Dissipativity Theory for Accelerating Stochastic Variance Reduction: A Unified Analysis of SVRG and Katyusha Using Semidefinite Programs

Bin Hu [1] **Stephen Wright** [1] **Laurent Lessard** [1]

Abstract

Techniques for reducing the variance of gradient estimates used in stochastic programming algorithms for convex finite-sum problems have received a great deal of attention in recent years. By leveraging dissipativity theory from control, we provide a new perspective on two important variance-reduction algorithms: SVRG and its direct accelerated variant Katyusha. Our perspective provides a physically intuitive understanding of the behavior of SVRG-like methods via a principle of energy conservation. The tools discussed here allow us to automate the convergence analysis of SVRG-like methods by capturing their essential properties in small semidefinite programs amenable to standard analysis and computational techniques. Our approach recovers existing convergence results for SVRG and Katyusha and generalizes the theory to alternative parameter choices. We also discuss how our approach complements the linear coupling technique. Our combination of perspectives leads to a better understanding of accelerated variance-reduced stochastic methods for finite-sum problems.

1. Introduction

Empirical risk minimization (ERM) is a key paradigm in machine learning (Bubeck, 2015; Bottou et al., 2016). Many learning problems, including ridge regression, logistic regression, and support vector machines, can be naturally formulated as the following finite-sum ERM

$$\min_{x\in\mathbb{R}^p} \; g(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where g is strongly convex. A standard approach for solving (1) is the stochastic gradient (SG) method (Robbins & Monro, 1951; Bottou & LeCun, 2003). Recently, a large family of variance-reduction methods have been developed to improve the convergence guarantees of SG. Such methods are typically classified into the following two categories:

1. SVRG-like methods are epoch-based, requiring evaluation of a complete gradient $\nabla g(\tilde{x})$ at the beginning of each epoch. The epoch length is typically set to be $2n$ but can be adaptive. Methods of this type include SVRG (Johnson & Zhang, 2013) and its direct accelerated variant Katyusha (Allen-Zhu, 2016).

2. SAGA-like methods do not involve epoch length tuning, and include SAG (Roux et al., 2012; Schmidt et al., 2013), SAGA (Defazio et al., 2014a), Finito (Defazio et al., 2014b), SDCA (Shalev-Shwartz & Zhang, 2013; Shalev-Shwartz, 2016), APCG (Lin et al., 2014), SPDC (Zhang & Xiao, 2017), and point-SAGA (Defazio, 2016). Due to storage issues, it may be difficult to apply SAGA-like methods to general learning problems other than linear prediction/classification, since their required storage for general learning tasks scales with the training set size.

This paper is motivated by the following two concerns. First, there has been recent interest in developing a unified, coherent set of tools for analyzing stochastic finite-sum methods. Traditionally, convergence proofs for variance-reduction methods have been developed in a case-by-case manner. More coherent techniques may facilitate the design of new finite-sum methods in more complicated setups. Recently, control theory has been used to derive linear matrix inequality (LMI) conditions that can be used to automate the analysis of a large family of first-order optimization methods (Lessard et al., 2016; Hu et al., 2017a;b; Hu & Lessard, 2017; Fazlyab et al., 2017). Specifically, Hu et al. (2017b) has tailored jump system theory to provide a unified analysis for SAGA, Finito, and SDCA. The analysis in Hu et al. (2017b) can potentially be extended to cover other SAGA-like methods, such as SAG, APCG, SPDC, and point-SAGA. However, as pointed out in Hu et al. (2017b), jump system theory may not be the most suitable tool for epoch-based methods. This paper aims in part to bridge the gap between control-oriented analysis and SVRG-like

[1]University of Wisconsin–Madison, United States. Correspondence to: Bin Hu <bhu38@wisc.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

methods by extending the deterministic dissipativity theory in Hu & Lessard (2017) to a stochastic setup. The approach of this paper allows us to formulate semidefinite programs for a unified analysis of SVRG-like methods. Together, dissipativity theory and the jump system theory described in Hu et al. (2017b) provide a complete picture of how control theory can be used to unify the analysis of stochastic finite-sum methods.

Second, there is still a need for better understanding of the role of momentum in the algorithms for the finite-sum problem (1). Nesterov's accelerated method (Nesterov, 2003) has received a great deal of attention for its ingenuity and its appealing theoretical and practical behavior. However, the original convergence rate proof of Nesterov's accelerated method relies on a technique of estimate sequences, and is not easy to interpret. Recently, new interpretations of Nesterov's accelerated method have been proposed from many different perspectives, for example, linear coupling (Allen-Zhu & Orecchia, 2014), geometric descent (Bubeck et al., 2015), control theory (Lessard et al., 2016; Hu & Lessard, 2017), continuous-time ODEs (Su et al., 2016; Wibisono et al., 2016; Wilson et al., 2016), and quadratic averaging (Drusvyatskiy et al., 2016). Among these new developments, linear coupling is the only one that has been extended to accelerate variance-reduction methods for the finite-sum problem (1); Katyusha momentum (Allen-Zhu, 2016) is based on this idea. Our current paper extends the control-oriented approach in Hu & Lessard (2017) to cover accelerated variance-reduction methods. The linear coupling framework of (Allen-Zhu & Orecchia, 2014; Allen-Zhu, 2016) provides useful and intuitive design guidelines for accelerating optimization methods. Our control approach complements linear coupling by providing a physical interpretation of accelerated variance-reduction methods, as well as automated convergence analysis via formulation and solution of small semidefinite programs. Both linear coupling and our control approach provide useful perspectives, and each has certain advantages from the viewpoint of analysis.

Our contributions can be summarized as follows. We present a unified analysis of SVRG and Katyusha by using the physically intuitive notion of dissipativity. We prove convergence results for SVRG by solving a 3×3 semidefinite program, and show that the existing convergence result for Katyusha can be recovered and generalized by solving a 6×6 semidefinite program. Numerical solutions of our proposed LMIs can be used to narrow the choices for various algorithm parameters (such as learning rate, momentum, and epoch length) at early stages of proof construction. We also present an energy-conservation interpretation for variance reduction and acceleration. Compared with Hu & Lessard (2017), the novelty of the present paper is the development of several new stochastic supply rate conditions that depend on the stochastic variance reduction mechanism.

2. Preliminaries

2.1. Notation

Let $\mathbb{R}$ and $\mathbb{R}_+$ denote the real and nonnegative real numbers, respectively. We denote the $p \times p$ identity matrix as I_p. The Kronecker product of two matrices is denoted as $A \otimes B$. Note hat $(A \otimes B)^\mathsf{T} = A^\mathsf{T} \otimes B^\mathsf{T}$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ when the matrices have compatible dimensions. A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is σ-strongly convex if $f(x) \geq f(y) + \nabla f(y)^\mathsf{T}(x-y) + \frac{\sigma}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^p$ and is L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$ for all $x, y \in \mathbb{R}^p$. Note that f is convex if f is 0-strongly convex. We use $x_\star$ to denote a point satisfying $\nabla f(x_\star) = 0$. When f is L-smooth and σ-strongly convex for some $\sigma > 0$, $x_\star$ is unique.

2.2. Dissipativity Theory for Stochastic Linear Systems

For completeness, we first review dissipativity theory for linear time-invariant (LTI) systems with stochastic inputs. Our development parallels that of Hu & Lessard (2017, Section 2.2), which reviews dissipativity theory for LTI systems with deterministic inputs.

Consider an LTI system governed by the state-space model

$$\xi_{k+1} = A\xi_k + Bw_k, \tag{2}$$

where $\xi_k \in \mathbb{R}^{n_\xi}$ is the state, $w_k \in \mathbb{R}^{n_w}$ is the input, and (A, B) are constant matrices with compatible dimensions, i.e. $A \in \mathbb{R}^{n_\xi \times n_\xi}$ and $B \in \mathbb{R}^{n_\xi \times n_w}$. The input sequence $\{w_k\}$ is assumed to be a stochastic process. Intuitively, we can interpret w_k as a stochastic force driving the state of the LTI model (2). Dissipativity theory describes how the input forces w_j, $j = 0, 1, 2, \ldots$ drive the internal energy stored in the states ξ_k, $k = 0, 1, 2, \ldots$. The theory hinges on two functions: a *supply rate* $S : \mathbb{R}^{n_\xi} \times \mathbb{R}^{n_w} \to \mathbb{R}$ and a *storage function* $V : \mathbb{R}^{n_\xi} \to \mathbb{R}_+$. Since w_k is stochastic, we adopt the following notion of almost sure dissipativity.

