Don't Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop

Dmitry Kovalev [1] **Samuel Horváth** [1] **Peter Richtárik** [1 2 3]

Abstract

The stochastic variance-reduced gradient method (SVRG) and its accelerated variant (Katyusha) have attracted enormous attention in the machine learning community in the last few years due to their superior theoretical properties and empirical behaviour on training supervised machine learning models via the empirical risk minimization paradigm. A key structural element in both of these methods is the inclusion of an outer loop at the beginning of which a full pass over the training data is made in order to compute the exact gradient, which is then used to construct a variance-reduced estimator of the gradient. In this work we design *loopless variants* of both of these methods. In particular, we remove the outer loop and replace its function by a coin flip performed in each iteration designed to trigger, with a small probability, the computation of the gradient. We prove that the new methods enjoy the same superior theoretical convergence properties as the original methods. However, we demonstrate through numerical experiments that our methods have substantially superior practical behavior.

1. Introduction

Empirical risk minimization (aka finite-sum) problems form the dominant paradigm for training supervised machine learning models such as ridge regression, support vector machines, logistic regression, and neural networks. In its most general form, a finite sum problem has the form

$$\min_{x\in\mathbb{R}^d}\left\{f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x)\right\}, \qquad (1)$$

where n refers to the number of training data points (e.g., videos, images, molecules, text corpora), x is the vector representation of a model using d features, and $f_i(x)$ is the loss of model x on data point i.

1.1. Variance-reduced methods

One of the most remarkable algorithmic breakthroughs in recent years was the development of *variance-reduced* stochastic gradient algorithms for solving (1). These methods are significantly faster than SGD (Nemirovsky & Yudin, 1983; Nemirovski et al., 2009; Takáč et al., 2013) in theory and practice on convex and strongly convex problems, and faster in theory on several classes on nonconvex problems (unfortunately, variance-reduced methods are not, as of yet, state-of-the-art methods for training production-grade neural networks).

Two of the most notable and popular methods belonging to the family of variance-reduced methods are SVRG (Johnson & Zhang, 2013) and its accelerated variant known as Katyusha (Allen-Zhu, 2017). The latter method accelerates the former via the employment of a novel "negative momentum" idea. Both of these methods have a double loop design. At the beginning of the outer loop, a full pass over the training data is made to compute the gradient of f at a reference point w^k, which is chosen as the freshest iterate (SVRG) or a weighted average of recent iterates (for Katyusha). This gradient is then used in the inner loop to *adjust* the stochastic gradient $\nabla f_i(x^k)$, where i is sampled uniformly at random from $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$, and x^k is the current iterate, so as to reduce its variance. In particular, both SVRG and Katyusha perform the adjustment

$$g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k).$$

Note that, like $\nabla f_i(x^k)$, the new search direction g^k is an unbiased estimator of $\nabla f(x^k)$. Indeed,

$$\mathrm{E}\left[g^k\right] = \nabla f(x^k) - \nabla f(w^k) + \nabla f(w^k) = \nabla f(x^k). \quad (2)$$

where the expectation is taken over random choice of $i \in [n]$. However, it turns out that as the methods progress, the variance of g^k, unlike that of $\nabla f_i(x^k)$, progressively decreases to zero. The total effect of this is significantly faster convergence.

[1]King Abdullah University of Science and Technology, Kingdom of Saudi Arabia [2]University of Edinburgh, United Kingdom [3]Moscow Institute of Physics and Technology, Russian Federation. Correspondence to: Dmitry Kovalev <dmitry.kovalev@kaust.edu.sa>, Samuel Horváth <samuel.horvath@kaust.edu.sa>, Peter Richtárik <peter.richtarik@kaust.edu.sa>.

January 25, 2019

1.2. Converegnce of SVRG and Katyusha for L–smooth and μ–strongly convex functions

For instance, consider the regime where f_i is L–smooth for each i, and f is μ–strongly convex:

Assumption 1.1 (L–smoothness)**.** Functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are L–smooth for some $L > 0$. That is, for all $x, y \in \mathbb{R}^d$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 . \quad (3)$$

Assumption 1.2 (μ–strong convexity)**.** Function $f : \mathbb{R}^d \to \mathbb{R}$ is μ–strongly convex for $\mu > 0$. That is, for all $x, y \in \mathbb{R}^d$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 . \quad (4)$$

In this regime, the iteration complexity of SVRG is

$$\mathcal{O}\left(\left(n + \frac{L}{\mu}\right) \log \frac{1}{\epsilon}\right),$$

which is a vast improvement on the linear rate of gradient descent (GD), which is $\mathcal{O}\left(\frac{nL}{\mu} \log \frac{1}{\epsilon}\right)$, and on the sublinear rate of SGD, which is $\mathcal{O}(\frac{L}{\mu} + \frac{\sigma^2}{\mu^2 \epsilon})$, where $\sigma^2 = \frac{1}{n} \sum_i \|\nabla f_i(x^*)\|^2$ and x^* is the (necessarily unique) minimizer of f. On the other hand, Katyusha enjoys the *accelerated* rate

$$\mathcal{O}\left(\left(n + \sqrt{\frac{nL}{\mu}}\right) \log \frac{1}{\epsilon}\right),$$

which is superior to that of SVRG in the ill-conditioned regime where $\frac{L}{\mu} \geq n$. This rate has been shown to be *optimal* in a certain precise sense (Nesterov, 2013).

In the past several years, an enormous effort of the machine learning and optimization communities was exerted into designing new efficient variance-reduced methods algorithms to tackle problem (1). These developments have brought about a renaissance in the field. The historically first provably variance-reduced method, the stochastic average gradient (SAG) method of (Roux et al., 2012; Schmidt et al., 2017), was awarded the Lagrange prize in continuous optimization in 2018. The SAG method was later modified to an unbiased variant called SAGA (Defazio et al., 2014a), achieving the same theoretical rates.

Alternative variance-reduced method include MISO (Mairal, 2015), FINITO (Defazio et al., 2014b), SDCA (Shalev-Shwartz, 2016), dfSDCA (Csiba & Richtárik, 2015), AdaSDCA (Csiba et al., 2015), QUARTZ (Qu et al., 2015), SBFGS (Gower et al., 2016), SDNA (Qu et al., 2016), SARAH (Nguyen et al., 2017) and S2GD (Konečný & Richtárik, 2017), mS2GD (Konečný et al., 2016), RBCN (Doikov & Richtárik, 2018), JacSketch (Gower et al., 2018) and SAGD (Bibi et al., 2018). Accelerated variance-reduced method were developed by Shalev-Shwartz & Zhang (2014), Defazio (2016), Zhou (2018) and Zhou et al. (2018).

2. Contributions

As explained in the introduction, a trade-mark structural feature of SVRG and its accelerated variant, Katyusha, is the presence of the outer loop in which a full pass over the data is made. However, the presence of this outer loop is the source of several issues. First, the methods are harder to analyze. Second, one needs to decide at which point the inner loop is terminated and the outer loop entered. For SVRG, the theoretically optimal inner loop size depends on both L and μ. However, μ is not always known. Moreover, even when an estimate is available, as is the case in regularized problems with an explicit strongly convex regularizer, the estimate can often be very loose. Because of these issues, one often chooses the inner loop size in a suboptimal way, such as by setting it to n or $\mathcal{O}(n)$.

