	Upper bounds	*Assumptions*
$\mathcal{L}_1$	$\frac{1}{n\tau}\max_i\frac{L_i}{\tilde{p}_i} + (1-\frac{1}{\tau})L_f$	*all f_i are convex, Assumption 8*
$\mathcal{L}_2$	$\frac{1}{n\tau}\max_i\frac{L_i}{\tilde{p}_i}$	*all f_i are convex, Assumption 8*
$\mathcal{L}_3$	$\frac{1}{n^2\tau}\sum_{i=1}^n\frac{L_i^2}{\tilde{p}_i}$	*Assumption 8*

6.4 ESO

In this subsection, we give the estimations of these expected smoothness parameters under the ESO inequality. Consider $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$, where $\mathbf{A}_i \in \mathbb{R}^{d\times m}$, $\phi_i : \mathbb{R}^m \to \mathbb{R}$ is $1/\gamma$-smooth.

The parameters $v_1, \ldots, v_n$ are assumed to satisfy the following expected separable over-approximation (ESO) inequality, which needs to hold for all $h_i \in \mathbb{R}^m$

$$\mathbb{E}_S\left[\left\|\sum_{i\in S}\mathbf{A}_i h_i\right\|^2\right] \leq \sum_{i=1}^n p_i v_i \|h_i\|^2. \tag{32}$$

Lemma 32 *If ϕ_i is $1/\gamma$-smooth and convex, then for any $x, y \in \mathbb{R}^d$, we have*

$$\left\|\nabla\phi_i(\mathbf{A}_i^\top x) - \nabla\phi_i(\mathbf{A}_i^\top y)\right\|^2 \leq \frac{2}{\gamma}\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x - y\rangle\right).$$

Lemma 33 *Let S be a proper sampling, and $\Theta_S^i = 1/p_i$. Let ϕ_i be $1/\gamma$-smooth and convex. If the ESO inequality (32) holds, then the expected smoothness constants $\mathcal{L}_1$ in Assumption 5 and $\mathcal{L}_2$ in Assumption 6 satisfy*

$$\mathcal{L}_i \leq \frac{1}{n\gamma}\max_i\left\{\frac{v_i}{p_i}\right\}, \quad i = 1, 2.$$

Lemma 34 *Let S be a proper sampling, and $\Theta_S^i = 1/p_i$. Let ϕ_i be $1/\gamma$-smooth. If the ESO inequality (32) holds, then the $\mathcal{L}_3$ in Assumption 7 satisfies*

$$\mathcal{L}_3 \leq \frac{1}{n^2\gamma^2}\sum_{i=1}^n \frac{v_i\|\mathbf{A}_i\|^2}{p_i}.$$

6.5 Importance Sampling

Since the complexity bounds of the algorithms increase with the constants $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$. It is natural to choose the sampling strategy minimizing those constants. In this subsection, we consider group sampling and sampling with replacement and study the influence of the importance sampling. Let $\tau = \mathbb{E}[|S|]$ be the expected cardinality of S, counting multiplicity. We denote $\bar{L} := \frac{1}{n}\sum_{i=1}^n L_i$. Note that $L_f \leq \bar{L}$.

6.5.1 Group Sampling

Let

$$q_i = \frac{L_i\tau}{\sum_{i=1}^n L_i}, \ \ \forall i \in [n], \tag{33}$$

and $T = \{i | q_i > 1\}$. If for each $i \in [n]$

$$p_i = \begin{cases} q_i + \delta_i & \text{if } i \notin T \\ 1 & \text{otherwise} \end{cases}$$

for some $\delta_i \in [0, 1-q_i]$ such that $\sum_{i=1}^n p_i = \tau$, then

$$\max\left\{\max_{i\in\mathcal{I}}\left(\frac{1}{p_i}-1\right)L_i,\ \max_{i\notin\mathcal{I}}\frac{L_i}{p_i}\right\} \le \max_{i\in[n]}\frac{L_i}{q_i} = \frac{n\bar{L}}{\tau},$$

and

$$\sum_{i\in\mathcal{I}}\left(\frac{1}{p_i}-1\right)L_i^2 + \sum_{i\notin\mathcal{I}}\frac{L_i^2}{p_i} \le \sum_{i\in[n]}\frac{L_i^2}{q_i} = \frac{n^2\bar{L}^2}{\tau}.$$

Here we used the fact that if $p_i = 1$, then necessarily $i \in \mathcal{I}$, otherwise $p_i \ge q_i$. Hence for this group sampling we can achieve

$$\mathcal{L}_1 \le \frac{\bar{L}}{\tau} + L_f, \qquad \mathcal{L}_2 \le \frac{\bar{L}}{\tau}, \qquad \mathcal{L}_3 \le \frac{\bar{L}^2}{\tau}. \tag{34}$$

We call this sampling importance group sampling. If instead we replace (33) by

$$q_i = \frac{\tau}{n},\ \ \forall i \in [n],$$

then the bounds are

$$\mathcal{L}_1 \le \frac{\max_i L_i}{\tau} + L_f, \qquad \mathcal{L}_2 \le \frac{\max_i L_i}{\tau}, \qquad \mathcal{L}_3 \le \frac{\sum_i L_i^2}{\tau n}. \tag{35}$$

6.5.2 Sampling With Replacement

If for each $i \in [n]$

$$\tilde{p}_i = \frac{L_i}{\sum_{i=1}^n L_i}, \tag{36}$$

then

$$\max_i \frac{L_i}{\tilde{p}_i} = n\bar{L}, \qquad \sum_{i=1}^n \frac{L_i^2}{\tilde{p}_i} = n^2\bar{L}^2.$$

Hence for this sampling with replacement we can achieve the same bounds as in (34). We call this sampling importance sampling with replacement. If instead we replace (36) by

$$\tilde{p}_i = \frac{1}{n},$$

then the we can only achieve the same bounds as in (35).

From (34) and (35), we know importance group sampling and importance sampling with replacement are preferable to uniform sampling in general. In order to construct importance sampling, we need to know the ratio L_i/L_j for $i, j \in [n]$. For instance, if $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$ as in Section 6.4, we have

$$L_i \le \frac{1}{\gamma}\|\mathbf{A}_i\mathbf{A}_i^\top\| \le \frac{1}{\gamma}\mathrm{tr}(\mathbf{A}_i\mathbf{A}_i^\top),$$

where $\mathrm{tr}(\mathbf{A}_i\mathbf{A}_i^\top)$ means the trace of $\mathbf{A}_i\mathbf{A}_i^\top$. Therefore, we can estimate L_i/L_j by computing $\|\mathbf{A}_i\mathbf{A}_i^\top\|$ or $\mathrm{tr}(\mathbf{A}_i\mathbf{A}_i^\top)$ for $i \in [n]$. On the other hand, if we could not get a good estimation of L_i/L_j, we can use uniform sampling instead.

6.6 Iteration Complexity Bounds

Next, we insert the bounds (34) into previous results to get explicit complexity bounds for each algorithm. Although our theory allows arbitrary changing probability p, for simplicity we consider $p = \tau/n$ in this section. In this case, the expected cost of each iteration is 2τ.

L-SVRG.

1. When f or ψ is strongly convex, by Theorem 6 and (34), the iteration complexity of L-SVRG can be bounded by

$$\mathcal{O}\left(\left(\frac{n}{\tau} + \frac{L_f}{\mu} + \frac{\bar{L}}{\mu\tau}\right)\log\frac{1}{\epsilon}\right). \tag{37}$$

   Such complexity bound is comparable with that of SAGA-AS with importance mini-batch sampling (Qian et al., 2019). Note that as SAGA-AS, L-SVRG does not need to know the strong convexity parameter μ. From (37) we see that linear speedup is achieved when $\tau \leq \frac{\bar{L}}{L_f}$.

2. When f is convex, by Theorem 17 and (34), the iteration complexity of L-SVRG can be bounded by

$$\mathcal{O}\left(\left(\frac{n}{\tau} + L_f + \frac{\bar{L}}{\tau}\right)\frac{1}{\epsilon}\right),$$

   which suggests a linear speedup up to $\tau \leq \frac{\bar{L}}{L_f}$.

3. When f is nonconvex and $\psi \equiv 0$, by Corollary 23 and (34), the iteration complexity of L-SVRG can be bounded by

$$\mathcal{O}\left(\left(L_f + \frac{n^{\frac{2}{3}}(L_f\bar{L}^2)^{\frac{1}{3}}}{\tau} + \frac{\sqrt{n}\bar{L}}{\tau}\right)\frac{1}{\epsilon}\right).$$

   In (Horváth and Richtárik, 2019), the iteration complexity for SVRG and SAGA with the optimal sampling is proved to be bounded by

$$\mathcal{O}\left(\frac{(1+\frac{n-\tau}{n})\bar{L}n^{\frac{2}{3}}}{\tau\epsilon}\right),$$

   for $\tau \leq \mathcal{O}(n^{2/3})$. We can see our bound is at least as good as theirs, and could be better if L_f is smaller than $\bar{L}$. Furthermore, our bound holds for any $1 \leq \tau \leq n$, while the one in (Horváth and Richtárik, 2019) only holds for $\tau \leq \mathcal{O}(n^{2/3})$.