Definition 1. *The system (2) is almost surely (a.s.) dissipative with respect to the supply rate $S : \mathbb{R}^{n_\xi} \times \mathbb{R}^{n_w} \to \mathbb{R}$ if there exists a storage function $V : \mathbb{R}^{n_\xi} \to \mathbb{R}_+$ such that*

$$V(\xi_{k+1}) - V(\xi_k) \leq S(\xi_k, w_k) \ \ a.s. \tag{3}$$

for all k. The inequality (3) is called an a.s. dissipation inequality.

We now discuss physical interpretations for the supply rate S, the storage function V, and the dissipation inequality (3). The storage function V quantifies the amount of internal energy stored in the system state ξ_k. The supply rate function S maps any state/input pair (ξ, w) to a scalar that characterizes the energy supplied from the input w to the state ξ. (Note that the supply rate can be negative, in which case the force w_k is extracting energy from the system.) The

a.s. dissipation inequality (3) states that there will always (technically "a.s.") be some energy dissipating from the system (3), and hence the internal energy increase, which is $V(\xi_{k+1}) - V(\xi_k)$, is bounded above by the energy supplied to the system. The dissipation inequality can be thought of as a restatement of the energy conservation law. A useful variant of (3) is the exponential dissipation inequality:

$$V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, w_k) \text{ a.s.}, \tag{4}$$

where $0 \leq \rho \leq 1$ is given. The exponential dissipation inequality (4) just states that at least a fraction $(1-\rho^2)$ of the internal energy will dissipate at every step k.

Remark 2. *It is often the case that the driving force w_k depends on the state ξ_k in some prescribed way, so we know some properties of the supply rate in advance. If the supply rate function S satisfies certain bounds, then the dissipation inequality can be used to obtain convergence guarantees for (2). For example, if we know there exists a positive constant M such that $\mathbb{E}S(\xi_k, w_k) \leq M$ for all k, then taking expectation of (4) leads to the conclusion that $\mathbb{E}V(\xi_{k+1}) \leq \rho^2 \mathbb{E}V(\xi_k) + M$. Based on this inequality, one can show that $\mathbb{E}V(\xi_k) \leq \rho^{2k} V(\xi_0) + \frac{M}{1-\rho^2}$. This suggests that the state ξ_k linearly converges to a ball centered at the origin, and the radius of the ball is related to $\frac{M}{1-\rho^2}$. Later we will demonstrate that the convergence of SG can be proved from a dissipation inequality argument of this type.*

A computational advantage of dissipativity theory is that if the supply rate S is quadratic, we can search over admissible quadratic storage functions V by solving a small semidefinite program. The following approach is standard in the controls literature. See Willems (1972a;b; 2007) for a more comprehensive treatment of dissipativity theory.

Theorem 3. *Suppose $X_j = X_j^\mathsf{T} \in \mathbb{R}^{(n_\xi+n_w)\times(n_\xi+n_w)}$ for $j = 1, 2, \cdots, J$. Define $S_j : \mathbb{R}^{n_\xi} \times \mathbb{R}^{n_w} \to \mathbb{R}$ as*

$$S_j(\xi, w) := \begin{bmatrix} \xi \\ w \end{bmatrix}^\mathsf{T} X_j \begin{bmatrix} \xi \\ w \end{bmatrix}. \tag{5}$$

If there exists a positive semidefinite matrix $P \in \mathbb{R}^{n_\xi \times n_\xi}$ and non-negative scalars λ_j such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - \sum_{j=1}^{J} \lambda_j X_j \preceq 0, \tag{6}$$

then the a.s. exponential dissipation inequality (4) holds for all sample paths of (2) with $V(\xi) := \xi^\mathsf{T} P \xi$ and $S(\xi, w) := \sum_{j=1}^{J} \lambda_j S_j(\xi, w)$. Further assuming that $\mathbb{E}S_j \leq \Lambda_j$ for all sample paths of (2), the following inequality always holds:

$$\mathbb{E}V(\xi_{k+1}) \leq \rho^2 \mathbb{E}V(\xi_k) + \sum_{j=1}^{J} \lambda_j \Lambda_j. \tag{7}$$

Proof. It is straightforward to verify

$$\begin{aligned} V(\xi_{k+1}) &= \xi_{k+1}^\mathsf{T} P \xi_{k+1} \\ &= (A\xi_k + Bw_k)^\mathsf{T} P (A\xi_k + Bw_k) \\ &= \begin{bmatrix} \xi_k \\ w_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} A^\mathsf{T}PA & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} \begin{bmatrix} \xi_k \\ w_k \end{bmatrix}. \end{aligned}$$

Hence, we can left- and right-multiply (6) by $\begin{bmatrix} \xi_k^\mathsf{T} & w_k^\mathsf{T} \end{bmatrix}$ and $\begin{bmatrix} \xi_k^\mathsf{T} & w_k^\mathsf{T} \end{bmatrix}^\mathsf{T}$ to obtain the desired dissipation inequality. Since λ_j is non-negative, we take expectations of the dissipation inequality and obtain (7). ■

If we fix (A, B, X_j, ρ), the condition (6) becomes an LMI with decision variables P and λ_j. For fixed (A, B, X_j, ρ), the feasibility of (6) can be numerically tested using semidefinite programs. When applied to analyze stochastic optimization methods, the resulting LMI is typically small, and can also be solved analytically.

If one only wants to construct the dissipation inequality (4), there is no need to enforce nonnegativity of λ_j. However, we need $\lambda_j \geq 0$ to ensure that the weighted supply rate $S = \sum_{j=1}^{J} \lambda_j S_j$ is useful in convergence analysis.

We will use Theorem 3 to unify the analysis of SVRG and Katyusha. The unified analysis follows four steps.

1. Rewrite the stochastic optimization methods in the form of a stochastic linear system (2).
2. Choose matrices X_j in a way that the supply rate functions (5) satisfy certain desired properties.
3. Solve the LMI (6) to obtain a dissipation inequality that directly yields the so-called one-iteration convergence result.
4. Apply some standard telescoping trick to convert the one-iteration convergence result into a rate bound for the analyzed method.

Step 1 is straightforward. Step 4 has been routinized in the literature. We will show how to perform Steps 2 and 3 for SVRG and Katyusha. Compared with Hu & Lessard (2017), the novelty of the present paper is the development of several new stochastic supply rate conditions that depend on the stochastic variance reduction mechanism.

For illustrative purposes, we first recall the LMI analysis for SG using dissipativity theory.

2.3. Demonstrative Example: Dissipativity for SG

To gain some insight, we first rephrase the LMI-based analysis for SG in (Hu et al., 2017a) using dissipativity theory. SG uses the following iteration:

$$x_{k+1} = x_k - \eta \nabla f_{i_k}(x_k), \tag{8}$$

where i_k is sampled uniformly from $\{1, 2, \ldots, n\}$ at every step. Note that (8) is equivalent to $x_{k+1} - x_\star = x_k - x_\star - \eta \nabla f_{i_k}(x_k)$. Hence we can define $\xi_k = x_k - x_\star$, $w_k = \nabla f_{i_k}(x_k) = \nabla f_{i_k}(\xi_k + x_\star)$, $A = I_p$, and $B = -\eta I_p$. Then the SG iteration (8) is equivalent to the LTI model (2). Based on the properties of f_i and g, we can choose the supply rate functions $S_j(\xi, w)$ based on the following lemma.