2.1. Two loopless methods

In this paper we address the above issues by developing *loopless* variants of both SVRG and Katyusha; we refer to them as L-SVRG and L-Katyusha, respectively. In these methods, we dispose of the outer loop and replace its role by a *biased coin-flip*, to be performed in every step of the methods, used to trigger the computation of the gradient $\nabla f(w^k)$ via a pass over the data. In particular, in each step, with (a small) probability $p > 0$ we perform a full pass over data and update the reference gradient $\nabla f(w^k)$. With probability $1 - p$ we keep the previous reference gradient. This procedure can alternatively be interpreted as *having an outer loop of a random length*. However, the resulting methods are easier to write down, comprehend and analyze.

2.2. Fast rates are preserved

We show that L-SVRG and L-Katyusha enjoy the same fast theoretical rates as their loopy forefathers. However, the proofs are different and the complexity results more insightful.

Convergence of L-SVRG. For L-SVRG with fixed stepsize $\eta = \frac{1}{6L}$ and probability $p = \frac{1}{n}$, we show (see Theorem 4.5) that for the Lyapunov function

$$\Phi^k = \left\|x^k - x^*\right\|^2 + \frac{4\eta^2}{pn} \sum_{i=1}^{n} \left\|\nabla f_i(w^k) - \nabla f_i(x^*)\right\|^2 . \quad (5)$$

we get $\mathrm{E}\left[\Phi^k\right] \leq \epsilon\Phi^0$ as long as

$$k = \mathcal{O}\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right).$$

In contrast, the classical SVRG result shows convergence of the expected functional suboptimality $\mathrm{E}\left[f(x^k) - f(x^*)\right]$ to zero at the same rate. Note that the classical result follows from our result by utilizing the inequality

$$f(x) - f(x^*) \leq \frac{L}{2}\|x^k - x^*\|^2,$$

which is a simple consequence of L–smoothness. However, our result is provides a deeper insight into the behavior of the method. In particular, it follows that the gradients $\nabla f_i(w^k)$ at the reference points w^k converge to the gradients at the optimum. This is a key intuition behind the workings of SVRG, one not revealed by the classical analysis. Hereby we close the gap in the theoretical understanding of the the SVRG convergence mechanism.

Our theory predicts that as log as

$$p \in \left[\min\left\{\frac{c}{n}, \frac{c\mu}{L}\right\}, \max\left\{\frac{c}{n}, \frac{c\mu}{L}\right\}\right], \tag{6}$$

where $c = \Theta(1)$, L-SVRG will enjoy the optimal complexity $\mathcal{O}\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$. In the ill-conditioned regime $\frac{L}{\mu} \gg n$, for instance, we roughly have $p \in [\frac{\mu}{L}, \frac{1}{n}]$. This is in contrast with the (loopy/standard) SVRG method the outer loop of which needs to be of the size $\approx \frac{L}{\mu}$. To the best of our knowledge, SVRG does not enjoy this rate for an outer loop of size n, even though this is the setting most often used in practice.

Convergence of L-Katyusha. For L-Katyusha with stepsize $\eta = \frac{\theta_2}{(1+\theta_2)\theta_1}$ we show convergence of the Lyapunov function

$$\Psi^k = \mathcal{Z}^k + \mathcal{Y}^k + \mathcal{W}^k, \tag{7}$$

where

$$\mathcal{Z}^k = \frac{L(1+\eta\sigma)}{2\eta}\left\|z^k - x^*\right\|^2, \ \mathcal{Y}^k = \frac{1}{\theta_1}(f(y^k) - f(x^*)), \tag{8}$$

and

$$\mathcal{W}^k = \frac{\theta_2(1+\theta_1)}{p\theta_1}(f(w^k) - f(x^*)), \tag{9}$$

and where x^k, y^k and w^k are iterates produced by the method, with the parameters defined by $\sigma = \frac{\mu}{L}$, $\theta_1 = \min\left\{\sqrt{\frac{2\sigma n}{3}}, \frac{1}{2}\right\}$, $\theta_2 = \frac{1}{2}$, $p = \frac{1}{n}$. Our main result (Theorem 5.6) states that $\mathrm{E}\left[\Psi^k\right] \leq \epsilon\Psi^0$ as long as

$$k = \mathcal{O}\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right).$$

2.3. Simplified Analysis

An advantage of the loopless approach is that the *analysis of the decrease of the Lyapunov function over a single iteration is sufficient to establish convergence*. In contrast, one needs to perform elaborate aggregation across the inner loop to prove the convergence of the original loopy methods.

2.4. Superior empirical behaviour

We show through extensive numerical testing on both synthetic and real data that our loopless methods are superior to their loopy variants.

We show through experiments that L-SVRG is *very robust to the choice of p from the optimal interval* (6) predicted by our theory. Moreover, *even the worst case for L-SVRG outperforms the best case for SVRG*. This shows how further randomization can significantly speed up and stabilize the algorithm.

3. Notations

Throughout the whole paper we use conditional expectation $\mathrm{E}\left[\mathcal{X} \mid x^k, w^k\right]$ for L-SVRG and $\mathrm{E}\left[\mathcal{X} \mid y^k, z^k, w^k\right]$ for L-Katyusha, but for simplicity we will denote these expectations as $\mathrm{E}\left[\mathcal{X}\right]$. If $\mathrm{E}\left[\mathcal{X}\right]$ refers to unconditional expectation, it is directly mentioned.

4. Loopless SVRG (L-SVRG)

In this section we describe in detail the Loopless SVRG method (L-SVRG), and its convergence properties.

4.1. The algorithm

The L-SVRG method, formalized as Algorithm 1, is inspired by the original SVRG (Johnson & Zhang, 2013) method. We remove the outer loop present in SVRG and instead use a probabilistic update of the full gradient.[1] This update can be also seen in a way that outer loop size is generated by geometric distribution similar to (Konečný & Richtárik, 2017; Lei et al., 2017).

Note that the reference point w^k (at which a full gradient is computed) is updated in each iteration with probability p to the current iterate x^k, and is left unchanged with probability $1 - p$. Alternatively, the probability p can be seen as a parameter that controls the expected time before next full pass over data. To be more precise, the expected time before next full pass over data is $1/p$. Intuitively, we wish to keep p small so that full passes over data are computed rarely enough. As we shall see next, the simple choice $p = \frac{1}{n}$

[1] This idea was indepdentnly explored by Hofmann et al. (2015); we have learned about this work after a first draft of our paper was finished.

Algorithm 1 Loopless SVRG (L-SVRG)

Parameters: stepsize $\eta > 0$, probability $p \in (0,1]$
Initialization: $x^0 = w^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
  Sample $i \in \{1, \ldots, n\}$ uniformly at random
  $g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$
  $x^{k+1} = x^k - \eta g^k$
  $w^{k+1} = \begin{cases} x^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}$
end for

leads to complexity identical to that of original SVRG.