L-Katyusha. When f or ψ is strongly convex, by Corollary 11 and (34), the iteration complexity of L-Katyusha can be bounded by

$$\mathcal{O}\left(\left(\frac{n}{\tau} + \sqrt{\frac{L_f}{\mu}} + \frac{1}{\tau}\sqrt{\frac{n\bar{L}}{\mu}}\right)\log\frac{1}{\epsilon}\right).$$

This is the same iteration complexity bound of the original Katyusha with importance sampling with replacement in (Allen-Zhu, 2017). The numerical experiments also confirm

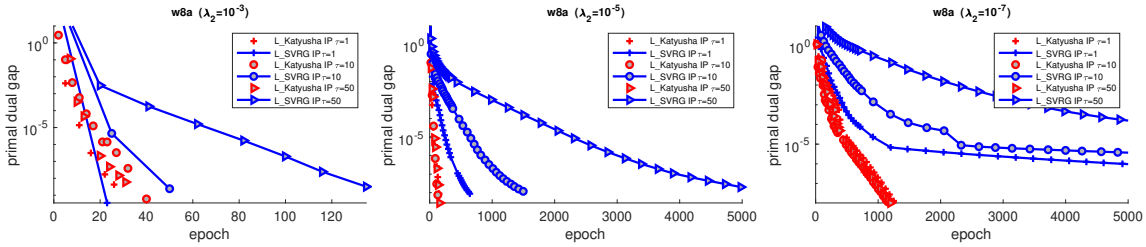
the similarity of the two methods in terms of iteration complexity, see Figure 7 and Figure 9. Note that linear speedup is achieved for $\tau \leq \sqrt{n\bar{L}/L_f}$.

AdapReg-L-Katyusha. When f is convex, by (26) and (34), the iteration complexity of AdapReg-L-Katyusha can be bounded by

$$\begin{cases} \mathcal{O}\left(\frac{n}{\tau\epsilon^{1/4}} + \frac{1}{\tau}\sqrt{\frac{n\bar{L}}{\epsilon}}\right) & \text{if } \tau \leq \sqrt{\frac{n\bar{L}}{L_f}} \text{ and } \mu_0 = \frac{\bar{L}}{n} \\ \mathcal{O}\left(\left(\frac{n}{\tau\epsilon^{1/4}} + \sqrt{\frac{\mathcal{L}_f}{\epsilon}}\right)\ln\frac{\tau^2 L_f}{n\bar{L}}\right) & \text{if } \tau > \sqrt{\frac{n\bar{L}}{L_f}} \text{ and } \mu_0 = L_f\frac{\tau^2}{4n^2}. \end{cases}$$

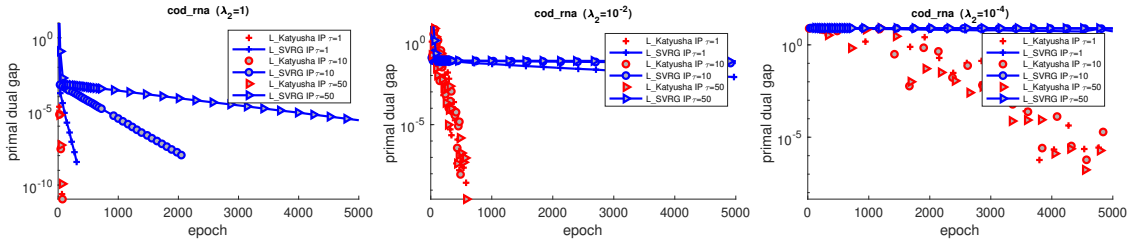
Clearly linear speedup is achieved for $\tau \leq \sqrt{n\bar{L}/L_f}$.

7. Numerical Experiments



(a) $\lambda_2 = 10^{-3}$ (b) $\lambda_2 = 10^{-5}$ (c) $\lambda_2 = 10^{-7}$

Figure 1: L-SVRG V.S. L-Katyusha, w8a



(a) $\lambda_2 = 1$ (b) $\lambda_2 = 10^{-2}$ (c) $\lambda_2 = 10^{-4}$

Figure 2: L-SVRG V.S. L-Katyusha, cod-rna

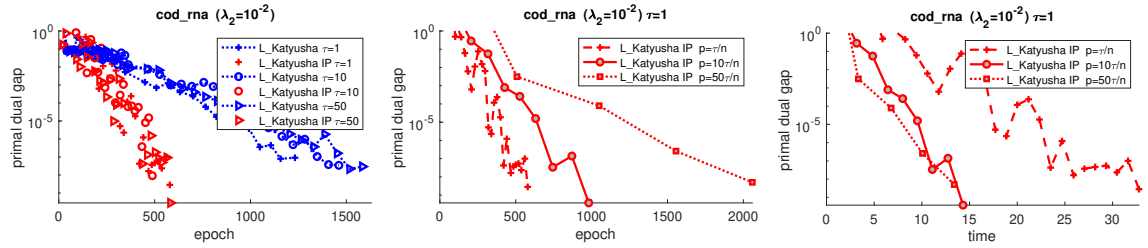
We tested L-SVRG (Algorithm 1) and L-Katyusha (Algorithm 2) on the logistic regression problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \log(1 + e^{b_i A_i^\top x}) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2, \tag{38}$$

with fixed $\lambda_1 = 10^{-4}$ and different values of λ_2. The data sets that we used are all downloaded from `https://www.csie.ntu.edu.tw/$\sim$cjlin/libsvmtools/datasets/`.

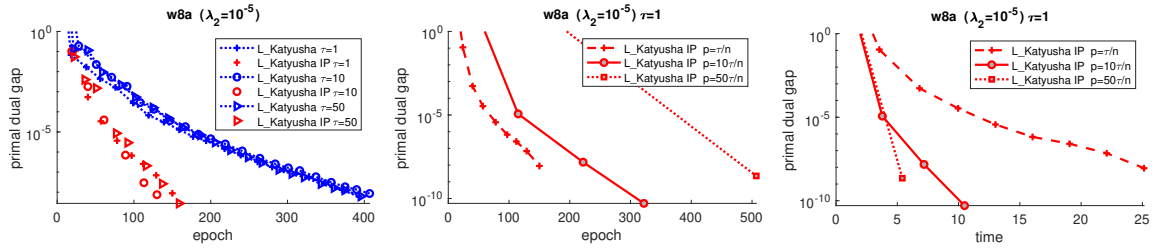
In all the plots, L-SVRG and L-Katyusha refer respectively to Algorithm 1 and Algorithm 2 with the uniform sampling strategy. L-SVRG IP and L-Katyusha IP mean that importance sampling with replacement as described in Section 6.5 is used. Katyusha refers to the original Katyusha algorithm proposed in (Allen-Zhu, 2017). Since in practice group sampling and sampling with replacement have similar convergence behaviour, here we only show the results obtained with sampling with replacement. In all the plots, the y-axis corresponds to the primal-dual gap of the iterates $\{x^k\}$. The x-axis may be the number of epochs, counted as $k\tau/n$ plus the number of times we change w^k, or the actual running time. The experiments were carried out on a MacBook (1.2 GHz Intel Core m3 with 16 GB RAM) running macOS High Sierra 10.13.1.

Comparison of L-SVRG and L-Katyusha: In Figure 1 and Figure 2 we compare L-SVRG with L-Katyusha, both with the importance sampling strategy for w8a and cod_rna and three different values of λ_2. In each plot, we compare three different minibatch sizes τ. The numerical results show that the number of epochs of L-SVRG generally increases with τ (since $\bar{L}/L_f$ is not large in these examples), while that of L-Katyusha is stable and thus achieves a **linear speedup** in terms of the number of epochs, as predicted in Section 6.6.



(a) uniform V.S. IP (b) different p, epoch (c) different p, time

Figure 3: cod-rna



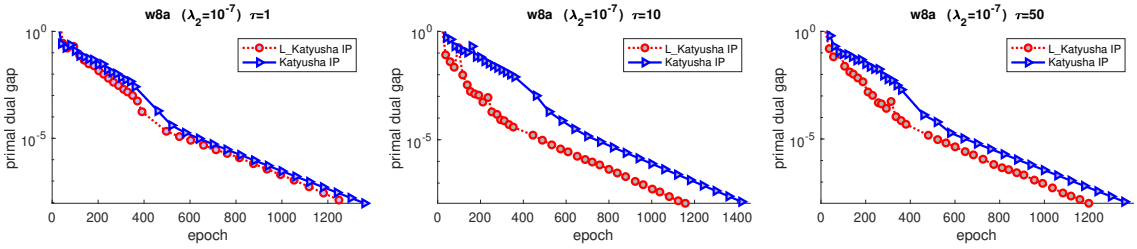
(a) uniform V.S. IP (b) different p, epoch (c) different p, time

Figure 4: w8a

Comparison of Uniform and Importance Sampling: Figure 3(a) and Figure 4(a) compare the uniform sampling strategy and the importance sampling strategy with three different values of τ, respectively, for the data set cod_rna and w8a. As predicted by theory, the importance sampling is better than the uniform sampling if $\bar{L}$ is smaller than $\max_i L_i$. Note that for cod_rna, $\bar{L} = 259,158$ and $\max_i L_i = 3,506,320$.

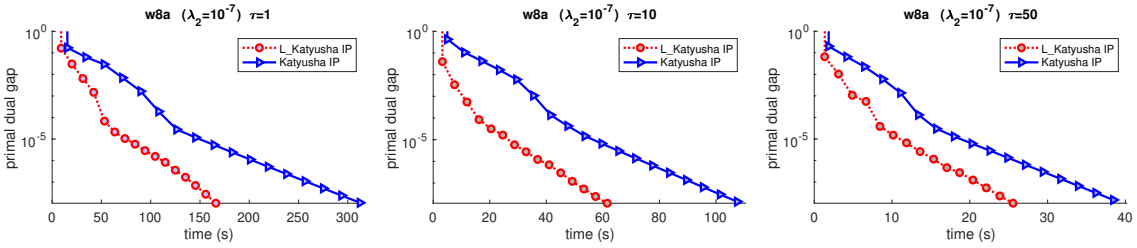
The updating probability p: Figure 3(b) and Figure 4(b) compare the performance of L-Katyusha for different choices of updating probability p of the reference point w^k, with x-axis being the number of epochs. Figure 3(c) and Figure 4(c) show the actual running time. Although the total number of epochs does increase with $p \geq \frac{\tau}{n}$, the running time can be significantly reduced by taking p larger than τ/n.