Lemma 4. *Let g be L-smooth and σ-strongly convex with $\sigma > 0$. Suppose f_i is L-smooth and convex. Let $x_\star$ be the point satisfying $\nabla g(x_\star) = 0$. Define $X_1 = \bar{X}_1 \otimes I_p$ and $X_2 = \bar{X}_2 \otimes I_p$, where*

$$\bar{X}_1 := \begin{bmatrix} 2\sigma & -1 \\ -1 & 0 \end{bmatrix}, \quad \bar{X}_2 := \begin{bmatrix} 0 & -L \\ -L & 1 \end{bmatrix}. \tag{9}$$

Consider $w_k = \nabla f_{i_k}(\xi_k + x_\star)$ where i_k is sampled uniformly. Define the supply rate functions $S_1(\xi, w)$ and $S_2(\xi, w)$ using (5). Then the following supply rate conditions hold

$$S_1 \le 0, \quad S_2 \le \frac{2}{n} \sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2.$$

Proof. The proof is given in Hu et al. (2017a). For completeness, we include it in the supplementary material. ■

Using Lemma 4, we can apply Theorem 3 to construct a dissipation inequality for SG. Setting $P = I_p$, the LMI condition (6) becomes

$$\begin{bmatrix} 1 - \rho^2 - 2\lambda_1 \sigma & -\eta + \lambda_1 + \lambda_2 L \\ -\eta + \lambda_1 + \lambda_2 L & \eta^2 - \lambda_2 \end{bmatrix} \otimes I_p \preceq 0. \tag{10}$$

Based on Remark 2, we can show $\mathbb{E}\|x_k - x_\star\|^2 \le \rho^{2k}\|x_0 - x_\star\|^2 + \frac{2\lambda_2}{n(1-\rho^2)} \sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2$ by finding non-negative $(\lambda_1, \lambda_2, \rho^2)$ satisfying the above LMI. In fact, the choices $\lambda_1 = \eta - L\eta^2$, $\lambda_2 = \eta^2$, and $\rho^2 = 1 - 2\lambda_1\sigma$ suffice, since they make the left-hand side of (10) zero. We thus obtain the conclusion

$$\mathbb{E}\|x_k - x_\star\|^2 \le (1 - 2\sigma\eta + 2\sigma L\eta^2)^k \|x_0 - x_\star\|^2 + \frac{\eta}{\sigma(1 - L\eta)} \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2 \right),$$

which is the standard convergence result for SG (Needell et al., 2014, Theorem 2.1). Since the supply rate S_2 continues to deliver energy into the system, the SG method with a constant stepsize can only converge to a ball around the optimal point. Later we will see that SVRG-like methods adopt different supply rate functions and eventually reduce their supply energy to 0, enabling linear convergence to the optimal point to be proved.

In this paper, we confine our scope to the case of constant learning rate. For algorithms with time-varying learning rates, one may need to adopt the dissipativity theory for linear time-varying (LTV) systems. This theory requires time-varying Lyapunov functions and infinite-dimensional LMIs. See Hu & Lessard (2017, Section 4.2) for further discussions of this point.

3. Dissipation Inequality for SVRG

In this section, we present a unified LMI-based analysis for SVRG using dissipativity theory. SVRG iterates as follows. Let $\tilde{x}^0 \in \mathbb{R}^p$ be an arbitrary initial point. For each epoch $s = 0, 1, \cdots$, we have $x_0^s = \tilde{x}^s$. For each s, SVRG performs the following steps for $k = 0, 1, \ldots, m-1$:

$$x_{k+1}^s = x_k^s - \eta \left(\nabla f_{i_k^s}(x_k^s) - \nabla f_{i_k^s}(\tilde{x}^s) + \nabla g(\tilde{x}^s) \right),$$

where i_k^s is uniformly sampled from $\{1, 2, \ldots, n\}$ in an IID manner, and m is a prescribed integer determining the epoch length. A popular choice for m is $m = 2n$. At the end of each epoch s, two typical options are available for updating $\tilde{x}^{s+1}$:

- Option I: Set $\tilde{x}^{s+1} = x_m^s$;
- Option II[1]: Set $\tilde{x}^{s+1} = \frac{1}{m} \sum_{k=0}^{m-1} x_k^s$.

When analyzing SVRG, one typically needs to show that there exist $0 \le \nu < 1$ such that

$$\mathbb{E}V(\tilde{x}^{s+1}) \le \nu \mathbb{E}V(\tilde{x}^s), \tag{11}$$

where $V(\tilde{x}^s)$ is set to be either $\|\tilde{x}^s - x_\star\|^2$ or $g(\tilde{x}^s) - g(x_\star)$. Since (11) needs to hold for all s, we can drop the superscript s in the so-called one-iteration analysis, and write each epoch of SVRG in the form of the LTI model (2). Specifically, for a fixed s, we have from the SVRG formula above that

$$x_{k+1} - x_\star = x_k - x_\star + B w_k, \quad k = 0, 1, \ldots, m-1, \tag{12}$$

where $B = \begin{bmatrix} -\eta I_p & -\eta I_p \end{bmatrix}$, and w_k is given as

$$w_k = \begin{bmatrix} \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_\star) \\ \nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x}) \end{bmatrix}. \tag{13}$$

With these choices of w_k and B, we can set $\xi_k = x_k - x_\star$ and $A = I_p$ to recast SVRG in the linear model (2). Next, we will show how to construct supply rate functions for SVRG and apply Theorem 3 to obtain various rate bounds in the form of (11). Our analysis recovers the existing bounds for SVRG, and leads to more general characterizations of the convergence properties of SVRG. We also give physical interpretations for the convergence mechanism of SVRG.

[1] A similar variant with similar analysis is to choose $\tilde{x}^{s+1}$ by sampling uniformly from the iterates in the last epoch.

3.1. Warm-up: Dissipativity for SVRG with Option I

Since we have already rewritten SVRG in the form of the linear model (2), we can construct the dissipation inequality efficiently for SVRG using semidefinite programs in Theorem 3. As before there matrices are derived from propoerties of f_i, $i = 1, 2, \ldots, n$ and g, as we show now.

Lemma 5. *Suppose that g is L-smooth and σ-strongly convex with $\sigma > 0$, and that f_i is L-smooth and convex for $i = 1, 2, \ldots, n$. Suppose that $x_\star$ satisfies $\nabla g(x_\star) = 0$. Set $X_j = \bar{X}_j \otimes I_p$, where $\bar{X}_j$, $j = 1, 2, 3, 4$ are defined as follows:*

$$\bar{X}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \bar{X}_2 = \begin{bmatrix} 2\sigma & -1 & -1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \\ \bar{X}_3 = \begin{bmatrix} 0 & -L & 0 \\ -L & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \bar{X}_4 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}. \tag{14}$$

Consider $\xi_k = x_k - x_\star$ and w_k defined by (13). Suppose the supply rate S_j is defined by (5) for $j = 1, 2, 3, 4$. Then $\mathbb{E}S_1 \leq L^2\mathbb{E}\|\tilde{x} - x_\star\|^2$, $\mathbb{E}S_2 \leq 0$, $\mathbb{E}S_3 \leq 0$, and $\mathbb{E}S_4 = 0$.

Proof. It is straightforward to verify that the proposed supply rate conditions are equivalent to standard inequalities (co-coercivity, etc) in the literature. ■

We will provide more guidelines for supply rate constructions in the supplementary material.

We now apply Theorem 3 to perform LMI-based convergence analysis for SVRG with Option I.