4.2. Convergence theory

A key role in the analysis is played by the *gradient learning* quantity

$$\mathcal{D}^k \stackrel{\text{def}}{=} \frac{4\eta^2}{pn} \sum_{i=1}^n \left\|\nabla f_i(w^k) - \nabla f_i(x^*)\right\|^2 \tag{10}$$

and the Lyapunov function

$$\Phi^k \stackrel{\text{def}}{=} \left\|x^k - x^*\right\|^2 + \mathcal{D}^k. \tag{11}$$

The analysis involves four lemmas, followed by the main theorem. We wish to mention the lemmas as they highlight the way in which the argument works. All lemmas combined, together with the main theorem, can be proved on a single page, which underlines the simplicity of our approach.

Our first lemma upper bounds the expected squared distance of x^{k+1} from x^* in terms of the same distance but for x^k, function suboptimality, and variance of g^k.

Lemma 4.1. We have

$$\begin{aligned} \mathrm{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] &\le (1-\eta\mu)\left\|x^k - x^*\right\|^2 \\ &\quad - 2\eta(f(x^k) - f(x^*)) + \eta^2 \mathrm{E}\left[\left\|g^k\right\|^2\right]. \end{aligned} \tag{12}$$

In our next lemma, we further bound the variance of g^k in terms of function suboptimality and $\mathcal{D}^k$.

Lemma 4.2. We have

$$\mathrm{E}\left[\left\|g^k\right\|^2\right] \le 4L(f(x^k) - f(x^*)) + \frac{p}{2\eta^2}\mathcal{D}^k. \tag{13}$$

Finally, we bound $\mathrm{E}\left[\mathcal{D}^{k+1}\right]$ in terms of $\mathcal{D}^k$ and function suboptimality.

Lemma 4.3. We have

$$\mathrm{E}\left[\mathcal{D}^{k+1}\right] \le (1-p)\mathcal{D}^k + 8L\eta^2(f(x^k) - f(x^*)). \tag{14}$$

Putting the above three lemmas together naturally leads to the following result involving Lyapunov function (5).

Lemma 4.4. Let the step size $\eta \le \frac{1}{6L}$. Then for all $k \ge 0$ the following inequality holds:

$$\mathrm{E}\left[\Phi^{k+1}\right] \le (1-\eta\mu)\left\|x^k - x^*\right\|^2 + \left(1 - \frac{p}{2}\right)\mathcal{D}^k. \tag{15}$$

In order to obtain a recursion involving the Lyapunov function on the right-hand side of (15)

Theorem 4.5. Let $\eta = \frac{1}{6L}$, $p = \frac{1}{n}$. Then $\mathrm{E}\left[\Phi^k\right] \le \varepsilon\Phi^0$ as long as

$$k \ge \mathcal{O}\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\varepsilon}\right).$$

Proof. As the corollary of Lemma 4.4 we have

$$\mathrm{E}\left[\Phi^k\right] \le \max\left\{1 - \eta\mu, 1 - \frac{p}{2}\right\}\Phi^{k-1}.$$

Setting $\eta = \frac{1}{6L}$, $p = \frac{1}{n}$ and unrolling conditional probability one obtains

$$\mathrm{E}\left[\Phi^k\right] \le \max\left\{1 - \frac{\mu}{6L}, 1 - \frac{1}{2n}\right\}^k \Phi^0,$$

which concludes the proof. □

Note that the step size does not depend on the strong convexity parameter μ and yet the resulting complexity adapts to it.

4.3. Discussion

Examining (15), we can see that contraction of the Lyapunov function is $\max\{1 - \eta\mu, 1 - p/2\}$. Due to the limitation of $\eta \le \frac{1}{6L}$, the first term is at least $1 - \frac{\eta}{6\mu}$, thus the complexity cannot better than $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$. In terms of total complexity (number of stochastic gradient calls), L-SVRG calls the stochastic gradient oracle in expectation $\mathcal{O}(1 + pn)$ times times in each iteration. Combining these two complexities together, one gets the total complexity

$$\mathcal{O}\left(\left(\frac{1}{p} + n + \frac{L}{\mu} + \frac{L}{\mu}pn\right)\log\frac{1}{\varepsilon}\right).$$

Note that any choice of

$$p \in \left[\min\left\{\frac{c}{n}, \frac{c\mu}{L}\right\}, \max\left\{\frac{c}{n}, \frac{c\mu}{L}\right\}\right],$$

where $c = \Theta(1)$, leads to the optimal total complexity $\mathcal{O}\left(\left(n + \frac{L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$. This fills the gap in SVRG theory, where the outer loop length (in our case $1/p$ in expectation) needs to be proportional to $\frac{L}{\mu}$. Moreover, analysis for L-SVRG is much simpler and provides more insights.

5. Loopless Katyusha (L-Katyusha)

In this section we describe in detail the Loopless Katyusha method (L-Katyusha), and its convergence properties.

5.1. The algorithm

The L-Katyusha method, formalized as Algorithm 2, is inspired by the original Katyusha (Allen-Zhu, 2017) method. We use the same technique as for Algorithm 1, where we remove the outer loop present in Katyusha and instead use a probabilistic update of the full gradient.

Algorithm 2 Loopless Katyusha (L-Katyusha)

Parameters: θ_1, θ_2, probability $p \in (0, 1]$
Initialization: Choose $y^0 = w^0 = z^0 \in \mathbb{R}^d$, stepsize $\eta = \frac{\theta_2}{(1+\theta_2)\theta_1}$ and set $\sigma = \frac{\mu}{L}$
for $k = 0, 1, 2, \ldots$ **do**
$x^k = \theta_1 z^k + \theta_2 w^k + (1 - \theta_1 - \theta_2) y^k$
Sample $i \in \{1, \ldots, n\}$ uniformly at random
$g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$
$z^{k+1} = \frac{1}{1+\eta\sigma}\left(\eta\sigma x^k + z^k - \frac{\eta}{L} g^k\right)$
$y^{k+1} = x^k + \theta_1(z^{k+1} - z^k)$
$w^{k+1} = \begin{cases} y^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}$
end for

The exact analogy applies to the reference point w^k (at which a full gradient is computed) as for L-SVRG. Instead of updating this point in a deterministic way every m iteration, we use the probabilistic update with parameter p, when we update w^{k+1} to the current iterate y^k with this probability and is left unchanged with probability $1 - p$. As we shall see next, the same choice $p = \frac{1}{n}$ as for L-SVRG leads to complexity identical to that of original Katyusha.

5.2. Convergence theory

In comparison to L-SVRG, we don't use *gradient mapping* as the key component of our analysis. Instead, we prove convergence of functional values in y^k, w^k and point-wise convergence of z^k. This is summarized in the following Lyapunov function:

$$\Psi^k = \mathcal{Z}^k + \mathcal{Y}^k + \mathcal{W}^k, \tag{16}$$

where

$$\mathcal{Z}^k = \frac{L(1+\eta\sigma)}{2\eta}\left\|z^k - x^*\right\|^2, \ \mathcal{Y}^k = \frac{1}{\theta_1}(f(y^k) - f(x^*)), \tag{17}$$

and

$$\mathcal{W}^k = \frac{\theta_2(1+\theta_1)}{p\theta_1}(f(w^k) - f(x^*)). \tag{18}$$

Note that even if x^k is not in this function, its point-wise convergence is directly implied by the convergence of Ψ^k due to the definition of x^k in Algorithm 2 and L-smoothness of f.