Comparison with Katyusha: Figure 5 to Figure 12 compare our loopless Katyusha with the original Katyusha proposed in (Allen-Zhu, 2017) for three different values of τ, based on the importance sampling strategy. We tested four data sets: w8a, real-sim, astro_ph, and a9a. While the performance of the two algorithms are similar in terms of epochs (Figure 5, 7, 9, 11), the actual running time of the loopless variant can be 20% to 50% less than that of Katyusha (Figure 6, 8, 10, 12). This is due to the additional averaging step in the original Katyusha method at the end of every inner loop, see Appendix A for further details.



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 5: L-Katyusha V.S. Katyusha, epoch plot, w8a

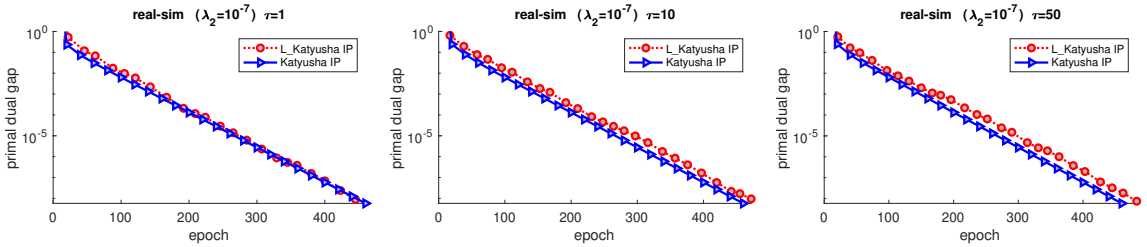


(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 6: L-Katyusha V.S. Katyusha, time plot, w8a

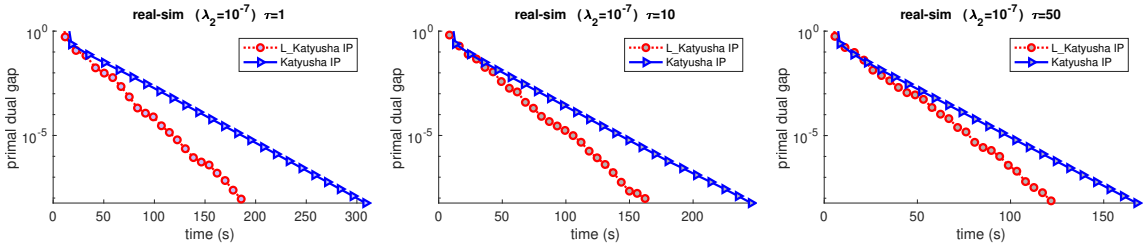
8. Conclusion

We presented loopless SVRG and loopless Katyusha for finite-sum optimization problems under the arbitrary sampling regime. We studied these algorithms in the strongly convex case and the non-strongly convex case, respectively, for composite objective functions. We analyzed loopless SVRG for the nonconvex and smooth case. The expected smoothness assumptions were used to detach the convergence analysis from the sampling strategy as well



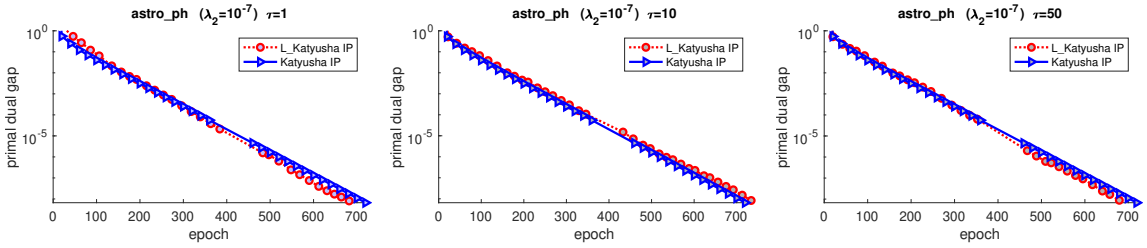
(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 7: L-Katyusha V.S. Katyusha, epoch plot, real-sim



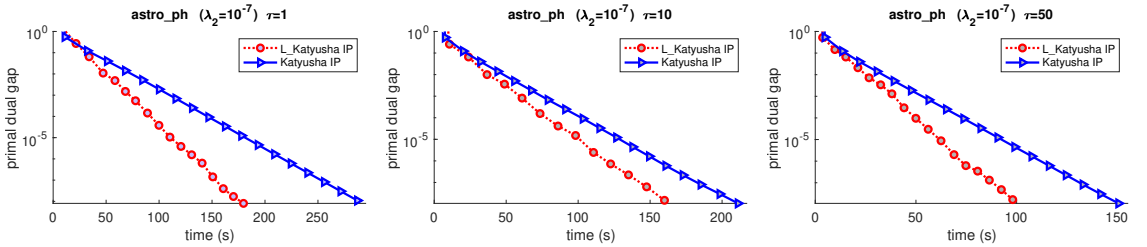
(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 8: L-Katyusha V.S. Katyusha, time plot, real-sim



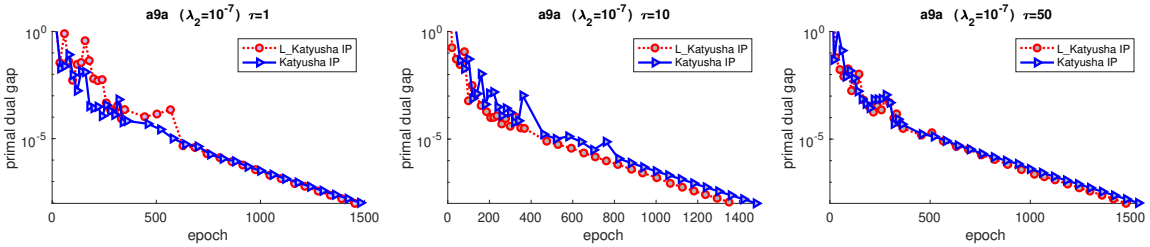
(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 9: L-Katyusha V.S. Katyusha, epoch plot, astro_ph



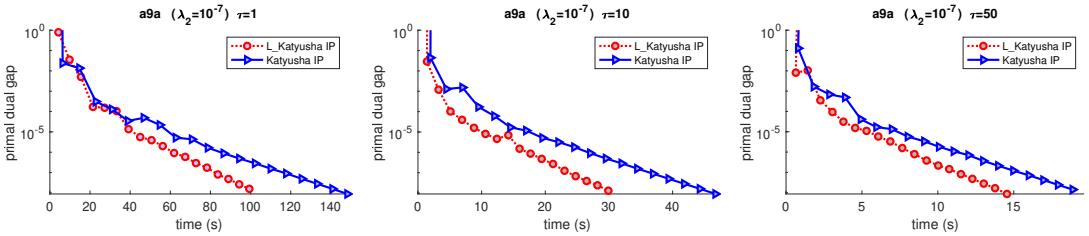
(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 10: L-Katyusha V.S. Katyusha, time plot, astro_ph



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 11: L-Katyusha V.S. Katyusha, epoch plot, a9a



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 50$

Figure 12: L-Katyusha V.S. Katyusha, time plot, a9a

as the convex and smooth properties of each f_i. Under the smoothness (and convexity) of each f_i, we estimated these expected smoothness parameters for arbitrary sampling strategies and established the connection between these expected smoothness parameters and ESO. Based on these estimations, we proposed importance group sampling and importance sampling with replacement which are preferable to uniform sampling generally. The linear speedup was also obtained when the expected minibatch size is in a certain range.

Acknowledgments

We thank the action editor and two anonymous referees for their valuable comments. All authors are thankful for support through the KAUST Baseline Research Funding Scheme. Xun Qian and Peter Richtárik acknowledge further funding by the Extreme Computing Research Center at KAUST, and administrative support from the Visual Computing Center at KAUST. Zheng Qu acknowledges further funding by Hong Kong Research Grants Council Early Career Scheme 27303016.

Appendix A. Efficient Implementation

The delayed update is a standard technique in stochastic variance reduced type methods for more efficiency when the Jacobian matrix $\mathbf{G}(x)$ is sparse. For the sake of completeness, we

provide details in the case when

$$\psi(x) \equiv \frac{\lambda_2}{2}\|x\|^2 + \lambda_1\|x\|_1,$$

for some $\lambda_2 > 0$. For Algorithm 1, the ith coordinate of the iterates $\{x^k\}$ satisfies

$$x_i^{k+1} = \arg\min_{a\in\mathbb{R}} \left\{ \frac{\lambda_2}{2}a^2 + \lambda_1|a| + \frac{1}{2\eta}\left(a - x_i^k + \eta g_i^k\right)^2 \right\}. \tag{39}$$

Let $t_0 < t_1$ be two positive integers. Suppose that

$$g_i^k = \hat{g}_i, \ \ \forall k = t_0, \ldots, t_1 - 1,$$

then the value of $x_i^{t_1}$ can be obtained without explicitly computing the value of $x_i^{t_0+1}, \ldots, x_i^{t_1-1}$. The details of computation can be found in (Zhang and Xiao, 2017). For convenience we give the pseudocode in Algorithm 5, so that

$$x_i^{t_1} = \text{delayed\_update}(t_0, t_1, \hat{g}_i, x_i^{t_0}, \eta).$$

Note that the complexity of Algorithm 5 is $\mathcal{O}(\log(t_1 - t_0))$ (Zhang and Xiao, 2017) while direct computation of $x_i^{t_1}$ from $x_i^{t_0}$ yields a time complexity $\mathcal{O}(t_1 - t_0)$. This is how the computation load can be reduced when $\mathbf{G}(x)$ is sparse.