Corollary 6. *Suppose g is σ-strongly convex and L-smooth. In addition, f_i is assumed to be convex and L-smooth. Let $0 \leq \rho^2 < 1$ be given. If there exist nonnegative scalars $(\lambda_1, \lambda_2, \lambda_3)$ and another scalar λ_4 (not necessarily nonnegative) such that*

$$\begin{bmatrix} 1 - \rho^2 - 2\sigma\lambda_2 & \lambda_2 - \eta + L\lambda_3 & \lambda_2 - \eta + \lambda_4 \\ \lambda_2 - \eta + L\lambda_3 & \eta^2 - 2\lambda_3 & \eta^2 \\ \lambda_2 - \eta + \lambda_4 & \eta^2 & \eta^2 - \lambda_1 \end{bmatrix} \preceq 0, \tag{15}$$

then SVRG with Option I satisfies

$$\mathbb{E}\|x_m - x_\star\|^2 \leq \left(\rho^{2m} + \frac{\lambda_1 L^2}{1 - \rho^2}\right) \mathbb{E}\|x_0 - x_\star\|^2. \tag{16}$$

Proof. We choose the supply rate functions S_j for $j = 1, 2, 3, 4$ as described in Lemma 5 and (5). Since $A = I_p$, and $B = \begin{bmatrix} -\eta I_p & -\eta I_p \end{bmatrix}$, we can set $P = I_p$ and show

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} = \begin{bmatrix} 1 - \rho^2 & -\eta & -\eta \\ -\eta & \eta^2 & \eta^2 \\ -\eta & \eta^2 & \eta^2 \end{bmatrix} \otimes I_p.$$

Thus the left-hand side of (15) satisfies (6) if $\lambda_4 \geq 0$. If $\lambda_4 < 0$, we can replace X_4 by $-X_4$ and λ_4 by $-\lambda_4$, and (6) will hold with λ_4 now positive. The conclusion (7) is not affected by the change of sign, since $\mathbb{E}S_4 = 0$, so we can set $\Lambda_4 = 0$ in (7). We have from the conclusion of Theorem 3 that $\mathbb{E}V(\xi_{k+1}) = \mathbb{E}\|x_{k+1} - x_\star\|^2 \leq \rho^2\mathbb{E}\|x_k - x_\star\|^2 + \lambda_1 L^2\mathbb{E}\|x_0 - x_\star\|^2$. We iterate this inequality over $k = 0, 1, \ldots, m-1$ to obtain the result. ■

We can immediately show linear convergence of SVRG with Option I by choosing $\lambda_1 = 2\eta^2$, $\lambda_2 = \eta - L\eta^2$, $\lambda_3 = \eta^2$, $\lambda_4 = L\eta^2$, and $\rho^2 = 1 - 2\sigma(\eta - L\eta^2)$. Then (15) becomes

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & -\eta^2 & \eta^2 \\ 0 & \eta^2 & -\eta^2 \end{bmatrix} \preceq 0,$$

which is clearly true. Hence (11) holds with $V(\tilde{x}^s) = \|\tilde{x}^s - x_\star\|^2$ and ν given by

$$\nu = (1 - 2\eta\sigma(1 - \eta L))^m + \frac{\eta L^2}{\sigma(1 - \eta L)}. \tag{17}$$

This bound slightly improves that of (Tan et al., 2016, Corollary 1). Other bounds under various assumptions are discussed in the supplementary material.

Remark 7. *The important physical insight is provided by the supply rate condition $\mathbb{E}S_1 \leq L^2\mathbb{E}\|\tilde{x} - x_\star\|^2$. Although the supply rate S_1 is delivering energy into the system, the energy supplied is bounded above by $L^2\mathbb{E}\|\tilde{x} - x_\star\|^2$, which diminishes as $\tilde{x}$ approaches $x_\star$. Eventually, the energy supplied by S_1 cannot overcome dissipation.*

3.2. LMI Analysis for SVRG with Option II

For SVRG with Option II, we require the following supply rate functions.

Lemma 8. *Suppose that g is L-smooth and σ-strongly convex with $\sigma > 0$, and that each f_i is L-smooth and convex. Let $x_\star$ be the point satisfying $\nabla g(x_\star) = 0$. Set $X_j = \bar{X}_j \otimes I_p$, where $\bar{X}_j$, $j = 1, 2, 3$ are defined as:*

$$\bar{X}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \bar{X}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \bar{X}_3 = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}. \tag{18}$$

Define $\xi_k = x_k - x_\star$ and define w_k as in (13). Suppose the supply rate S_j is defined by (5) for $j = 1, 2, 3$. Then the following supply rate conditions hold

$$\mathbb{E}S_1 \leq 2L(\mathbb{E}g(x_k) - g(x_\star)), \tag{19a}$$
$$\mathbb{E}S_2 \leq 2L(\mathbb{E}g(\tilde{x}) - g(x_\star)), \tag{19b}$$
$$\mathbb{E}S_3 \leq -\mathbb{E}g(x_k) + g(x_\star). \tag{19c}$$

Proof. These are also standard inequalities in the literature. See Tan et al. (2016) and Bubeck (2015, Lemma 6.4) for more details. (Further discussions are provided in the supplementary material.) ■

Corollary 9. *Suppose that g is σ-strongly convex and L-smooth, and that each f_i is convex and L-smooth. If there exist non-negative scalars λ_j, $j = 1, 2, 3$, such that $\lambda_3 - L\lambda_1 > 0$ and*

$$\begin{bmatrix} 0 & \lambda_3 - \eta & \lambda_3 - \eta \\ \lambda_3 - \eta & \eta^2 - \lambda_1 & \eta^2 \\ \lambda_3 - \eta & \eta^2 & \eta^2 - \lambda_2 \end{bmatrix} \preceq 0, \tag{20}$$

then SVRG with Option II satisfies

$$\mathbb{E}g\left(\frac{1}{m}\sum_{k=0}^{m-1} x_k\right) - g(x_\star) \leq \left(\frac{\sigma^{-1} + mL\lambda_2}{(\lambda_3 - L\lambda_1)m}\right)(\mathbb{E}g(\tilde{x}) - g(x_\star)). \tag{21}$$

Proof. Recall that $A = I_p$, and $B = \begin{bmatrix} -\eta I_p & -\eta I_p \end{bmatrix}$ for the state-space representation of SVRG. Let S_1, S_2, and S_3 be the supply rate functions defined from X_1, X_2, and X_3 of Lemma 8 via (5). Setting $P = I_p$ and $\rho = 1$, the left-hand side of the LMI (6) becomes

$$\begin{bmatrix} 0 & \lambda_3 - \eta & \lambda_3 - \eta \\ \lambda_3 - \eta & \eta^2 - \lambda_1 & \eta^2 \\ \lambda_3 - \eta & \eta^2 & \eta^2 - \lambda_2 \end{bmatrix} \otimes I_p.$$

Since (20) holds, can apply Theorem 3 to show that

$$\mathbb{E}\|x_{k+1} - x_\star\|^2 \leq \mathbb{E}\|x_k - x_\star\|^2 - (2\lambda_3 - 2L\lambda_1)(\mathbb{E}g(x_k) - g(x_\star)) + 2L\lambda_2(\mathbb{E}g(\tilde{x}) - g(x_\star)).$$

We can sum the above inequality from $k = 0$ to $m - 1$ and show that

$$(2\lambda_3 - 2L\lambda_1)\sum_{k=0}^{m-1}(\mathbb{E}g(x_k) - g(x_\star)) \leq \mathbb{E}\|x_0 - x_\star\|^2 + 2mL\lambda_2\mathbb{E}(g(\tilde{x}) - g(x_\star)). \tag{22}$$

By convexity of g, we have

$$g\left(\frac{1}{m}\sum_{k=0}^{m-1} x_k\right) \leq \frac{1}{m}\sum_{k=0}^{m-1} g(x_k).$$

Since g is σ-strongly convex, we also have $\|x_0 - x_\star\|^2 \leq \frac{2}{\sigma}(\mathbb{E}g(x_0) - g(x_\star))$. By substituting these inequalities into (22), and using the assumption $\lambda_3 - L\lambda_1 > 0$, we obtain the result. ■

We can recover the standard rate result for SVRG by choosing $\lambda_1 = \lambda_2 = 2\eta^2$, and $\lambda_3 = \eta$. We have $\lambda_3 - L\lambda_1 = \eta - L\eta^2 \geq 0$ for $\eta \leq \frac{1}{L}$, and (20) becomes

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & -\eta^2 & \eta^2 \\ 0 & \eta^2 & -\eta^2 \end{bmatrix} \preceq 0,$$

which is clearly true. Additionally, we have

$$\frac{\sigma^{-1} + mL\lambda_2}{(\lambda_3 - L\lambda_1)m} = \frac{1}{m\sigma\eta(1 - 2L\eta)} + \frac{2L\eta}{1 - 2L\eta}, \tag{23}$$

which is exactly the rate in (Johnson & Zhang, 2013, Theorem 1). This result states that the iteration complexity of SVRG with Option II is $\mathcal{O}\left((\frac{L}{\sigma} + n)\log(\frac{1}{\varepsilon})\right)$ if we choose $m = \frac{20L}{\sigma}$.