The analysis involves five lemmas, followed by the iteration complexity summarized in the main theorem. The lemmas highlight important steps of our analysis. The simplicity of our approach is still preserved, where all lemmas combined, together with the main theorem, can be proved on not more than two pages.

Our first lemma upper bounds the variance of the gradient estimator g^k, which eventually goes to zero as our algorithm progresses.

Lemma 5.1. We have

$$\begin{aligned} &\mathrm{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\ &\leq 2L\left(f(w^k) - f(x^k) - \left\langle \nabla f(x^k), w^k - x^k\right\rangle\right). \end{aligned} \tag{19}$$

Next two lemmas are more technical, but essential for proving the convergence.

Lemma 5.2. We have

$$\begin{aligned} &\left\langle g^k, x^* - z^{k+1}\right\rangle + \frac{\mu}{2}\left\|x^k - x^*\right\|^2 \\ &\geq \frac{L}{2\eta}\left\|z^k - z^{k+1}\right\|^2 + \mathcal{Z}^{k+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^k. \end{aligned} \tag{20}$$

Lemma 5.3. We have

$$\begin{aligned} &\frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) - \frac{\theta_2}{2L\theta_1}\left\|g^k - \nabla f(x^k)\right\|^2 \\ &\leq \frac{L}{2\eta}\left\|z^{k+1} - z^k\right\|^2 + \left\langle g^k, z^{k+1} - z^k\right\rangle. \end{aligned} \tag{21}$$

Finally, we use the update of Algorithm 2 to decompose $\mathcal{W}^{k+1}$ in terms of $\mathcal{W}^k$ and $\mathcal{Y}^k$, which is one of the main components that allow for simpler analysis than the one of original Katyusha.

Lemma 5.4. We have

$$\mathrm{E}\left[\mathcal{W}^{k+1}\right] = (1-p)\mathcal{W}^k + \theta_2(1+\theta_1)\mathcal{Y}^k. \tag{22}$$

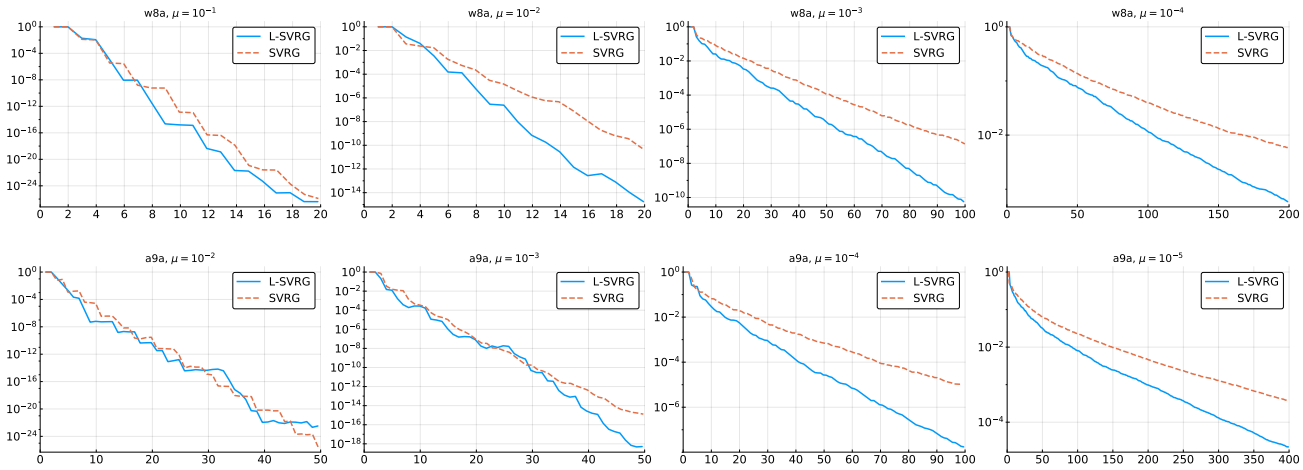

Figure 1. Comparison of SVRG and L-SVRG for different datasets and regularizer weights μ.

Putting all lemmas together, we obtain the following contraction of the Lyapunov function defined in (7).

Lemma 5.5. Let $\theta_1, \theta_2 > 0$, $\theta_1 + \theta_2 \leq 1$, $\sigma = \frac{\mu}{L}$ and $\eta = \frac{\theta_2}{(1+\theta_2)\theta_1}$, then we have

$$\begin{aligned} &\mathrm{E}\left[\mathcal{Z}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1}\right] \\ &\leq \frac{1}{1+\eta\sigma}\mathcal{Z}^k + (1-\theta_1(1-\theta_2))\mathcal{Y}^k \\ &\quad + \left(1 - \frac{p\theta_1}{1+\theta_1}\right)\mathcal{W}^k. \end{aligned} \tag{23}$$

In order to obtain a recursion involving the Lyapunov function on the right-hand side of (23)

Theorem 5.6. Let $\theta_1 = \min\left\{\sqrt{\frac{2\sigma n}{3}}, \frac{1}{2}\right\}$, $\theta_2 = \frac{1}{2}$, $p = \frac{1}{n}$. Then $\mathrm{E}\left[\Psi^k\right] \leq \varepsilon\Psi^0$ after the following number of iterations:

$$k = O\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\varepsilon}\right).$$

Proof. As a corollary of Lemma 5.5, we have

$$\mathrm{E}\left[\Psi^{k+1}\right] \leq \max\left\{\tfrac{1}{1+\eta\sigma}, 1-\theta_1(1-\theta_2), 1-\tfrac{p\theta_1}{1+\theta_1}\right\}\Psi^k.$$

Setting $p = 1/n$, $\theta_1 = \min\left\{\sqrt{\frac{2\sigma n}{3}}, \frac{1}{2}\right\}$, $\theta_2 = \frac{1}{2}$, and unrolling conditional probability one obtains $\mathrm{E}\left[\Psi^{k+1}\right] \leq (1-\theta)\mathrm{E}\left[\Psi^k\right]$, where $\theta = \min\left\{\frac{\sigma}{6\theta_1}, \frac{\theta_1}{2n}\right\}$. Choosing $\sigma = \frac{\mu}{L}$ concludes the proof. □

5.3. Discussion

One can show by analyzing (23) that for ill-conditioned problems ($n < L/\mu$), the iteration complexity is $\mathcal{O}\left(\sqrt{\frac{L}{\mu p}}\log\frac{1}{\varepsilon}\right)$. Algorithm 2 calls stochastic gradient oracle $\mathcal{O}(1+pn)$ times per iteration in expectation. Thus, the total complexity is $\mathcal{O}\left((1+pn)\sqrt{\frac{L}{\mu p}}\log\frac{1}{\varepsilon}\right)$. One can see that $p = \Theta\left(\frac{1}{n}\right)$ leads to optimal rate.