A.1 Efficient Implementation for L-Katyusha

For Algorithm 2, the ith coordinate of the iterates $\{x^k, y^k, z^k\}$ satisfy

$$\begin{cases} x_i^k = \theta_1 z_i^k + \theta_2 w_i^k + (1 - \theta_1 - \theta_2) y_i^k \\ z_i^{k+1} = \arg\min_{a\in\mathbb{R}} \left\{ \frac{\lambda_2}{2}a^2 + \lambda_1|a| + \frac{(1+\eta\sigma_1)L}{2\eta}\left(a - \frac{\eta\sigma_1 x_i^k + z_i^k}{1+\eta\sigma_1} + \frac{\eta g_i^k}{(1+\eta\sigma_1)L}\right)^2 \right\} \\ y_i^{k+1} = x_i^k + \theta_1(z_i^{k+1} - z_i^k) \end{cases}$$

We eliminate x_i^k and obtain

$$\begin{cases} z_i^{k+1} = \arg\min_{a\in\mathbb{R}} \left\{ \frac{\lambda_2}{2}a^2 + \lambda_1|a| + \frac{(1+\eta\sigma_1)L}{2\eta} \right. \\ \qquad\qquad \left. \cdot \left(a - \frac{(\eta\sigma_1\theta_1+1)z_i^k + \eta\sigma_1\theta_2 w_i^k + \eta\sigma_1(1-\theta_1-\theta_2)y_i^k}{1+\eta\sigma_1} + \frac{\eta g_i^k}{(1+\eta\sigma_1)L}\right)^2 \right\} \\ y_i^{k+1} = \theta_1 z_i^{k+1} + \theta_2 w_i^k + (1 - \theta_1 - \theta_2) y_i^k. \end{cases}$$

- If $\lambda_1 = 0$, then the above system can be written as

$$\begin{cases} z_i^{k+1} = \frac{(\eta\sigma_1\theta_1+1)L}{\eta\lambda_2 + L(1+\eta\sigma_1)} z_i^k + \frac{\eta\sigma_1(1-\theta_1-\theta_2)L}{\eta\lambda_2+L(1+\eta\sigma_1)} y_i^k + \frac{\eta\sigma_1\theta_2 L w_i^k - \eta g_i^k}{\eta\lambda_2+L(1+\eta\sigma_1)} \\ y_i^{k+1} = \frac{\theta_1(\eta\sigma_1\theta_1+1)L}{\eta\lambda_2+L(1+\eta\sigma_1)} z_i^k + \left(1 - \theta_1 - \theta_2 + \frac{\theta_1\eta\sigma_1(1-\theta_1-\theta_2)L}{\eta\lambda_2+L(1+\eta\sigma_1)}\right) y_i^k + \theta_2 w_i^k \\ \qquad + \frac{\theta_1(\eta\sigma_1\theta_2 L w_i^k - \eta g_i^k)}{\eta\lambda_2+L(1+\eta\sigma_1)}. \end{cases}$$

  Then if

$$g_i^k = \hat{g}_i, w_i^k = \hat{w}_i \ \ \forall k = t_0, \ldots, t_1 - 1,$$

Algorithm 5 $\tilde{x}$ =delayed_update(t_0, t_1, u, x, η)

if $t_1 = t_0$ **then** $\tilde{x} = x$; return;
end if
$\alpha = (1+\eta\lambda_2)^{t_0-t_1}$
if $x = 0$ **then**
  if $\lambda_1 + u < 0$ **then** $\tilde{x} = \alpha x - (1-\alpha)(u+\lambda_1)/\lambda_2$
  else
    if $u - \lambda_1 > 0$ **then** $\tilde{x} = \alpha x - (1-\alpha)(u-\lambda_1)/\lambda_2$
    else
      $\tilde{x} = 0$
    end if
  end if
else
  if $x > 0$ **then**
    if $\lambda_1 + u \le 0$ **then** $\tilde{x} = \alpha x - (1-\alpha)(u+\lambda_1)/\lambda_2$
    else $t = t_0 + \log\left(1 + \frac{\lambda_2 x}{\lambda_1 + u}\right) / \log(1+\eta\lambda_2)$
      if $t < t_1$ **then**
        $t' = \lfloor t \rfloor$
        $\alpha' = (1+\eta\lambda_2)^{t_0-t'}$
        $x' = \alpha' x - (1-\alpha')(u+\lambda_1)/\lambda_2$
        $x'' = \arg\min_{a\in\mathbb{R}} \left\{ \frac{\lambda_2}{2}a^2 + \lambda_1|a| + au + \frac{1}{2\eta}(a-x')^2 \right\}$
        $\tilde{x} =$ delayed_update$(t'+1, t_1, u, x'', \eta)$
      else
        $\tilde{x} = \alpha x - (1-\alpha)(u+\lambda_1)/\lambda_2$
      end if
    end if
  else
    $\tilde{x} = -$delayed_update$(t_0, t_1, -u, -x, \eta)$
  end if
end if
Output $\tilde{x}$

$z_i^{t_1}$ and $y_i^{t_1}$ can be computed by

$$\begin{pmatrix} z_i^{t_1} \\ y_i^{t_1} \end{pmatrix} = A^{t_1-t_0} \begin{pmatrix} z_i^{t_0} \\ y_i^{t_0} \end{pmatrix} + \left(\sum_{s=0}^{t_1-t_0-1} A^s \right) b, \tag{40}$$

with

$$A = \begin{pmatrix} \frac{(\eta\sigma_1\theta_1+1)L}{\eta\lambda_2+L(1+\eta\sigma_1)} & \frac{\eta\sigma_1(1-\theta_1-\theta_2)L}{\eta\lambda_2+L(1+\eta\sigma_1)} \\ \frac{\theta_1(\eta\sigma_1\theta_1+1)L}{\eta\lambda_2+L(1+\eta\sigma_1)} & 1-\theta_1-\theta_2+\frac{\theta_1\eta\sigma_1(1-\theta_1-\theta_2)L}{\eta\lambda_2+L(1+\eta\sigma_1)} \end{pmatrix}, \quad b = \begin{pmatrix} \frac{\eta\sigma_1\theta_2 L w_i^k - \eta g_i^k}{\eta\lambda_2+L(1+\eta\sigma_1)} \\ \frac{\theta_1(\eta\sigma_1\theta_2 L w_i^k - \eta g_i^k)}{\eta\lambda_2+L(1+\eta\sigma_1)} \end{pmatrix}.$$

It is clear that (40) can be computed in $\mathcal{O}(\log(t_1 - t_0))$ time.

- If $\lambda_1 > 0$, we need to require $\sigma_1 = 0$ to have reduced computation load. In this case, we have a simplified recursive relation

$$\begin{cases} z_i^{k+1} = \arg\min_{a\in\mathbb{R}} \left\{ \frac{\lambda_2}{2}a^2 + \lambda_1|a| + \frac{L}{2\eta}\left(a - z_i^k + \frac{\eta g_i^k}{L}\right)^2 \right\} \\ y_i^{k+1} = \theta_1 z_i^{k+1} + \theta_2 w_i^k + (1-\theta_1-\theta_2) y_i^k. \end{cases}$$

  Suppose that

$$g_i^k = \hat{g}_i, w_i^k = \hat{w}_i \ \ \forall k = t_0, \ldots, t_1 - 1.$$

  Since the z_i^k follows the same recursive formula as (39), we can apply Algorithm 5 to compute $z_i^{t_1}$, i.e.,

$$z_i^{t_1} = \text{delayed\_update}(t_0, t_1, \hat{g}_i, z_i^{t_0}, \eta/L).$$

  Let $\theta_3 = 1 - \theta_1 - \theta_2$. Then for any integers $k \geq t_0$ and $0 < s \leq t_1 - k$

$$y_i^{k+s} = \theta_1\left(z_i^{k+s} + \theta_3 z_i^{k+s-1} + \cdots + \theta_3^{s-1} z_i^{k+1}\right) + \theta_2\left(1 + \theta_3 + \cdots + \theta_3^{s-1}\right) w_i^k + \theta_3^s y_i^k.$$

  If

$$z_i^{k+l} = q^l(z_i^k + h) - h, \ \ \forall l = 1, \ldots, s, \tag{41}$$

  for some $q > 0$ and $h \in \mathbb{R}$, then

$$\begin{aligned} y_i^{k+s} &= \theta_1\left(\sum_{l=1}^{s} \theta_3^{s-l}\left(q^l(z_i^k + h) - h\right)\right) + \frac{\theta_2(1-\theta_3^s)}{1-\theta_3} w_i^k + \theta_3^s y_i^k \\ &= \theta_1\theta_3^s(z_i^k + h)\sum_{l=1}^{s}\left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w_i^k - \theta_1 h)(1-\theta_3^s)}{1-\theta_3} + \theta_3^s y_i^k. \end{aligned} \tag{42}$$

  Based on the above computation we can write down the efficient implementation for L-Katyusha, given in Algorithm 6, and we have

$$(y_i^{t_1}, z_i^{t_1}) = \text{delayed\_update2}(t_0, t_1, \hat{g}_i, y_i^{t_0}, z_i^{t_0}, \hat{w}_i, \eta/L).$$

  It is easy to check that the computational complexity of Algorithm 6 is $\mathcal{O}(\log(t_1 - t_0))$.