Remark 10. *Some important physical insight is provided by the supply rate condition $\mathbb{E}S_2 \leq 2L\mathbb{E}(g(\tilde{x}) - g(x_\star))$. As $\tilde{x}$ approaches $x_\star$, the energy supplied by S_2 drops and is unable to overcome dissipation, leading to convergence. One may add more supply rate functions and improve the convergence guarantees by some constant factor. In principle, the introduction of more supply rate functions may reduce the conservatism in the analysis. Other choices of λ_j may also change the iteration complexity by a constant factor. In addition, new choices of m may require different choices of λ_j. Note that LMI (6) in Theorem 3 can be implemented and solved numerically, leading to numerical clues for how to construct P and λ_j for proving rate results. Therefore, our proposed LMI provides an efficient tool for constructing bounds of the form (21).*

4. Dissipativity Theory for Katyusha

Katyusha solves the following problem:

$$\min_{x\in\mathbb{R}^p} F(x) := f(x) + \psi(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x), \tag{24}$$

where ψ is σ-strongly convex and possibly nonsmooth, while each f_i, $i = 1, 2, \ldots, n$ is L-smooth and convex.

For each epoch $s = 0, 1, \cdots$, we have $y_0^s = z_0^s = \tilde{x}^s$. For any fixed s, and positive parameters τ_1, τ_2, and α, Katyusha applies the following iteration for $k = 0, 1, \ldots, m - 1$:

$$x_{k+1}^s = \tau_1 z_k^s + \tau_2 \tilde{x}^s + (1 - \tau_1 - \tau_2)y_k^s, \tag{25a}$$

$$v_k^s = \nabla f_{i_k^s}(x_{k+1}^s) - \nabla f_{i_k^s}(\tilde{x}^s) + \nabla f(\tilde{x}^s), \tag{25b}$$

$$z_{k+1}^s = \arg\min_z \left\{\frac{1}{2\alpha}\|z - z_k^s\|^2 + (v_k^s)^\mathsf{T} z + \psi(z)\right\}, \tag{25c}$$

$$y_{k+1}^s = \arg\min_y \left\{\frac{3L}{2}\|y - x_{k+1}^s\|^2 + (v_k^s)^\mathsf{T} y + \psi(y)\right\}, \tag{25d}$$

where i_k^s is uniformly sampled from $\{1, 2, \ldots, n\}$ in an i.i.d. manner, and m is a prescribed integer determining the length of the epoch. (A popular choice is $m = 2n$.) At the end of each epoch s, we set

$$\tilde{x}^{s+1} = \left(\sum_{j=0}^{m-1}(1+\sigma\alpha)^j\right)^{-1}\left(\sum_{j=0}^{m-1}(1+\alpha\sigma)^j y_{j+1}^s\right). \tag{26}$$

Allen-Zhu (2016) shows that the iteration complexity for Katyusha is $\mathcal{O}\left(\left(\sqrt{\frac{Ln}{\sigma}}+n\right)\log(\frac{1}{\varepsilon})\right)$ if one chooses $\tau_2 = \frac{1}{2}$, $\tau_1 = \min\{\sqrt{\frac{m\sigma}{3L}}, \frac{1}{2}\}$, $\alpha = \frac{1}{3\tau_1 L}$, and $m = 2n$. The key of the proof is the coupling lemma (Allen-Zhu, 2016, Lemma 3.7), which states the following holds for Katyusha with $\tau_1 \le \frac{1}{3\alpha L}$ and $\tau_2 = \frac{1}{2}$:

$$\begin{aligned}&\frac{1+\alpha\sigma}{2}\mathbb{E}\|z_{k+1}-x_\star\|^2 + \frac{\alpha}{\tau_1}\left(\mathbb{E}F(y_{k+1}) - F(x_\star)\right)\\ &-\frac{1}{2}\mathbb{E}\|z_k - x_\star\|^2 - \frac{\alpha(1-\tau_1-\tau_2)}{\tau_1}\left(\mathbb{E}F(y_k)-F(x_\star)\right)\\ &\le \frac{\alpha\tau_2}{\tau_1}\left(\mathbb{E}F(\tilde{x}) - F_\star\right).\end{aligned} \tag{27}$$

We analyze a single epoch, dropping the superscript s to simplify the notation. As stated in (Allen-Zhu, 2016, Section 3.2), once the above one-iteration convergence result is established, a telescoping trick can be applied to show the improved iteration complexity of Katyusha. We show how to provide a general proof for (27) using dissipativity, with Theorem 3 again being our main technical tool.

4.1. Katyusha as a Stochastic System

At a given epoch s (subscript dropped), a single "inner" iteration of Katyusha can be written as follows:

$$x_{k+1} = \tau_1 z_k + \tau_2 \tilde{x} + (1-\tau_1-\tau_2)y_k, \tag{28a}$$
$$v_k = \nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(\tilde{x}) + \nabla f(\tilde{x}), \tag{28b}$$
$$z_{k+1} = z_k - \alpha v_k - \alpha g_k, \tag{28c}$$
$$y_{k+1} = x_{k+1} - \zeta v_k - \zeta h_k, \tag{28d}$$

where g_k is some subgradient of ψ evaluated at z_{k+1}, and h_k is some subgradient of ψ evaluated at y_{k+1}. We can set $\zeta = \frac{1}{3L}$ to recover the standard Katyusha iteration 25a.

We can rewrite (28) as

$$\begin{bmatrix} z_{k+1}-x_\star \\ y_{k+1}-x_\star \\ \tilde{x}-x_\star\end{bmatrix} = A\begin{bmatrix} z_k-x_\star \\ y_k-x_\star \\ \tilde{x}-x_\star\end{bmatrix} + B\begin{bmatrix} v_k \\ g_k \\ h_k\end{bmatrix},$$

where $A = \bar{A}\otimes I_p$ and $B = \bar{B}\otimes I_p$, and $\bar{A}$ and $\bar{B}$ are defined as follows:

$$\bar{A} = \begin{bmatrix} 1 & 0 & 0 \\ \tau_1 & 1-\tau_1-\tau_2 & \tau_2 \\ 0 & 0 & 1\end{bmatrix},\quad \bar{B} = \begin{bmatrix} -\alpha & -\alpha & 0 \\ -\zeta & 0 & -\zeta \\ 0 & 0 & 0\end{bmatrix}.$$

Based on the iteration above, it is straightforward to check that Katyusha (28) is equivalent to the stochastic linear system (2) with

$$\xi_k = \begin{bmatrix} z_k - x_\star \\ y_k - x_\star \\ \tilde{x} - x_\star\end{bmatrix},\quad w_k = \begin{bmatrix} v_k \\ g_k \\ h_k\end{bmatrix}. \tag{29}$$

4.2. Supply Rate Functions for Katyusha

Katyusha extracts energy out of the system much faster than SVRG, as can be shown by the use of more advanced supply rate functions.

Lemma 11. *Let ψ be σ-strongly convex with $\sigma > 0$. Suppose f_i is L-smooth and convex. Let $x_\star$ be the optimal point of F. Define $X_1 = \bar{X}_1 \otimes I_p$, where $\bar{X}_1$ is the following sum of four matrices:*

$$\begin{aligned}\bar{X}_1 = &-\begin{bmatrix} -\frac{\sigma\tau_1}{2} & 0 & 0 & \frac{\tau_1(\alpha\sigma+1)}{2} & \frac{\tau_1(\alpha\sigma+1)}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{\tau_1(\alpha\sigma+1)}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{\tau_1(\alpha\sigma+1)}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}\\ &+\left(\zeta - \frac{\alpha\tau_1}{2} - \frac{L\zeta^2(1+\tau_2)}{2\tau_2}\right)\begin{bmatrix} 0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&1\\0&0&0&0&0&0\\0&0&0&1&0&1\end{bmatrix}\\ &+\frac{\alpha\tau_1(\alpha\sigma+1)}{2}\begin{bmatrix} 0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&1&0\\0&0&0&1&1&0\\0&0&0&0&0&0\end{bmatrix}\\ &+\frac{\alpha\tau_1}{2}\begin{bmatrix} 0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&1&-1\\0&0&0&0&-1&1\end{bmatrix}.\end{aligned} \tag{30}$$