6. Numerical Experiments

In this section, we perform experiments with logistic regression for binary classification with L_2 regularizer, where our loss function has the form

$$f_i(x) = \log\left(1 + \exp(-b_i a_i^\top x)\right) + \tfrac{\mu}{2}\|x\|^2,$$

where $a_i \in \mathbb{R}^d, b_i \in \{-1, +1\}, i = 1, \ldots, n$. Hence, f is smooth and $\mu-$strongly convex. We use four LIBSVM library[2]: *a9a, w8a, mushrooms, phishing, cod-rna*.

We compare our methods L-SVRG and L-Katyusha with their original version. It is well-known that whenever practical, SAGA is a bit faster than SVRG. While a comparison to SAGA seems natural as it also does not have a double loop structure, we position our loopless methods for applications where the high memory requirements of SAGA prevent it to be applied. Thus, we do not compare to SAGA.

For Section 6.1, the parameters are chosen as suggested by theory. For L-SVRG and L-Katyusha they are chosen based on Theorems 4.5 and 5.6, respectively. For SVRG and Katyusha we also choose the parameters based on the

[2]The LIBSVM dataset collection is available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

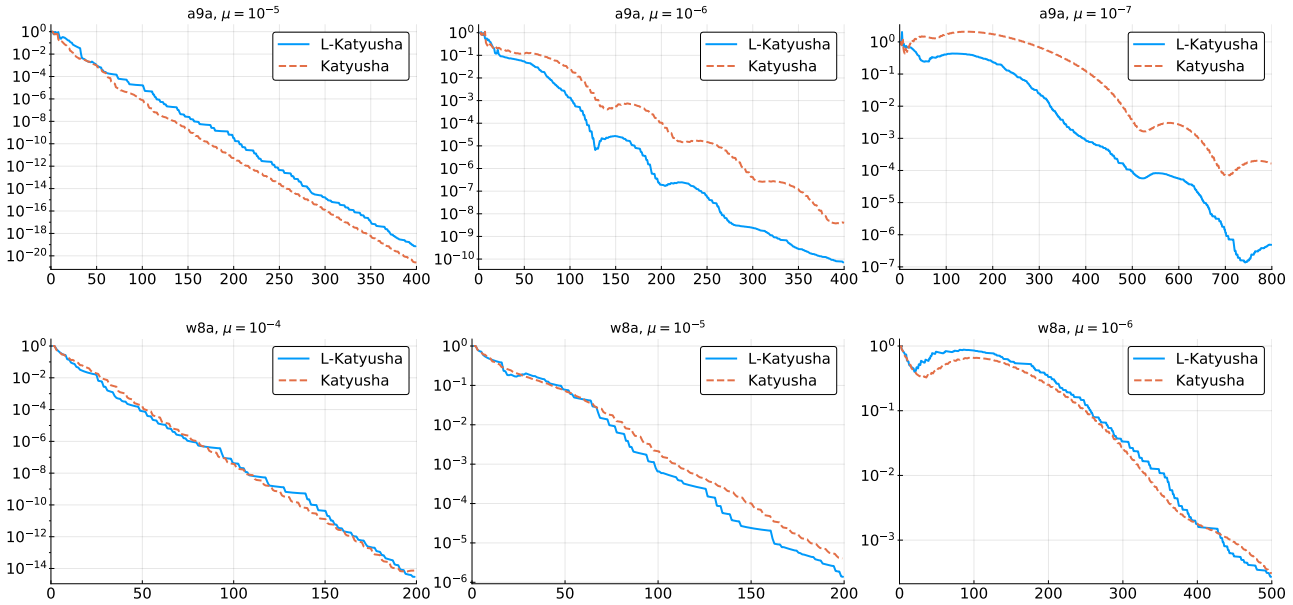

Figure 2. Comparison of `Katyusha` and `L-Katyusha` for different datasets and regularizer weights μ.

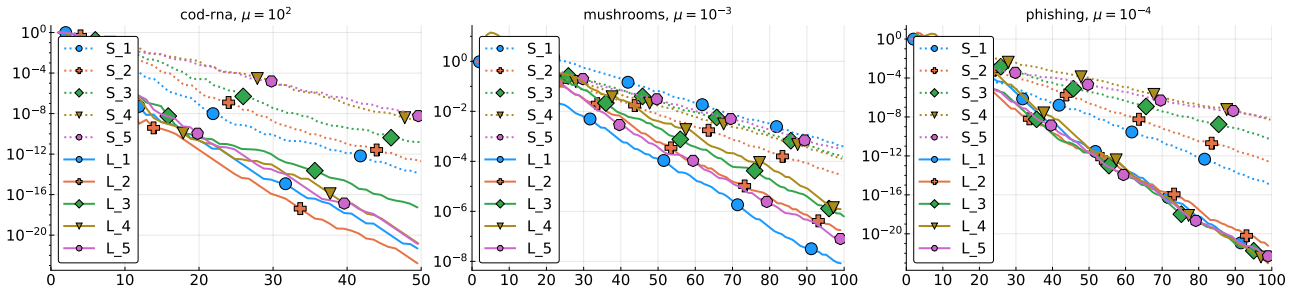

Figure 3. Comparison of `SVRG` (S) and `L-SVRG` (L) for several choices of expected outer loop length (`L-SVRG`) or deterministic outer loop length (`SVRG`). Numbers 1–5 correspondent to loop-lengths $n, \sqrt[4]{\kappa n^3}, \sqrt{\kappa n}, \sqrt[4]{\kappa^3 n}, \kappa$, respectively, where $\kappa = L/\mu$.

theory, as described in the original papers. The initial point x^0 is chosen to be the origin. In Section 6.2, we provide several choices of parameters. Plots are constructed in such a way that the y-axis displays $\left\|x^k - x^\star\right\|^2$ for `L-SVRG` and $\left\|y^k - x^\star\right\|^2$ for `L-Katyusha`, where $x^\star$ were obtained by running gradient descent for a large number of epochs. The x-axis displays the number of epochs (full gradient evaluations). That is, n computations of $\nabla f_i(x)$ equals one epoch.

6.1. Superior practical behaviour of the loopless approach

In this section, we show that the replacement of the outer loop in `SVRG` and `Katyusha` brings not just a simpler analysis but also speed up in the experiments.

In theory, both the loopy and the loopless methods are the same. However, as we can see from Figure 1, the improvement of the loopless approach can be significant. One can see that for these datasets, `L-SVRG` *is always better than* `SVRG`*, and can be faster by several orders of magnitude!*

Looking at Figure 2, we see that *the performance of* `L-Katyusha` *is at least as good that of* `Katyusha`*, and can be significantly faster in some cases.* These experiments support the claim that not just the theoretical analysis is simpler, but also the performance is affected in a positive way.