A.2 Efficient Implementation for Katyusha

As mentioned, one major difference between the original Katyusha (Allen-Zhu, 2017) and our loopless variant is in the update of the reference point. Let m be the size of the inner loop in Katyusha. After s outer loops, Katyusha requires to compute a convex combination of y^k

$$\tilde{x}^{s+1} = \frac{\sum_{j=0}^{m-1}\theta^j y^{sm+j+1}}{\sum_{j=0}^{m-1}\theta^j},$$

for some $\theta > 1$. For any $1 \leq t \leq m$, define

$$\hat{x}^t = \frac{\sum_{j=0}^{t-1}\theta^j y^{sm+j+1}}{\sum_{j=0}^{t-1}\theta^j}.$$

Algorithm 6 $(\tilde{y}, \tilde{z})$ =delayed_update2(t_0, t_1, u, y, z, w, η)

if $t_1 = t_0$ **then** $\tilde{z} = z$; $\tilde{y} = y$; return;
end if
$\alpha = (1+\eta\lambda_2)^{t_0-t_1}$
$q = 1/(1+\eta\lambda_2)$
if $z = 0$ **then**
  if $\lambda_1 + u < 0$ **then** $\tilde{z} = \alpha z - (1-\alpha)(u+\lambda_1)/\lambda_2$; $h = (u+\lambda_1)/\lambda_2$
  else
    if $u - \lambda_1 > 0$ **then** $\tilde{z} = \alpha z - (1-\alpha)(u-\lambda_1)/\lambda_2$; $h = (u-\lambda_1)/\lambda_2$
    else
      $\tilde{z} = 0$; $q = 0$; $h = 0$
    end if
  end if
  $\tilde{y} = \theta_1\theta_3^{t_1-t_0}(z+h)\sum_{l=1}^{t_1-t_0}\left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w - \theta_1 h)(1-\theta_3^{t_1-t_0})}{1-\theta_3} + \theta_3^{t_1-t_0} y.$
else
  $h = (u+\lambda_1)/\lambda_2$
  if $z > 0$ **then**
    if $\lambda_1 + u \le 0$ **then**
      $\tilde{z} = \alpha z - (1-\alpha)(u+\lambda_1)/\lambda_2$
      $\tilde{y} = \theta_1\theta_3^{t_1-t_0}(z+h)\sum_{l=1}^{t_1-t_0}\left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w - \theta_1 h)(1-\theta_3^{t_1-t_0})}{1-\theta_3} + \theta_3^{t_1-t_0} y.$
    else $t = t_0 + \log\left(1 + \frac{\lambda_2 z}{\lambda_1 + u}\right) / \log(1+\eta\lambda_2)$
      if $t < t_1$ **then**
        $t' = \lfloor t \rfloor$
        $\alpha' = (1+\eta\lambda_2)^{t_0-t'}$
        $z' = \alpha' z - (1-\alpha')(u+\lambda_1)/\lambda_2$
        $y' = \theta_1\theta_3^{t'-t_0}(z+h)\sum_{l=1}^{t'-t_0}\left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w - \theta_1 h)(1-\theta_3^{t'-t_0})}{1-\theta_3} + \theta_3^{t'-t_0} y.$
        $z'' = \arg\min_{a\in\mathbb{R}}\left\{\frac{\lambda_2}{2}a^2 + \lambda_1|a| + au + \frac{1}{2\eta}(a-z')^2\right\}$
        $y'' = \theta_1 z'' + \theta_2 w + (1-\theta_1-\theta_2)y'$
        $(\tilde{y}, \tilde{z}) = \text{delayed\_update2}(t'+1, t_1, u, y'', z'', w, \eta)$
      else
        $\tilde{z} = \alpha z - (1-\alpha)(u+\lambda_1)/\lambda_2$
        $\tilde{y} = \theta_1\theta_3^{t_1-t_0}(z+h)\sum_{l=1}^{t_1-t_0}\left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w - \theta_1 h)(1-\theta_3^{t_1-t_0})}{1-\theta_3} + \theta_3^{t_1-t_0} y.$
      end if
    end if
  else
    $(\tilde{y}, \tilde{z}) = -\text{delayed\_update2}(t_0, t_1, -u, -y, -z, -w, \eta)$
  end if
end if
Output $(\tilde{y}, \tilde{z})$

Suppose that

$$g_i^{sm+k} = \hat{g}_i, \ \ \forall k = t_0, \ldots, t_1 - 1.^1$$

1. Recall that our g^k corresponds to $\tilde{\nabla}_k$ in Katyusha (Allen-Zhu, 2017).

In order to compute $\hat{x}^{t_1}$ from $\hat{x}^{t_0}$ in $\mathcal{O}(\log(t_1 - t_0))$, we consider the case when $\sigma_1 = 0$. First note that

$$\hat{x}^{t_1} = \left(1 - \frac{\theta^{t_1} - \theta^{t_0}}{\theta^{t_1} - 1}\right)\hat{x}^{t_0} + \frac{y^{sm+t_0+1} + \theta y^{sm+t_0+2} + \cdots + \theta^{t_1-t_0-1} y^{sm+t_1}}{\theta^{t_1-t_0-1} + \cdots + \theta^{-t_0}}.$$

Assume that

$$z_i^{sm+t_0+l} = q^l(z_i^{sm+t_0} + h) - h, \ \ \forall l = 1, \ldots, t_1 - t_0. \tag{43}$$

The same as (42) we have

$$\begin{aligned} & y^{sm+t_0+1} + \theta y^{sm+t_0+2} + \cdots + \theta^{t_1-t_0-1} y^{sm+t_1} \\ &= \sum_{k=1}^{t_1-t_0} \theta^{k-1} \left(\theta_1 \theta_3^k (z_i^{sm+t_0} + h) \sum_{l=1}^{k} \left(q\theta_3^{-1}\right)^l + \frac{(\theta_2 w_i^k - \theta_1 h)(1 - \theta_3^k)}{1 - \theta_3} + \theta_3^k y_i^{sm+t_0} \right). \end{aligned} \tag{44}$$

After rearranging, we can compute (44) and then $\hat{x}^{t_1}$ from $\hat{x}^{t_0}$ in $\mathcal{O}(\log(t_1 - t_0))$ time when (43) holds. Then we can update $\hat{x}^t$ in the same efficient way as we update the three inner iterates $\{x^k, y^k, z^k\}$ in Algorithm 6. We omit further details as this is not the main topic of our paper. However, the above discussion shows that the implementation of the original Katyusha is more complicated than our loopless variant, due to the use of the weighted average as the reference point.

Appendix B. Proofs of Lemma 3 and Inequality (4)

B.1 Proof of Lemma 3

Proof If $\tau = 1$, then every $i \in [n]$ can be in the same group $[n]$, and the number of groups is equal to 1. Next we consider the case where $\tau > 1$. We construct a group sampling S as follows. We partition $[n]$ into groups as follows. For the ordered sequence $p_1, \ldots, p_n$, we add them from p_1 consecutively, until the summation is greater than one at p_{i_1}. We collect $\{1, \ldots, i_1 - 1\}$ as a group C_1. In such way, $\sum_{i \in C_1} p_i$ is less than or equal to one. Next we repeat this procedure to the ordered sequence $p_{i_1}, \ldots, p_n$ until every index is divided into some group. The rest of the formation of the sampling is the same as the final step in the definition of group sampling.

Assume the number of the groups is t, and the groups we get from the above construction are ordered sets $C_1, \ldots, C_t$. According to the construction, we know

$$\sum_{i \in C_j} p_i + \sum_{i \in C_{j+1}} p_i > 1,$$

for any $1 \leq j < t$. Next, we consider two cases.

Case 1. Suppose t is even. Then $\frac{t}{2} < \tau$. If τ is an integer, then $t \leq 2\tau - 2$, otherwise, $t < 2\tau$.

Case 2. Suppose t is odd. Then $\frac{t-1}{2} < \tau$. If τ is an integer, then $t \leq 2\tau - 1$, otherwise, $t < 2\tau + 1$. ■

B.2 Proof of Inequality (4)

Proof Let $L_m = \max(L_f, |\mu_f|)$ and $g(x) = f(x) + \frac{L_m}{2}\|x\|^2,\ \forall x \in \mathbb{R}^d$. Notice that for any $x, y \in \mathbb{R}^d$

$$\frac{L_m}{2}\|x\|^2 = \frac{L_m}{2}\|y\|^2 + L_m\langle y, x-y\rangle + \frac{L_m}{2}\|x-y\|^2.$$

Then from Assumption 1, we have

$$g(y) + \langle \nabla g(y), x-y\rangle + \frac{L_m+\mu_f}{2}\|x-y\|^2 \le g(x) \le g(y) + \langle \nabla g(y), x-y\rangle + L_m\|x-y\|^2,$$

for any $x, y \in \mathbb{R}^d$. Since $L_m + \mu_f \ge 0$, we know g is convex. Then we arrive at

$$\begin{aligned}
& \|\nabla f(x) - \nabla f(y)\|^2 + 2L_m\langle \nabla f(x) - \nabla f(y), x-y\rangle + L_m^2\|x-y\|^2 \\
= \ & \|\nabla f(x) - \nabla f(y) + L_m(x-y)\|^2 \\
= \ & \|\nabla g(x) - \nabla g(y)\|^2 \\
\overset{(5)}{\le} \ & 4L_m\left(g(x) - g(y) - \langle \nabla g(y), x-y\rangle\right) \\
= \ & 4L_m\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle + \frac{L_m}{2}\|x-y\|^2\right) \\
\le \ & 4L_m\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle\right) + 2L_m^2\|x-y\|^2,
\end{aligned}$$

for any $x, y \in \mathbb{R}^d$. Cancelling the term $L_m^2\|x-y\|^2$, we can obtain

$$\begin{aligned}
& \|\nabla f(x) - \nabla f(y)\|^2 + 2L_m\langle \nabla f(x) - \nabla f(y), x-y\rangle \\
\le \ & 4L_m\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle\right) + L_m^2\|x-y\|^2,
\end{aligned}$$

and by changing the positions of x and y, we also have

$$\begin{aligned}
& \|\nabla f(x) - \nabla f(y)\|^2 + 2L_m\langle \nabla f(x) - \nabla f(y), x-y\rangle \\
\le \ & 4L_m\left(f(y) - f(x) - \langle \nabla f(x), y-x\rangle\right) + L_m^2\|x-y\|^2.
\end{aligned}$$