Consider ξ_k and w_k defined by (29). Suppose the supply rate S_j is defined by (5) for $j = 1$. Then the following supply rate condition holds for Katyusha

$$\begin{aligned}&\mathbb{E}S_1(\xi_k, w_k) \le (1-\tau_1-\tau_2)(\mathbb{E}F(y_k) - F(x_\star))\\ &-(\mathbb{E}F(y_{k+1}) - F(x_\star)) + \tau_2(\mathbb{E}F(\tilde{x}) - F(x_\star)).\end{aligned} \tag{31}$$

Proof. The proof is based on the strong-convexity of ψ, and the smoothness and convexity of f_i. The detailed proof is presented in the supplementary material. ■

The physical interpretation for the above supply rate is as follows. There is some hidden energy in the system that takes the form of $F(y_k) - F(x_\star)$. There is also some initial energy in the form of $F(\tilde{x}) - F(x_\star)$. The above supply rate condition states that the delivered energy is bounded by a weighted decrease of the hidden energy plus some amount of the initial energy. Such a supply rate can efficiently extract energy out of the systems due to its coupling with the hidden energy and the initial energy. The supply rate construction in Lemma 11 is quite similar to the supply rate construction for Nesterov's accelerated method (Hu & Lessard, 2017). From a physical viewpoint, the essential property of momentum terms can extract the hidden energy out of the system in a more efficient way.

Remark 12. *Although the supply rate in Lemma 11 is complicated, there are some general guidelines for constructing and choosing supply rates. We discuss these guidelines in the supplementary materials.*

4.3. Analysis of Katyusha Using Dissipativity

Using the supply rate function in Lemma 11, we can immediately recover the one-iteration result (27) as follows. Suppose $\tau_2 = \frac{1}{2}$ and $\zeta = \frac{1}{3L}$. We choose $\rho^2 = \frac{1}{1+\alpha\sigma}$, $\lambda_1 = \frac{\alpha}{\tau_1}$, and

$$P = \frac{1+\alpha\sigma}{2}\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \otimes I_p. \tag{32}$$

Then the left-hand side of the LMI condition (6) becomes

$$\frac{\alpha}{2}\left(\alpha - \frac{1}{3L\tau_1}\right)\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&1\\0&0&0&0&0&0\\0&0&0&1&0&1\end{bmatrix} \otimes I_p + \frac{\alpha^2}{2}\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&-1&1\\0&0&0&0&1&-1\end{bmatrix} \otimes I_p,$$

which is clearly negative semidefinite when $\tau_1 \leq \frac{1}{3\alpha L}$.

Therefore, we can apply Theorem 3 to prove that

$$\frac{1+\alpha\sigma}{2}\mathbb{E}\|z_{k+1} - x_\star\|^2 - \frac{1}{2}\mathbb{E}\|z_k - x_\star\|^2 \leq \frac{\alpha}{\tau_1}\mathbb{E}S(\xi_k, w_k).$$

From the supply rate condition (31), we immediately recover the one-iteration analysis result (27), which can be easily transferred into the iteration complexity result by applying the telescoping trick in Allen-Zhu (2016).

We emphasize that Lemma 11 works for general choices of ζ and τ_2. Due to the generality of Lemma 11, our LMI approach can be used to generalize (Allen-Zhu, 2016, Lemma 3.7) for many more choices of (τ_1, τ_2) . This could lead to other choices of $(\tau_1, \tau_2, \alpha, \zeta)$, which yields the same accelerated iteration complexity. However, those choices of parameters will at most improve the iteration complexity by a constant factor. For example, consider $\zeta = \frac{1}{3L}$ and any $\tau_2 \geq \frac{1}{5}$. We can still choose $\rho^2 = \frac{1}{1+\alpha\sigma}$, $\lambda_1 = \frac{\alpha}{\tau_1}$, and P as defined in (32) to prove (27). In this case, the left-hand side of the LMI condition (6) becomes

$$\frac{\alpha}{2}\left(\alpha - \frac{5\tau_2 - 1}{9L\tau_1\tau_2}\right)\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&1\\0&0&0&0&0&0\\0&0&0&1&0&1\end{bmatrix} \otimes I_p + \frac{\alpha^2}{2}\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&-1&1\\0&0&0&0&1&-1\end{bmatrix} \otimes I_p,$$

which is clearly negative semidefinite when $\tau_1 \leq \frac{5\tau_2-1}{9\alpha L\tau_2}$. Therefore, the one-iteration convergence result (27) holds for any $\frac{1}{5} \leq \tau_2 < 1$ and $\tau_1 \leq \min\{\frac{5\tau_2-1}{9\alpha L\tau_2}, 1-\tau_2\}$. This generalizes the coupling lemma (Allen-Zhu, 2016, Lemma 3.7) to more general choices of (τ_1, τ_2). Based on this, one can use the telescoping trick to show Katyusha with $\tau_2 \neq \frac{1}{2}$ can also achieve the iteration complexity of $\mathcal{O}\left((\sqrt{\frac{Ln}{\sigma}} + n)\log(\frac{1}{\varepsilon})\right)$. More details are provided in the supplementary material.

5. Future Work

We plan to use our techniques to study optimal tuning of Katyusha X (Allen-Zhu, 2018) for ERM problems in which the component functions f_i are not individually convex. In addition, we are interested in investigating how to accelerate other recently-developed methods such as SARAH (Nguyen et al., 2017) using our LMI approach. It is also important to extend our control framework for understanding other accelerating mechanism such as Catalyst (Lin et al., 2015).

Notice that deterministic continuous-time algorithms have also been understood as dissipative dynamical systems (Attouch et al., 2000; Alvarez et al., 2002; Hu & Lessard, 2017). It is possible that one can modify the proposed framework to study stochastic continuous-time dynamics. This is another important future direction.

Acknowledgments

Bin Hu and Laurent Lessard are supported by the National Science Foundation (NSF) under Grants No. 1656951 and 1750162. Bin Hu and Laurent Lessard also acknowledge support from the Wisconsin Institute for Discovery, the College of Engineering, and the Department of Electrical and Computer Engineering at the University of Wisconsin–Madison. Stephen Wright was supported by NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; AFOSR Award FA9550-13-1-0138; Subcontracts 3F-30222 and 8F-30039 from Argonne National Laboratory; and DARPA Award N660011824020.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *arXiv preprint arXiv:1603.05953*, 2016.

Allen-Zhu, Z. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.

Allen-Zhu, Z. and Orecchia, L. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Alvarez, F., Attouch, H., Bolte, J., and Redont, P. A second-order gradient-like dissipative dynamical system with hessian-driven damping.-application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–780, 2002.

Attouch, H., Goudou, X., and Redont, P. The heavy ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

Bottou, L. and LeCun, Y. Large scale online learning. In *Advances in neural information processing systems*, pp. 217–224, 2003.

Bottou, L., Curtis, F., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (3-4):231–357, 2015.

Bubeck, S., Lee, Y., and Singh, M. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 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*, 2014a.

Defazio, A., Domke, J., and Caetano, T. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1125–1133, 2014b.

Drusvyatskiy, D., Fazel, M., and Roy, S. An optimal first order method based on optimal quadratic averaging. *arXiv preprint arXiv:1604.06543*, 2016.

Fazlyab, M., Ribeiro, A., Morari, M., and Preciado, V. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *arXiv preprint arXiv:1705.03615*, 2017.

Hu, B. and Lessard, L. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Hu, B., Seiler, P., and Lessard, L. Analysis of approximate stochastic gradient using quadratic constraints and sequential semidefinite programs. *arXiv preprint arXiv:1711.00987*, 2017a.

Hu, B., Seiler, P., and Rantzer, A. A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In *Conference on Learning Theory*, pp. 1157–1189, 2017b.

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

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015.

Lin, Q., Lu, Z., and Xiao, L. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pp. 3059–3067, 2014.

Needell, D., Ward, R., and Srebro, N. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pp. 1017–1025, 2014.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2003.

Nguyen, L., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2613–2621, 2017.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems*, 2012.

Schmidt, M., Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *ArXiv preprint*, 2013.

Shalev-Shwartz, S. SDCA without duality, regularization, and individual convexity. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 747–754, 2016.