6.2. Different choices of probability/ outer loop size

In this section, we compare several choices of the probability p of updating the full gradient for `SVRG` and several outer loop sizes m for `SVRG`. Since our analysis guarantees the

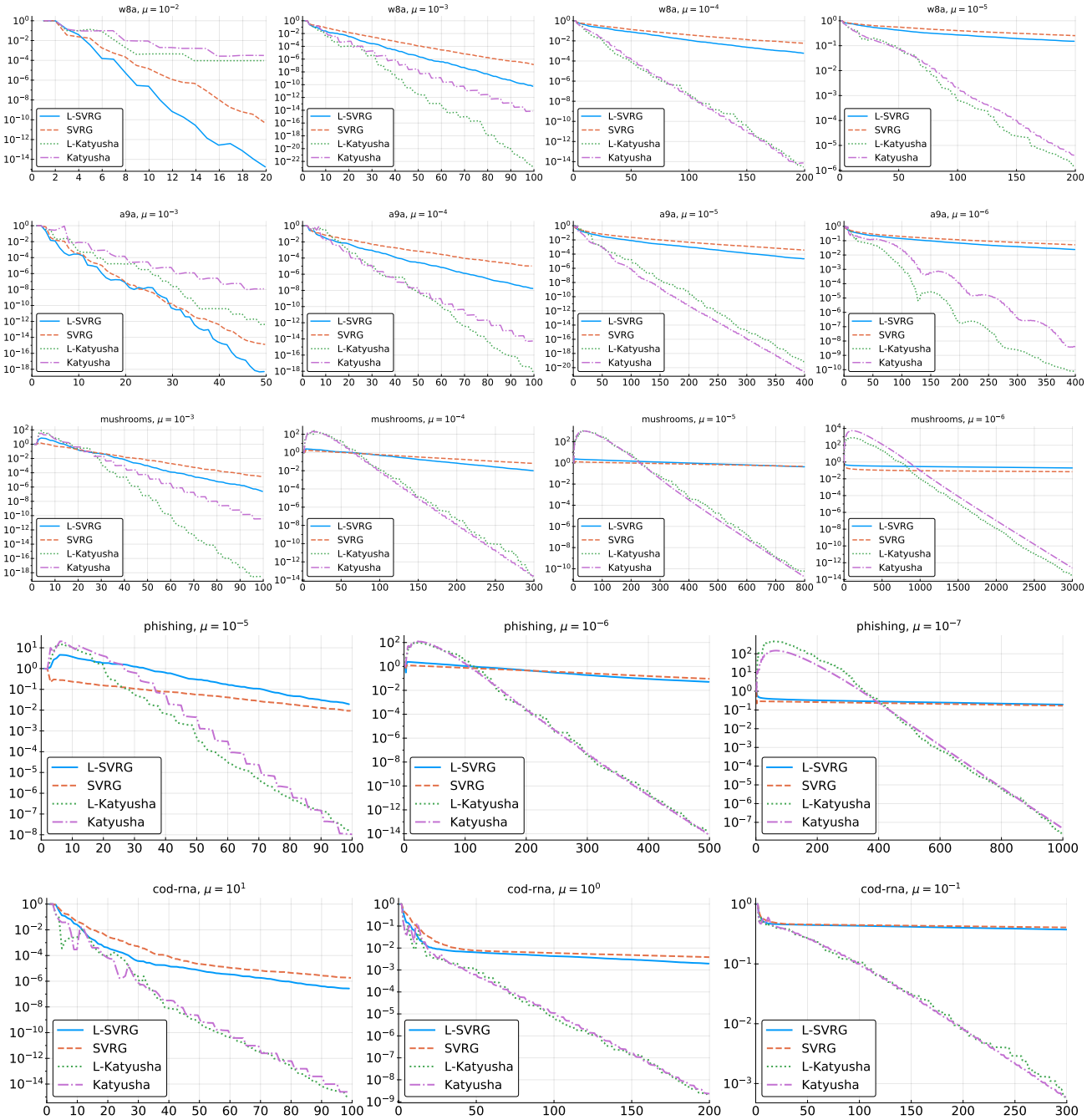

Figure 4. All methods together for different datasets and different regularizer weights.

optimal rate for any choice of p between $1/n$ and μ/L for well condition problems, we decided to perform experiments for p within this range. More precisely, we choose 5 values of p, uniformly distributed in logarithmic scale across this interval, and thus our choices are $n, \sqrt[4]{\kappa n^3}$, $\sqrt{\kappa n}$, $\sqrt[4]{\kappa^3 n}$, and κ, where $\kappa = L/\mu$, denoted in the figures by $1, 2, 3, 4, 5$, respectively. Since the expected "outer loop" length (length for which reference point stays the same) is $1/p$, for SVRG we choose $m = 1/p$.

Looking at Figure 3, one can see that L−SVRG is very *robust* to the choice of p from the "optimal interval" predicted by our theory. Moreover, *even the worst case for* L−SVRG *outperforms the best case for* SVRG.

6.3. All methods together

As the last visualization we provide all algorithms together in one graph for different datasets with different regularizer weight, thus with different condition number, displayed in Figure 4. As for the previous experiments, loopless methods are not worse and sometimes significantly better.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017.

Bibi, A., Sailanbayev, A., Ghanem, B., Gower, R. M., and Richtárik, P. Improving SAGA via a probabilistic interpolation with gradient descent. *arXiv: 1806.05633*, 2018.

Csiba, D. and Richtárik, P. Primal method for ERM with flexible mini-batching schemes and non-convex losses. *arXiv:1506.02227*, 2015.

Csiba, D., Qu, Z., and Richtárik, P. Stochastic dual coordinate ascent with adaptive probabilities. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 674–683, 2015.

Defazio, A. A simple practical accelerated method for finite sums. In *Advances in Neural Information Processing Systems*, pp. 676–684, 2016.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014a.

Defazio, A., Caetano, T., and Domke, J. Finito: A faster, permutable incremental gradient method for Big Data problems. *The 31st International Conference on Machine Learning*, 2014b.

Doikov, N. and Richtárik, P. Randomized block cubic Newton method. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Gower, R. M., Goldfarb, D., and Richtárik, P. Stochastic block BFGS: squeezing more curvature out of data. In *33rd International Conference on Machine Learning*, pp. 1869–1878, 2016.

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. *arXiv:1805.02632*, 2018.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pp. 2305–2313, 2015.

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

Konečný, J. and Richtárik, P. S2GD: Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, pp. 1–14, 2017.

Konečný, J., Lu, J., Richtárik, P., and Takáč, M. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10 (2):242–255, 2016.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pp. 2348–2358, 2017.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

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

Nemirovsky, A. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621, 2017.

Qu, Z., Richtárik, P., and Zhang, T. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems 28*, pp. 865–873, 2015.

Qu, Z., Richtárik, P., Takáč, M., and Fercoq, O. SDNA: Stochastic dual Newton ascent for empirical risk minimization. In *The 33rd International Conference on Machine Learning*, pp. 1823–1832, 2016.

Roux, N. L., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for

finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shalev-Shwartz, S. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pp. 747–754, 2016.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, pp. 64–72, 2014.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. Mini-batch primal and dual methods for SVMs. In *30th International Conference on Machine Learning*, pp. 537–552, 2013.

Zhou, K. Direct acceleration of SAGA using sampled negative momentum. *arXiv preprint arXiv:1806.11048*, 2018.

Zhou, K., Shang, F., and Cheng, J. A simple stochastic variance reduced algorithm with fast convergence rates. *arXiv preprint arXiv:1806.11027*, 2018.