Adding the above two inequalities yields the desired result. ■

Appendix C. Strongly Convex Case: Proof of Theorem 6

C.1 Proof of Lemma 5

Proof By definition, we have

$$\mathbb{E}_k[\mathcal{D}^{k+1}] = \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(w^{k+1}) - \mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2 \,\middle|\, w^{k+1}\right] \,\middle|\, (x^k, w^k)\right],$$

where $S \sim \mathcal{S}$ is independent of w^{k+1}, x^k and w^k. Therefore

$$\begin{aligned}\mathbb{E}_k[\mathcal{D}^{k+1}] = \quad & p\,\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x^k)-\mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2 \;\middle|\; (x^k,w^k)\right] \\ & +(1-p)\,\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(w^k)-\mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2 \;\middle|\; (x^k,w^k)\right] \\ = \quad & p\,\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x^k)-\mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2 \;\middle|\; (x^k,w^k)\right] + (1-p)\mathcal{D}^k.\end{aligned}$$

Finally, because of the independence between x^k and S and Assumption 5, we deduce that

$$\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x^k)-\mathbf{G}(x^*)\right)\Theta_S\mathbf{I}_S e\right\|^2 \;\middle|\; (x^k,w^k)\right] \leq 2\mathcal{L}_1(f(x^k)-f(x^*)-\langle\nabla f(x^*),x^k-x^*\rangle).$$

Similarly, since S_k is independent of x^k and w^k, we have

$$\begin{aligned}& \mathbb{E}_k[\|g^k-\nabla f(x^*)\|^2] \\ = \;& \mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x^k)-\mathbf{G}(w^k)\right)\Theta_{S_k}\mathbf{I}_{S_k}e+\frac{1}{n}\mathbf{G}(w^k)e-\nabla f(x^*)\right\|^2\right] \\ \leq \;& 2\mathbb{E}_k\left[\left\|\frac{1}{n}\left(\mathbf{G}(x^k)-\mathbf{G}(x^*)\right)\Theta_{S_k}\mathbf{I}_{S_k}e\right\|^2\right] \\ & +2\mathbb{E}_k\left[\left\|\frac{1}{n}\mathbf{G}(w^k)e-\frac{1}{n}\mathbf{G}(x^*)e-\frac{1}{n}\left(\mathbf{G}(w^k)-\mathbf{G}(x^*)\right)\Theta_{S_k}\mathbf{I}_{S_k}e\right\|^2\right] \\ \leq \;& 4\mathcal{L}_1(f(x^k)-f(x^*)-\langle\nabla f(x^*),x^k-x^*\rangle)+2\mathbb{E}_k\left[\left\|\frac{1}{n}\left(\mathbf{G}(w^k)-\mathbf{G}(x^*)\right)\Theta_{S_k}\mathbf{I}_{S_k}e\right\|^2\right] \\ = \;& 4\mathcal{L}_1(f(x^k)-f(x^*)-\langle\nabla f(x^*),x^k-x^*\rangle)+2\mathcal{D}^k.\end{aligned}$$

■

Appendix D. Strongly Convex Case: Proof of Theorem 10

D.1 Proof of Lemma 9

Proof By the computation formula of z^{k+1} in Algorithm 2, there is $g \in \partial\psi(z^{k+1})$ such that

$$z^{k+1}-\frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1 x^k+z^k-\frac{\eta}{L}g^k\right)+\frac{\eta}{(1+\eta\sigma_1)L}g=0.$$

Together with $\mu_f = L\sigma_1$, we obtain

$$g^k=\frac{L}{\eta}(z^k-z^{k+1})+\mu_f(x^k-z^{k+1})-g. \tag{45}$$

Therefore, we have

$$\begin{aligned}
\langle g^k, z^{k+1}-x^*\rangle &= \mu_f\langle x^k-z^{k+1}, z^{k+1}-x^*\rangle + \frac{L}{\eta}\langle z^k-z^{k+1}, z^{k+1}-x^*\rangle - \langle g, z^{k+1}-x^*\rangle \\
&= \frac{\mu_f}{2}\left(\|x^k-x^*\|^2 - \|x^k-z^{k+1}\|^2 - \|z^{k+1}-x^*\|^2\right) \\
&\quad + \frac{L}{2\eta}\left(\|z^k-x^*\|^2 - \|z^k-z^{k+1}\|^2 - \|z^{k+1}-x^*\|^2\right) - \langle z^{k+1}-x^*, g\rangle \\
&\le \frac{\mu_f}{2}\|x^k-x^*\|^2 + \frac{L}{2\eta}\left(\|z^k-x^*\|^2 - (1+\eta\sigma_1)\|z^{k+1}-x^*\|^2\right) \\
&\quad - \frac{L}{2\eta}\|z^k-z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}) - \frac{\mu_\psi}{2}\|z^{k+1}-x^*\|^2 \\
&= \frac{\mu_f}{2}\|x^k-x^*\|^2 + \frac{L}{2\eta}\left(\|z^k-x^*\|^2 - \left(1+\frac{\eta\mu}{L}\right)\|z^{k+1}-x^*\|^2\right) \\
&\quad - \frac{L}{2\eta}\|z^k-z^{k+1}\|^2 + \psi(x^*) - \psi(z^{k+1}),
\end{aligned}$$

where the last inequality comes from $-\|x^k-z^{k+1}\|^2 \le 0$ and ψ is μ_ψ-strongly convex, the last equality comes from $\sigma_1 = \mu_f/L$ and $\mu = \mu_f + \mu_\psi$. By the definition of $\mathcal{Z}^k$, we can obtain the result. ■

D.2 Proof of Corollary 11

Proof We consider two cases:

Case 1. Suppose $L_f \le \frac{\mathcal{L}_2}{p}$. In this case, $\theta_1 = \min\left(\sqrt{\frac{\mu}{\mathcal{L}_{2p}}}\theta_2, \theta_2\right)$ and $\theta_2 = \frac{\mathcal{L}_2}{2\max(L_f, \mathcal{L}_2)} \ge \frac{p}{2}$.

Case 1.1. Suppose $\frac{\mu}{\mathcal{L}_{2p}} \ge 1$. In this subcase, $\theta_1 = \theta_2$ and thus $\frac{\mu}{3\theta_1 L+\mu} = \frac{2\mu}{3\mathcal{L}_2+2\mu} \ge \frac{2p}{3+2p} \ge \frac{2p}{5}$. By choosing $q = \frac{2}{3}$, we have $\theta_1+\theta_2-\frac{\theta_2}{q} = \frac{\theta_2}{2} \ge \frac{p}{4}$ and $p(1-q) = \frac{p}{3}$.

Case 1.2. Suppose $\frac{\mu}{\mathcal{L}_{2p}} < 1$. In this subcase, $\theta_1 = \sqrt{\frac{\mu}{\mathcal{L}_{2p}}}\theta_2$ and $\frac{\mu}{3\theta_1 L+\mu} = \frac{2\mu}{3\sqrt{\mu\mathcal{L}_2/p}+2\mu} \ge \frac{2}{5}\sqrt{\frac{\mu p}{\mathcal{L}_2}}$. By choosing $q = 1-\frac{1}{3}\sqrt{\frac{\mu}{\mathcal{L}_{2p}}} \ge \frac{2}{3}$ so that $3\theta_2(1-q) = \theta_1$, we have $\theta_1+\theta_2-\frac{\theta_2}{q} = \theta_1(1-\frac{1}{3q}) \ge \frac{\theta_1}{2} = \frac{\theta_2}{2}\sqrt{\frac{\mu}{\mathcal{L}_{2p}}} \ge \frac{1}{4}\sqrt{\frac{\mu p}{\mathcal{L}_2}}$ and $p(1-q) = \frac{1}{3}\sqrt{\frac{\mu p}{\mathcal{L}_2}}$.

Case 2. Suppose $L_f > \frac{\mathcal{L}_2}{p}$. In this case, $\theta_1 = \min\left(\sqrt{\frac{\mu}{L_f}}, \frac{p}{2}\right)$, $L = L_f$, and $\theta_2 = \frac{\mathcal{L}_2}{2L_f} < \frac{p}{2}$.

Case 2.1. Suppose $\sqrt{\frac{\mu}{L_f}} \ge \frac{p}{2}$. In this subcase, $\theta_1 = \frac{p}{2}$ and $\frac{\mu}{3\theta_1 L+\mu} = \frac{2\mu}{3pL_f+2\mu} \ge \frac{p}{6+p} \ge \frac{p}{7}$. Let $q = \frac{2}{3}$. Then $\theta_1+\theta_2-\frac{\theta_2}{q} = \theta_1-\frac{\theta_2}{2} > \frac{p}{4}$ and $p(1-q) = \frac{p}{3}$.