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

Su, W., Boyd, S., and Candès, E. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Tan, C., Ma, S., Dai, Y., and Qian, Y. Barzilai-borwein step size for stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 685–693, 2016.

Wibisono, A., Wilson, A., and Jordan, M. A variational perspective on accelerated methods in optimization. *arXiv preprint arXiv:1603.04245*, 2016.

Willems, J. Dissipative dynamical systems part i: General theory. *Archive for Rational Mech. and Analysis*, 45(5): 321–351, 1972a.

Willems, J. Dissipative dynamical systems part ii: Linear systems with quadratic supply rates. *Archive for Rational Mech. and Analysis*, 45(5):352–393, 1972b.

Willems, J. Dissipative dynamical systems. *European Journal of Control*, 13(2-3):134–151, 2007.

Wilson, A., Recht, B., and Jordan, M. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Supplementary Material

The underlying probability space for the sampling index i_k is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $\mathcal{F}_k$ the σ-algebra generated by $(i_0, i_1, \ldots, i_k)$. Clearly, i_k is $\mathcal{F}_k$-adapted and we obtain a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$ on which the stochastic optimization method is defined.

A. Proof of Lemma 4

The proof is straightforward and included here only for completeness. Note that x_k does not depend on i_k, so we have $\mathbb{E}\left[(x_k - x_\star)^\mathsf{T} \nabla f_{i_k}(x_k) \middle| \mathcal{F}_{k-1}\right] = (x_k - x_\star)^\mathsf{T} \nabla g(x_k)$. If g is σ-strongly convex, we directly have

$$\mathbb{E}\left[\begin{bmatrix} x_k - x_\star \\ \nabla f_{i_k}(x_k) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 2\sigma & -1 \\ -1 & 0 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ \nabla f_{i_k}(x_k) \end{bmatrix}\right] = \mathbb{E}\left[\begin{bmatrix} x_k - x_\star \\ \nabla g(x_k) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 2\sigma & -1 \\ -1 & 0 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ \nabla g(x_k) \end{bmatrix}\right] \leq 0.$$

Next, if f_i is convex and L-smooth, the co-coercivity property implies

$$\begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 0 & -L \\ -L & 2 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) - \nabla f_i(x_\star) \end{bmatrix} \leq 0.$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left(\begin{bmatrix} x_k - x_\star \\ \nabla f_{i_k}(x_k) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 0 & -L \\ -L & 1 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ \nabla f_{i_k}(x_k) \end{bmatrix} \middle| \mathcal{F}_{k-1}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} 0 & -L \\ -L & 0 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ \nabla f_i(x_k) \end{bmatrix} + \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_k)\|^2 \\
&\leq -\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_k) - \nabla f_i(x_\star)\|^2 + \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_k)\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2.
\end{aligned}$$

Taking the expectation of the above inequality leads to the desired conclusion.

B. Proof of Lemma 5 and Lemma 8

We summarize some existing function inequalities that can be used to directly show Lemma 5 and Lemma 8.

Lemma S1. *Assume $\nabla g(x_\star) = 0$. Suppose i_k is uniformly sampled from $\{1, \ldots, n\}$ in an i.i.d. manner. Let $\{x_k : k = 0, 1, \ldots\}$ be an $\mathcal{F}_n$- predictable process whose sample path satisfies $x_k \in \mathbb{R}^p$ almost surely. In addition, $r_k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_\star)$ and $u_k = \nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x})$, where $\tilde{x}$ is $\mathcal{F}_0$-measurable.*

1. *The following always holds due to the uniform sampling strategy:*

$$\mathbb{E}\left[(x_k - x_\star)^\mathsf{T}(\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x}))\right] = 0. \tag{S1}$$

2. *If f_i is L-smooth, then*

$$\mathbb{E}\|\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x})\|^2 \leq L^2 \mathbb{E}\|\tilde{x} - x_\star\|^2. \tag{S2}$$

3. *If f_i is convex and L-smooth, then*

$$\mathbb{E}\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_\star)\|^2 \leq 2L(\mathbb{E}g(x_k) - g(x_\star)), \tag{S3}$$

$$\mathbb{E}\|\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x})\|^2 \leq 2L(\mathbb{E}g(\tilde{x}) - g(x_\star)). \tag{S4}$$

4. *The following inequality holds*

$$\mathbb{E}\left[\begin{bmatrix} x_k - x_\star \\ r_k \end{bmatrix}^{\mathsf{T}} (M \otimes I_p) \begin{bmatrix} x_k - x_\star \\ r_k \end{bmatrix}\right] \leq 0, \tag{S5}$$

where M is computed according to the assumption on f_i as follows

$$M := \begin{cases} \begin{bmatrix} 2\sigma L & -(\sigma+L) \\ -(\sigma+L) & 2 \end{bmatrix} & \text{if } f_i \text{is } L\text{-smooth and } \sigma\text{-strongly convex}, \\ \begin{bmatrix} 0 & -L \\ -L & 2 \end{bmatrix} & \text{if } f_i \text{is } L\text{-smooth and convex}, \\ \begin{bmatrix} -2L^2 & 0 \\ 0 & 2 \end{bmatrix} & \text{if } f_i \text{ is } L\text{-smooth}. \end{cases} \tag{S6}$$

5. *If g is σ-strongly convex, we have*

$$\mathbb{E}\left[\begin{bmatrix} x_k - x_\star \\ r_k \end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix} 2\sigma & -1 \\ -1 & 0 \end{bmatrix} \otimes I_p\right) \begin{bmatrix} x_k - x_\star \\ r_k \end{bmatrix}\right] \leq 0. \tag{S7}$$

6. *If g is convex, then*

$$\mathbb{E}\left[(x_k - x_\star)^{\mathsf{T}}(\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x}))\right] \geq \mathbb{E}g(x_k) - g(x_\star). \tag{S8}$$

7. *If g is σ-strongly convex, then*

$$\mathbb{E}\|\tilde{x} - x_\star\|^2 \leq \frac{2}{\sigma}\left(\mathbb{E}g(\tilde{x}) - g(x_\star)\right). \tag{S9}$$

Proof. The proof is standard and based on the fact that i_k and x_k are independent. For example, we have

$$\mathbb{E}\left[(x_k - x_\star)^{\mathsf{T}}(\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x}))\middle|\, \mathcal{F}_{k-1}\right] = (x_k - x_\star)^{\mathsf{T}}\nabla g(x_\star) = 0,$$

which directly leads to Statement 1. Note that $\mathbb{E}\left[\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x})\right] = -\mathbb{E}\nabla g(\tilde{x})$. Hence, we have

$$\mathbb{E}\|\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x}) + \nabla g(\tilde{x})\|^2 \leq \mathbb{E}\|\nabla f_{i_k}(x_\star) - \nabla f_{i_k}(\tilde{x})\|^2 \leq L^2\mathbb{E}\|\tilde{x} - x_\star\|^2,$$

which proves Statement 2. The other statements follow from taking expectations of well known function inequalities. ■

The proofs of Lemma 5 and Lemma 8 directly follow from the lemma above.

C. Further Discussion on SVRG

One can automate the convergence analysis for SVRG under various assumptions on f_i. For example, consider the analysis of SVRG with Option I. If f_i is assumed only to be L-smooth, we can modify $\bar{X}_3$ in Lemma 5 as

$$\bar{X}_3 = \begin{bmatrix} -2L^2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We still assume that g is L-smooth and σ-strongly convex, so we choose $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_4$ as in Lemma 5. For these choices, it is still true that $\mathbb{E}S_1 \leq L^2\mathbb{E}\|\tilde{x} - x_\star\|^2$, $\mathbb{E}S_2 \leq 0$, $\mathbb{E}S_3 \leq 0$, and $\mathbb{E}S_4 = 0$. The usual analysis route leads to the following bound:

$$\mathbb{E}\|x_m - x_\star\|^2 \leq \left((1 - 2\sigma\eta + 2L^2\eta^2)^m + \frac{\eta L^2}{\sigma - \eta L^2}\right)\mathbb{E}\|x_0 - x_\star\|^2.$$

This example demonstrates that one can modify the supply rate functions to reflect various assumptions on the cost functions. For SVRG with Option II, one can perform similar LMI analysis when the assumptions on f_i are changed.