SUPPLEMENTARY MATERIAL
SVRG and Katyusha are Better Without the Outer Loop

A. Auxiliary Lemmas

Lemma A.1. For random vector $X \in \mathbb{R}^d$ and any $y \in \mathbb{R}^d$, the variance can be decomposed as

$$\mathrm{E}\left[\|X - \mathrm{E}[X]\|^2\right] = \mathrm{E}\left[\|X - y\|^2\right] - \mathrm{E}\left[\|\mathrm{E}[X] - y\|^2\right] . \tag{24}$$

Lemma A.2. For any vectors $a_1, a_2, \ldots, a_k \in \mathbb{R}^d$, we have as a consequence of Jensen's inequality applied to $x \mapsto \|x\|^2$:

$$\left\|\sum_{i=1}^k a_i\right\|^2 \leq k \sum_{i=1}^k \|a_i\|^2 . \tag{25}$$

B. Proofs for Algorithm 1 (`L-SVRG`)

Let $f^* = f(x^*)$.

Proof of Lemma 4.1. Definition of x^{k+1} and unbiasness of g^k guarantee that

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] &= \mathrm{E}\left[\left\|x^k - x^* - \eta g^k\right\|\right]^2 \\
&\overset{\text{Alg. 1}}{=} \left\|x^k - x^*\right\|^2 + \mathrm{E}\left[2\eta\left\langle g^k, x^* - x^k\right\rangle\right] + \eta^2 \mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\overset{(2)}{=} \left\|x^k - x^*\right\|^2 + 2\eta\left\langle \nabla f(x^k), x^* - x^k\right\rangle + \eta^2 \mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\overset{(4)}{\leq} \left\|x^k - x^*\right\|^2 + 2\eta\left(f^* - f(x^k) - \frac{\mu}{2}\left\|x^k - x^*\right\|\right) + \eta^2 \mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&= \left\|x^k - x^*\right\|^2 (1 - \eta\mu) + 2\eta\left(f^* - f(x^k)\right) + \eta^2 \mathrm{E}\left[\left\|g^k\right\|^2\right].
\end{aligned}$$

□

Proof of Lemma 4.2. Using definition of g^k

$$\begin{aligned}
\mathrm{E}\left[\left\|g^k\right\|^2\right] &\overset{\text{Alg. 1}}{=} \mathrm{E}\left[\left\|\nabla f_i(x^k) - \nabla f_i(x^*) + \nabla f_i(x^*) - \nabla f_i(w^k) + \nabla f(w^k)\right\|^2\right] \\
&\overset{(25)}{\leq} 2\mathrm{E}\left[\left\|\nabla f_i(x^k) - \nabla f_i(x^*)\right\|^2\right] + 2\mathrm{E}\left[\left\|\nabla f_i(x^*) - \nabla f_i(w^k) - \mathrm{E}\left[\nabla f_i(x^*) - \nabla f_i(w^k)\right]\right\|^2\right] \\
&\overset{(3),(24)}{\leq} 4L(f(x^k) - f^*) + 2\mathrm{E}\left[\left\|\nabla f_i(w^k) - \nabla f_i(x^*)\right\|^2\right] \\
&\overset{(10)}{=} 4L(f(x^k) - f^*) + \frac{p}{2\eta^2}\mathcal{D}^k.
\end{aligned}$$

□

Proof of Lemma 4.3.

$$\begin{aligned}
\mathrm{E}\left[\mathcal{D}^{k+1}\right] &\overset{\text{Alg. 1}}{=} (1-p)\mathcal{D}^k + p\frac{4\eta^2}{pn}\sum_{i=1}^n \left\|\nabla f(x^k) - \nabla f(x^*)\right\|^2 \\
&\overset{(3)}{\leq} (1-p)\mathcal{D}^k + 8L\eta^2(f(x^k) - f^*).
\end{aligned}$$

□

Proof of Lemma 4.4. Combining Lemmas 4.1 and 4.3 we obtain

$$\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1}-x^*\right\|^2+\mathcal{D}^{k+1}\right] &\overset{(12),(14)}{\leq} (1-\mu\eta)\left\|x^k-x^*\right\|^2+2\eta(f^*-f(x^k))+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\quad +(1-p)\mathcal{D}^k+8L\eta^2(f(x^k)-f^*) \\
&\overset{(13)}{\leq} (1-\mu\eta)\left\|x^k-x^*\right\|^2+(1-p)\mathcal{D}^k+(2\eta-8L\eta^2)(f^*-f(x^k)) \\
&\quad +\eta^2\left(4L(f(x^k)-f^*)+\frac{p}{2\eta^2}\mathcal{D}^k\right) \\
&= (1-\mu\eta)\left\|x^k-x^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}^k+(2\eta-12L\eta^2)(f^*-f(x^k)).
\end{aligned}$$

Now we use the fact that $\eta \leq \frac{1}{6L}$ and obtain the desired inequality:

$$\mathrm{E}\left[\left\|x^{k+1}-x^*\right\|^2+\mathcal{D}^{k+1}\right] \leq (1-\mu\eta)\left\|x^k-x^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}^k.$$

□

C. Proofs for Algorithm 2 (`L-Katyusha`)

Proof of Lemma 5.1. To upper bound the variance of g^k we first uses its definition

$$\begin{aligned}
\mathrm{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] &\overset{\text{Alg. 2}}{=} \mathrm{E}\left[\left\|\nabla f_i(x^k)-\nabla f_i(w^k)-\mathrm{E}\left[\nabla f_i(x^k)-\nabla f_i(w^k)\right]\right\|^2\right] \\
&\overset{(24)}{\leq} \mathrm{E}\left[\left\|\nabla f_i(x^k)-\nabla f_i(w^k)\right\|^2\right] \\
&\overset{(3)}{\leq} 2L\left(f(w^k)-f(x^k)-\left\langle\nabla f(x^k),w^k-x^k\right\rangle\right).
\end{aligned}$$

□

Proof of Lemma 5.2. We start with the definition of z^{k+1}

$$z^{k+1} \overset{\text{Alg. 2}}{=} \frac{1}{1+\eta\sigma}\left(\eta\sigma x^k+z^k-\frac{\eta}{L}g^k\right),$$

which implies

$$\frac{\eta}{L}g^k=\eta\sigma(x^k-z^{k+1})+(z^k-z^{k+1}),$$

which further implies that

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

□

Proof of Lemma 5.3.