Case 2.2. Suppose $\sqrt{\frac{\mu}{L_f}} < \frac{p}{2}$. In this subcase, $\theta_1 = \sqrt{\frac{\mu}{L_f}}$ and $\frac{\mu}{3\theta_1 L+\mu} = \frac{\mu}{3\sqrt{\mu L_f}+\mu} \ge \frac{1}{4}\sqrt{\frac{\mu}{L_f}}$. Let $q = 1-\frac{2}{3p}\sqrt{\frac{\mu}{L_f}} > \frac{2}{3}$ so that $1-q = \frac{2}{3p}\theta_1$. Then $\theta_1+\theta_2-\frac{\theta_2}{q} = \theta_1-\frac{2}{3pq}\theta_1\theta_2 > (1-\frac{1}{3q})\theta_1 > \frac{\theta_1}{2} = \frac{1}{2}\sqrt{\frac{\mu}{L_f}}$ and $p(1-q) = \frac{2}{3}\sqrt{\frac{\mu}{L_f}}$. ■

D.3 Proof of Corollary 12

Proof By Theorem 10, after running k iterations L-Katyusha, we can get a vector w^k satisfying

$$\frac{\theta_2}{pq\theta_1}\mathbb{E}\Big[P(w^k)-P^*\Big] \leq \rho^k\left(\frac{L+\eta\mu}{2\eta}\|x^0-x^*\|^2+\left(\frac{1}{\theta_1}+\frac{\theta_2}{pq\theta_1}\right)\big(P(x^0)-P^*\big)\right),$$

with $\rho := 1-\min\left(\frac{\mu}{\mu+3\theta_1 L}, \theta_1+\theta_2-\frac{\theta_2}{q}, p(1-q)\right)$. Consequently

$$\begin{aligned}\mathbb{E}\Big[P(w^k)-P^*\Big] &\leq \rho^k\left(\frac{(L+\eta\mu)pq\theta_1}{\eta\mu\theta_2}+\frac{pq}{\theta_2}+1\right)\big(P(x^0)-P^*\big)\\ &= \rho^k\left(\frac{3Lpq\theta_1^2}{\mu\theta_2}+\frac{pq\theta_1}{\theta_2}+\frac{pq}{\theta_2}+1\right)\big(P(x^0)-P^*\big).\end{aligned}$$

Case 1. Suppose $L_f \leq \frac{\mathcal{L}_2}{p}$. In this case, $\mathcal{L}_2 p\theta_1^2 \leq \mu\theta_2^2$, $\theta_1 \leq \theta_2$ and thus

$$\frac{3Lpq\theta_1^2}{\mu\theta_2} \leq \frac{3Lq\theta_2}{\mathcal{L}_2} = \frac{3q}{2}, \qquad \frac{pq\theta_1}{\theta_2} \leq p.$$

In addition,

$$\frac{p}{\theta_2} = \frac{2Lp}{\mathcal{L}_2} \leq \max\left(2p, \frac{2L_f p}{\mathcal{L}_2}\right) \leq 2.$$

Hence

$$\frac{3Lpq\theta_1^2}{\mu\theta_2}+\frac{pq\theta_1}{\theta_2}+\frac{pq}{\theta_2}+1 \leq \frac{3p}{2}+p+2+1 \leq 6.$$

Case 2. Suppose $L_f > \frac{\mathcal{L}_2}{p}$. In this case, $L = L_f$, $\theta_1 \leq \sqrt{\frac{\mu}{L_f}}$ and thus

$$\frac{3Lpq\theta_1^2}{\mu\theta_2} = \frac{3L_f pq\theta_1^2}{\mu\theta_2} \leq \frac{3p}{\theta_2}.$$

In addition, $\theta_2 = \frac{\mathcal{L}_2}{2L_f} < \frac{p}{2}$ and hence

$$\frac{3Lpq\theta_1^2}{\mu\theta_2}+\frac{pq\theta_1}{\theta_2}+\frac{pq}{\theta_2}+1 \leq \frac{6p}{\theta_2} = \frac{12pL_f}{\mathcal{L}_2}.$$

Then we conclude from Corollary 11. ■

D.4 Proof of Corollary 13

Proof It suffices to note that for any $\beta \in (0,1)$ and $\alpha > 0$, we have

$$\alpha(1-\beta)^k \leq \frac{1}{4}, \ \ \forall k \geq \frac{\ln(4\alpha)}{\beta}.$$

■

Appendix E. Nonconvex and Smooth Case: Proof of Theorem 22

E.1 Proof of Lemma 20

Proof We have

$$\begin{aligned}\mathbb{E}_k\left[\|g^k\|^2\right] &= \mathbb{E}_k\left[\|g^k-\nabla f(x^k)+\nabla f(x^k)\|^2\right]\\ &\le 2\|\nabla f(x^k)\|^2+2\mathbb{E}_k\left[\|g^k-\nabla f(x^k)\|^2\right]\\ &\overset{\text{Assumption 7}}{\le} 2\|\nabla f(x^k)\|^2+2\mathcal{L}_3\|x^k-w^k\|^2.\end{aligned}$$

■

E.2 Proof of Lemma 21

Proof First, note that

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1}-w^{k+1}\|^2\right] &= p\mathbb{E}_k\left[\|x^{k+1}-x^k\|^2\right]+(1-p)\mathbb{E}_k\left[\|x^{k+1}-w^k\|^2\right]\\ &= p\eta^2\mathbb{E}_k\left[\|g^k\|^2\right]+(1-p)\mathbb{E}_k\left[\|x^{k+1}-w^k\|^2\right].\end{aligned}$$

For $\mathbb{E}_k\left[\|x^{k+1}-w^k\|^2\right]$, we have

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1}-w^k\|^2\right] &= \mathbb{E}_k\left[\|x^k-\eta g^k-w^k\|^2\right]\\ &= \|x^k-w^k\|^2+\eta^2\mathbb{E}_k\left[\|g^k\|^2\right]-2\eta\mathbb{E}_k\left[\langle x^k-w^k,\nabla f(x^k)\rangle\right]\\ &\le \|x^k-w^k\|^2+\eta^2\mathbb{E}_k\left[\|g^k\|^2\right]+\eta\left(\frac{1}{\beta}\|\nabla f(x^k)\|^2+\beta\|x^k-w^k\|^2\right)\\ &= (1+\eta\beta)\|x^k-w^k\|^2+\eta^2\mathbb{E}_k\left[\|g^k\|^2\right]+\frac{\eta}{\beta}\|\nabla f(x^k)\|^2,\end{aligned}$$

where the inequality is from $|2\langle a,b\rangle|\le\frac{1}{\beta}\|a\|^2+\beta\|b\|^2$ for any $\beta>0$. Combining all the above results, we can obtain the result. ■

E.3 Proof of Corollary 23

Proof If the stepsize η satisfies (27), then from Theorem 22, we have

$$\mathbb{E}\left[\|\nabla f(x^k)\|\right]\le\frac{4}{\eta}\left(\mathbb{E}\left[\Upsilon^k\right]-\mathbb{E}[\Upsilon^{k+1}]\right),$$

which implies that

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x^a)\|^2\right] &= \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\|\nabla f(x^i)\|^2\right]\\ &\leq \frac{1}{k+1}\cdot\frac{4}{\eta}\left(\Upsilon^0-\mathbb{E}\left[\Upsilon^{k+1}\right]\right)\\ &= \frac{1}{k+1}\cdot\frac{4}{\eta}\left(f(x^0)-\mathbb{E}\left[f(x^{k+1})\right]-\alpha\mathbb{E}\left[\|x^{k+1}-w^{k+1}\|^2\right]\right)\\ &\leq \frac{4}{\eta}\cdot\frac{f(x^0)-f(x^*)}{k+1}.\end{aligned}$$

■

Appendix F. Proofs in Section 6

F.1 Proof of Lemma 24

Proof From $\|\nabla f_S(x)-\nabla f_S(y)\|^2\leq 2L_S(f_S(x)-f_S(y)-\langle\nabla f_S(y),x-y\rangle)$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))\Theta_S\mathbf{I}_S e\right\|^2\right] &\leq 2\sum_C p_C L_C(f_C(x)-f_C(y)-\langle\nabla f_C(y),x-y\rangle)\\ &= 2\sum_C p_C L_C\sum_{i\in C}\frac{\Theta_C^i}{n}(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle)\\ &= \frac{2}{n}\sum_{i\in[n]}\sum_{C:i\in C}p_C L_C\Theta_C^i(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle)\\ &\leq \frac{2}{n}\mathcal{L}_{\max}\sum_{i\in[n]}(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle)\\ &= 2\mathcal{L}_{\max}(f(x)-f(y)-\langle\nabla f(y),x-y\rangle).\end{aligned}$$

Hence $\mathcal{L}_1\leq\mathcal{L}_{\max}=\max_{i\in[n]}\sum_{C:i\in C}p_C L_C\Theta_C^i$. For all i, since $L_S\leq\frac{1}{n}\sum_{i\in S}L_i\Theta_S^i$, we have

$$\sum_{C:i\in C}p_C L_C\Theta_C^i \leq \sum_{C:i\in C}p_C\Theta_C^i\frac{1}{n}\sum_{j\in C}L_j\Theta_C^j=\frac{1}{n}\sum_{j\in[n]}\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j L_j,$$

which implies

$$\mathcal{L}_{\max}\leq\frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j L_j\right\}.$$

Furthermore, if $\Theta_C^i=\frac{1}{p_i}$ for all i and C, then

$$\mathcal{L}_{\max}\leq\frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j L_j\right\}=\frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\frac{\mathbf{P}_{ij}}{p_i p_j}L_j\right\}.$$

For $\mathcal{L}_2$, since $\mathbb{E}\left[\|X-\mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))\Theta_S\mathbf{I}_S e-\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))e\right\|^2\right] \leq \mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))\Theta_S\mathbf{I}_S e\right\|^2\right]. \tag{46}$$

Then we get the same upper bound for $\mathcal{L}_2$. ■

F.2 Proof of Lemma 25

Proof First, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))\Theta_S\mathbf{I}_S e\right\|^2\right] \\
=\ & \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i\in S}\Theta_S^i(\nabla f_i(x)-\nabla f_i(y))\right\|^2\right] \\
=\ & \frac{1}{n^2}\sum_C p_C\left\langle\sum_{i\in C}\Theta_C^i(\nabla f_i(x)-\nabla f_i(y)),\sum_{i\in C}\Theta_C^i(\nabla f_i(x)-\nabla f_i(y))\right\rangle \\
=\ & \frac{1}{n^2}\sum_C p_C\sum_{i,j\in C}\Theta_C^i\Theta_C^j\left\langle\nabla f_i(x)-\nabla f_i(y),\nabla f_j(x)-\nabla f_j(y)\right\rangle \\
=\ & \frac{1}{n^2}\sum_{i,j=1}^n\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j\left\langle\nabla f_i(x)-\nabla f_i(y),\nabla f_j(x)-\nabla f_j(y)\right\rangle \qquad (47)\\
\leq\ & \frac{1}{n^2}\sum_{i,j=1}^n\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j L_iL_j\|x-y\|^2.
\end{aligned}$$