$$
\begin{aligned}
& \frac{L}{2\eta}\left\|z^{k+1}-z^k\right\|^2+\left\langle g^k, z^{k+1}-z^k\right\rangle \\
= \quad & \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|\theta_1(z^{k+1}-z^k)\right\|^2+\left\langle g^k, \theta_1(z^{k+1}-z^k)\right\rangle\right) \\
\overset{\text{Alg. 2}}{=} \quad & \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|y^{k+1}-x^k\right\|^2+\left\langle g^k, y^{k+1}-x^k\right\rangle\right) \\
= \quad & \frac{1}{\theta_1}\left(\frac{L}{2\eta\theta_1}\left\|y^{k+1}-x^k\right\|^2+\left\langle \nabla f(x^k), y^{k+1}-x^k\right\rangle+\left\langle g^k-\nabla f(x^k), y^{k+1}-x^k\right\rangle\right) \\
= \quad & \frac{1}{\theta_1}\left(\frac{L}{2}\left\|y^{k+1}-x^k\right\|^2+\left\langle \nabla f(x^k), y^{k+1}-x^k\right\rangle+\frac{L}{2}\left(\frac{1}{\eta\theta_1}-1\right)\left\|y^{k+1}-x^k\right\|^2+\left\langle g^k-\nabla f(x^k), y^{k+1}-x^k\right\rangle\right) \\
\overset{(3)}{\geq} \quad & \frac{1}{\theta_1}\left(f(y^{k+1})-f(x^k)+\frac{L}{2}\left(\frac{1}{\eta\theta_1}-1\right)\left\|y^{k+1}-x^k\right\|^2+\left\langle g^k-\nabla f(x^k), y^{k+1}-x^k\right\rangle\right) \\
\geq \quad & \frac{1}{\theta_1}\left(f(y^{k+1})-f(x^k)-\frac{\eta\theta_1}{2L(1-\eta\theta_1)}\left\|g^k-\nabla f(x^k)\right\|^2\right) \\
= \quad & \frac{1}{\theta_1}\left(f(y^{k+1})-f(x^k)-\frac{\theta_2}{2L}\left\|g^k-\nabla f(x^k)\right\|^2\right),
\end{aligned}
$$

where the last inequality uses the Young's inequality in the form of $\langle a, b\rangle \geq -\frac{\|a\|^2}{2\beta}-\frac{\beta\|b\|^2}{2}$ for $\beta=\frac{\eta\theta_1}{L(1-\eta\theta_1)}$, which concludes the proof. □

Proof of Lemma 5.4. From the definition of w^{k+1} in Algorithm 2 we have

$$
\mathrm{E}\left[f(w^{k+1})\right] \overset{\text{Alg. 2}}{=} (1-p)f(w^k)+pf(y^k). \tag{26}
$$

The rest of proof follows from the definition of $\mathcal{W}^k$ (22). □

Proof of Lemma 5.5. Combining all the previous lemmas together we obtain

$$
\begin{aligned}
f^* &\overset{(4)}{\geq} f(x^k) + \left\langle \nabla f(x^k), x^* - x^k \right\rangle + \frac{\mu}{2}\left\|x^k - x^*\right\|^2 \\
&= f(x^k) + \frac{\mu}{2}\left\|x^k - x^*\right\|^2 + \left\langle \nabla f(x^k), x^* - z^k + z^k - x^k \right\rangle \\
&\overset{\text{Alg. 2}}{=} f(x^k) + \frac{\mu}{2}\left\|x^k - x^*\right\|^2 + \left\langle \nabla f(x^k), x^* - z^k \right\rangle + \frac{\theta_2}{\theta_1}\left\langle \nabla f(x^k), x^k - w^k \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}\left\langle \nabla f(x^k), x^k - y^k \right\rangle \\
&\overset{(2)}{\geq} f(x^k) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(x^k), x^k - w^k \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathrm{E}\left[\frac{\mu}{2}\left\|x^k - x^*\right\|^2 + \left\langle g^k, x^* - z^{k+1} \right\rangle + \left\langle g^k, z^{k+1} - z^k \right\rangle\right] \\
&\overset{(20)}{\geq} f(x^k) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(x^k), x^k - w^k \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathrm{E}\left[\mathcal{Z}^{k+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^k\right] + \mathrm{E}\left[\left\langle g^k, z^{k+1} - z^k \right\rangle + \frac{L}{2\eta}\left\|z^k - z^{k+1}\right\|^2\right] \\
&\overset{(21)}{\geq} f(x^k) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(x^k), x^k - w^k \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathrm{E}\left[\mathcal{Z}^{k+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^k\right] + \mathrm{E}\left[\frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) - \frac{\theta_2}{2L\theta_1}\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&\overset{(19)}{\geq} f(x^k) + \frac{\theta_2}{\theta_1}\left\langle \nabla f(x^k), x^k - w^k \right\rangle + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(x^k) - f(y^k)) \\
&\quad + \mathrm{E}\left[\mathcal{Z}^{k+1} - \frac{1}{1+\eta\sigma}\mathcal{Z}^k\right] + \mathrm{E}\left[\frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right) - \frac{\theta_2}{\theta_1}\left(f(w^k) - f(x^k) - \left\langle \nabla f(x^k), w^k - x^k \right\rangle\right)\right] \\
&= f(x^k) + \frac{(1-\theta_1-\theta_2)}{\theta_1}(f(x^k) - f(y^k)) - \frac{1}{1+\eta\sigma}\mathcal{Z}^k - \frac{\theta_2}{\theta_1}(f(w^k) - f(x^k)) \\
&\quad + \mathrm{E}\left[\mathcal{Z}^{k+1} + \frac{1}{\theta_1}\left(f(y^{k+1}) - f(x^k)\right)\right],
\end{aligned}
$$

where in the second inequality we use also convexity of $f(x)$.

$$
\begin{aligned}
x^k &\overset{\text{Alg. 2}}{=} \theta_1 z^k + \theta_2 w^k + (1-\theta_1-\theta_2) y^k \\
z^k - x^k &= \frac{\theta_2}{\theta_1}(x^k - w^k) + \frac{1-\theta_1-\theta_2}{\theta_1}(x^k - y^k).
\end{aligned}
$$

After rearranging we get

$$
\frac{1}{1+\eta\sigma}\mathcal{Z}^k + (1-\theta_1-\theta_2)\mathcal{Y}^k + \frac{\theta_2}{\theta_1}(f(w^k - f^*)) \geq \mathrm{E}\left[\mathcal{Z}^{k+1} + \mathcal{Y}^{k+1}\right].
$$

Using definition of $\mathcal{W}^k$ we get

$$
\mathrm{E}\left[\mathcal{Z}^{k+1} + \mathcal{Y}^{k+1}\right] \leq \frac{1}{1+\eta\sigma}\mathcal{Z}^k + (1-\theta_1-\theta_2)\mathcal{Y}^k + \frac{p}{(1+\theta_1)}\mathcal{W}^k. \tag{27}
$$

Finally, using Lemma 5.4 we get

$$
\begin{aligned}
\mathrm{E}\left[\mathcal{Z}^{k+1} + \mathcal{Y}^{k+1} + \mathcal{W}^{k+1}\right] &\leq \frac{1}{1+\eta\sigma}\mathcal{Z}^k + (1-\theta_1-\theta_2)\mathcal{Y}^k + \frac{p}{(1+\theta_1)}\mathcal{W}^k + (1-p)\mathcal{W}^k + \theta_2(1+\theta_1)\mathcal{Y}^k \\
&= \frac{1}{1+\eta\sigma}\mathcal{Z}^k + (1-\theta_1(1-\theta_2))\mathcal{Y}^k + \left(1 - \frac{p\theta_1}{1+\theta_1}\right)\mathcal{W}^k,
\end{aligned}
$$

which concludes the proof. □