This along with (46) implies the result. If $\Theta_C^i=\frac{1}{p_i}$, then

$$\sum_{C:i,j\in C}p_C\Theta_C^i\Theta_C^j=\frac{\mathbf{P}_{ij}}{p_ip_j}.$$

■

F.3 Proof of Theorem 26

Proof Let any $x,y\in\mathbb{R}^d$. For ease of notation, we denote $X=\frac{1}{n}\left(\mathbf{G}(x)-\mathbf{G}(y)\right)\Theta_S\mathbf{I}_S e$. In view of (3), we have

$$\mathbb{E}[X]=\frac{1}{n}\left(\mathbf{G}(x)-\mathbf{G}(y)\right)e=\nabla f(x)-\nabla f(y),$$

and therefore

$$\mathbb{E}\left[\left\|\frac{1}{n}\left(\mathbf{G}(x)-\mathbf{G}(y)\right)\Theta_S\mathbf{I}_S e-\frac{1}{n}\left(\mathbf{G}(x)-\mathbf{G}(y)\right)e\right\|^2\right]=\mathbb{E}\left[\|X-\mathbb{E}[X]\|^2\right].$$

Applying (30) with $\mathbf{M} = \mathbf{G}(x) - \mathbf{G}(y)$, we obtain directly

$$\mathbb{E}\left[\|X\|^2\right] \leq \sum_{i=1}^{n} \mathcal{A}_i \left\| \frac{1}{n} \left(\nabla f_i(x) - \nabla f_i(y)\right)\right\|^2 + \mathcal{B}\left\|\nabla f(x) - \nabla f(y)\right\|^2,$$

and thus

$$\mathbb{E}\left[\|X\|^2\right] - \|\mathbb{E}[X]\|^2 \leq \sum_{i=1}^{n} \mathcal{A}_i \left\| \frac{1}{n} \left(\nabla f_i(x) - \nabla f_i(y)\right)\right\|^2 + \max\left(\mathcal{B}-1, 0\right)\left\|\nabla f(x) - \nabla f(y)\right\|^2.$$

Then by Assumption (8) and (4), we deduce

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \left(\frac{1}{n^2}\sum_{i=1}^{n} \mathcal{A}_i L_i^2 + \max(\mathcal{B}-1, 0)\max\left(L_f, |\mu_f|\right)\right)\|x-y\|^2.$$

If in addition, f_i is convex for each $i \in [n]$, then by (5) and (28),

$$\begin{aligned}\mathbb{E}\left[\|X\|^2\right] &\leq \sum_{i=1}^{n} \frac{2\mathcal{A}_i L_i}{n^2}\left(f_i(x) - f_i(y) - \langle \nabla f_i(y), x-y\rangle\right) \\ &\quad + 2\mathcal{B}L_f\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle\right) \\ &\leq 2\left(\frac{1}{n}\max_i \mathcal{A}_i L_i + \mathcal{B}L_f\right)\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle\right).\end{aligned}$$

Similarly, we have

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq 2\left(\frac{1}{n}\max_i \mathcal{A}_i L_i + \max\left(\mathcal{B}-1, 0\right)L_f\right)\left(f(x) - f(y) - \langle \nabla f(y), x-y\rangle\right).$$

■

F.4 Proof of Lemma 29

Proof By (31) we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{M}\Theta_S\mathbf{\Pi}_S e\|^2] &= \sum_{i,j=1}^{n} \frac{\mathbf{P}_{ij}}{p_i p_j}\langle \mathbf{M}_{:i}, \mathbf{M}_{:j}\rangle = \sum_{i\neq j} \frac{\mathbf{P}_{ij}}{p_i p_j}\langle \mathbf{M}_{:i}, \mathbf{M}_{:j}\rangle + \sum_{i=1}^{n}\frac{1}{p_i}\|\mathbf{M}_{:i}\|^2 \\ &= \sum_{i,j=1}^{n}\langle \mathbf{M}_{:i}, \mathbf{M}_{:j}\rangle - \sum_{k=1}^{t}\sum_{i,j\in C_k}\langle \mathbf{M}_{:i}, \mathbf{M}_{:j}\rangle + \sum_{i=1}^{n}\frac{1}{p_i}\|\mathbf{M}_{:i}\|^2 \\ &\leq \|\mathbf{M}e\|^2 - \sum_{i\in\mathcal{I}}\|\mathbf{M}_{:i}\|^2 + \sum_{i=1}^{n}\frac{1}{p_i}\|\mathbf{M}_{:i}\|^2.\end{aligned}$$

■

F.5 Proof of Lemma 32

Proof Since ϕ_i is $1/\gamma$-smooth, we have

$$\|\nabla\phi_i(\tilde{x}) - \nabla\phi_i(\tilde{y})\|^2 \leq \frac{2}{\gamma}\left(\phi_i(\tilde{x}) - \phi_i(\tilde{y}) - \langle\nabla\phi_i(\tilde{y}), \tilde{x} - \tilde{y}\right).$$

Letting $\tilde{x} = \mathbf{A}_i^\top x$, and $\tilde{y} = \mathbf{A}_i^\top y$ in the above inequlity yields

$$\begin{aligned}\left\|\nabla\phi_i(\mathbf{A}_i^\top x) - \nabla\phi_i(\mathbf{A}_i^\top y)\right\|^2 &\leq \frac{2}{\gamma}\left(\phi_i(\mathbf{A}_i^\top x) - \phi_i(\mathbf{A}_i^\top y) - \langle\nabla\phi_i(\mathbf{A}_i^\top y), \mathbf{A}_i^\top x - \mathbf{A}_i^\top y\right)\\ &= \frac{2}{\gamma}\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x - y\rangle\right).\end{aligned}$$

■

F.6 Proof of Lemma 33

Proof First, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x) - \mathbf{G}(y))\Theta_S\mathbf{I}_S e\right\|^2\right]\\ =\ &\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i\in S}\Theta_S^i(\nabla f_i(x) - \nabla f_i(y))\right\|^2\right]\\ =\ &\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i\in S}\frac{1}{p_i}\mathbf{A}_i(\nabla\phi_i(\mathbf{A}_i^\top x) - \nabla\phi_i(\mathbf{A}_i^\top y))\right\|^2\right]\\ \overset{(32)}{\leq}\ &\frac{1}{n^2}\sum_{i=1}^n p_i v_i\cdot\frac{1}{p_i^2}\left\|\nabla\phi_i(\mathbf{A}_i^\top x) - \nabla\phi_i(\mathbf{A}_i^\top y)\right\|^2 \qquad (48)\\ \overset{\textit{Lemma 32}}{\leq}\ &\frac{1}{n^2}\sum_{i=1}^n\frac{v_i}{p_i}\cdot\frac{2}{\gamma}\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x - y\rangle\right)\\ \leq\ &\frac{2}{n\gamma}\max_i\frac{v_i}{p_i}\cdot\frac{1}{n}\sum_{i=1}^n\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x - y\rangle\right)\\ =\ &\frac{2}{n\gamma}\max_i\frac{v_i}{p_i}\left(f(x) - f(y) - \langle\nabla f(y), x - y\rangle\right).\end{aligned}$$

This along with (46) implies the results. ■

F.7 Proof of Lemma 34

Proof From (48), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{G}(x)-\mathbf{G}(y))\Theta_S\mathbf{I}_S e\right\|^2\right] \\
\leq\ & \frac{1}{n^2}\sum_{i=1}^n p_i v_i\cdot\frac{1}{p_i^2}\left\|\nabla\phi_i(\mathbf{A}_i^\top x)-\nabla\phi_i(\mathbf{A}_i^\top y)\right\|^2 \\
\leq\ & \frac{1}{n^2}\sum_{i=1}^n \frac{v_i}{p_i}\cdot\frac{\|\mathbf{A}_i\|^2}{\gamma^2}\|x-y\|^2.
\end{aligned}$$

This along with (46) implies the result. ■

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *The Journal of Machine Learning Research*, volume 18(1), pages 8194–8244, 2017.

Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems*, NIPS'16, pages 1614–1622, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL `http://dl.acm.org/citation.cfm?id=3157096.3157277`.

Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2017.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1646–1654. Curran Associates, Inc., 2014.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/qian19b.html`.

Filip Hanzely and Peter Richtárik. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *International Conference on Artificial Intelligence and Statistics*, pages 304–312. PMLR, 2019.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

Samuel Horváth and Peter Richtárik. Nonconvex variance reduced optimization with arbitrary sampling. In *International Conference on Machine Learning*, pages 2781–2789. PMLR, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *arXiv: 1901.08689*, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Kluwer Academic Publishers, 2004.

Xun Qian, Zheng Qu, and Peter Richtárik. SAGA with arbitrary sampling. In *International Conference on Machine Learning*, pages 5190–5199. PMLR, 2019.

Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016a.

Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016b.

Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 865–873. Curran Associates, Inc., 2015.

Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

Fanhua Shang, Kaiwen Zhou, Hongying Liu, James Cheng, Ivor W Tsang, Lijun Zhang, Dacheng Tao, and Licheng Jiao. VR-SGD: A simple stochastic variance reduction method for machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 32(1): 188–202, 2018.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM Journal on Optimization*, 2008.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *J. Mach. Learn. Res.*, 18(1):2939–2980, January 2017. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=3122009.3176828`.