D. Proof of Lemma 11

We first set

$$q_k = \begin{bmatrix} \tau_1 & 1-\tau_1-\tau_2 & \tau_2 \end{bmatrix} \begin{bmatrix} z_k \\ y_k \\ \tilde{x} \end{bmatrix}. \tag{S10}$$

From the definition of Katyusha, we have $\mathbb{E}v_k = \mathbb{E}\nabla f(q_k)$. Since f is L-smooth and convex, it is straightforward to verify the following:

$$\mathbb{E}f(q_k) - \mathbb{E}f(y_k) \leq \mathbb{E}\nabla f(q_k)^{\mathsf{T}}(q_k - y_k) = \mathbb{E}\big[\mathbb{E}[v_k^{\mathsf{T}}(q_k - y_k)|\mathcal{F}_{i_{k-1}}]\big] = \mathbb{E}v_k^{\mathsf{T}}(q_k - y_k), \tag{S11}$$

$$\mathbb{E}f(q_k) - \mathbb{E}f(x_\star) \leq \mathbb{E}\nabla f(q_k)^{\mathsf{T}}(q_k - x_\star) = \mathbb{E}v_k^{\mathsf{T}}(q_k - x_\star), \tag{S12}$$

$$\begin{aligned}
\mathbb{E}f(y_{k+1}) - \mathbb{E}f(q_k) &\leq \mathbb{E}\left[\nabla f(q_k)^{\mathsf{T}}(y_{k+1} - q_k) + \frac{L}{2}\|y_{k+1} - q_k\|^2\right] \\
&= \mathbb{E}\left[(\nabla f(q_k) - v_k)^{\mathsf{T}}(y_{k+1} - q_k) + v_k^{\mathsf{T}}(y_{k+1} - q_k) + \frac{L}{2}\|y_{k+1} - q_k\|^2\right] \\
&\leq \frac{\tau_2}{2L}\mathbb{E}\|v_k - \nabla f(q_k)\|^2 + \frac{L}{2}\left(1 + \frac{1}{\tau_2}\right)\mathbb{E}\|y_{k+1} - q_k\|^2 + \mathbb{E}v_k^{\mathsf{T}}(y_{k+1} - q_k) \\
&\leq \tau_2(\mathbb{E}f(\tilde{x}) - \mathbb{E}f(q_k) - \mathbb{E}v_k^{\mathsf{T}}(\tilde{x} - q_k)) + \frac{L}{2}\left(1 + \frac{1}{\tau_2}\right)\mathbb{E}\|y_{k+1} - q_k\|^2 + \mathbb{E}v_k^{\mathsf{T}}(y_{k+1} - q_k),
\end{aligned} \tag{S13}$$

where the second-last inequality follows from the identity $a^T b \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, and the final step follows from the so-called variance upper bound in the literature (Lemma 3.4 of (Allen-Zhu, 2016)).

To prove Lemma 11, we need to show that

$$(\mathbb{E}F(y_{k+1}) - F(x_\star)) - (1-\tau_1-\tau_2)(\mathbb{E}F(y_k) - F(x_\star)) - \tau_2(\mathbb{E}F(\tilde{x}) - F(x_\star)) \leq -\mathbb{E}S_1(\xi_k, w_k). \tag{S14}$$

For brevity, define $\tilde{\tau} := 1 - \tau_1 - \tau_2$. The left side of (S14) can be rewritten as

$$\begin{aligned}
&(\mathbb{E}F(y_{k+1}) - F(x_\star)) - (1-\tau_1-\tau_2)(\mathbb{E}F(y_k) - F(x_\star)) - \tau_2(\mathbb{E}F(\tilde{x}) - F(x_\star)) \\
&= \mathbb{E}f(y_{k+1}) + \mathbb{E}\psi(y_{k+1}) - \tilde{\tau}\mathbb{E}f(y_k) - \tilde{\tau}\mathbb{E}\psi(y_k) - \tau_1 f(x_\star) - \tau_1\psi(x_\star) - \tau_2\mathbb{E}f(\tilde{x}) - \tau_2\mathbb{E}\psi(\tilde{x}) \\
&= \big(\mathbb{E}f(y_{k+1}) - \tilde{\tau}\mathbb{E}f(y_k) - \tau_1 f(x_\star) - \tau_2\mathbb{E}f(\tilde{x})\big) + \big(\mathbb{E}\psi(y_{k+1}) - \tilde{\tau}\mathbb{E}\psi(y_k) - \tau_1\psi(x_\star) - \tau_2\mathbb{E}\psi(\tilde{x})\big).
\end{aligned} \tag{S15}$$

We have decoupled the left side of (S14) into the sum of two terms, the first involving only f, and the second involving only ψ. We will use the properties of f and ψ to provide upper bounds in the quadratic forms for the first and second terms, respectively.

Bounding the first term in (S15), we obtain

$$\begin{aligned}
&\mathbb{E}f(y_{k+1}) - \tilde{\tau}\mathbb{E}f(y_k) - \tau_1 f(x_\star) - \tau_2\mathbb{E}f(\tilde{x}) \\
&= \mathbb{E}\left[f(y_{k+1}) - f(q_k) + \tau_2(f(q_k) - f(\tilde{x})) + \tau_1(f(q_k) - f(x_\star)) + \tilde{\tau}(f(q_k) - f(y_k))\right] \\
&\leq \frac{L}{2}\left(1 + \frac{1}{\tau_2}\right)\mathbb{E}\|y_{k+1} - q_k\|^2 + \mathbb{E}v_k^{\mathsf{T}}(y_{k+1} - q_k) + \tau_2\mathbb{E}v_k^{\mathsf{T}}(q_k - \tilde{x}) + \tau_1\mathbb{E}v_k^{\mathsf{T}}(q_k - x_\star) + \tilde{\tau}\mathbb{E}v_k^{\mathsf{T}}(q_k - y_k),
\end{aligned} \tag{S16}$$

where the last step follows from the three bounds (S11), (S12), and (S13). Next, strong convexity of ψ leads to an upper bound for the second term in (S15):

$$\begin{aligned}
&\mathbb{E}\psi(y_{k+1}) - \tilde{\tau}\mathbb{E}\psi(y_k) - \tau_1\psi(x_\star) - \tau_2\mathbb{E}\psi(\tilde{x}) \\
&= \mathbb{E}\left[\tilde{\tau}(\psi(y_{k+1}) - \psi(y_k)) + \tau_1(\psi(y_{k+1}) - \psi(z_{k+1})) + \tau_1(\psi(z_{k+1}) - \psi(x_\star)) + \tau_2(\psi(y_{k+1}) - \psi(\tilde{x}))\right] \\
&\leq \mathbb{E}\big[\tilde{\tau}\, h_k^{\mathsf{T}}(y_{k+1} - y_k) + \tau_1\, h_k^{\mathsf{T}}(y_{k+1} - z_{k+1}) + \tau_1\left(g_k^{\mathsf{T}}(z_{k+1} - x_\star) - \frac{\sigma}{2}\|z_{k+1} - x_\star\|^2\right) + \tau_2\, h_k^{\mathsf{T}}(y_{k+1} - \tilde{x})\big].
\end{aligned} \tag{S17}$$

Combining (S16)–(S17), we see that the left side of (S14) is bounded above by the expected value of the following sum:

$$\frac{L}{2}\left(1+\frac{1}{\tau_2}\right)\|y_{k+1}-q_k\|^2 + v_k^{\mathsf{T}}(y_{k+1}-q_k) + \tau_2\, v_k^{\mathsf{T}}(q_k-\tilde{x}) + \tau_1\, v_k^{\mathsf{T}}(q_k-x_\star) + \tilde{\tau}\, v_k^{\mathsf{T}}(q_k-y_k)$$
$$+ \tilde{\tau}\, h_k^{\mathsf{T}}(y_{k+1}-y_k) + \tau_1\, h_k^{\mathsf{T}}(y_{k+1}-z_{k+1}) + \tau_1\left(g_k^{\mathsf{T}}(z_{k+1}-x_\star) - \frac{\sigma}{2}\|z_{k+1}-x_\star\|^2\right) + \tau_2\, h_k^{\mathsf{T}}(y_{k+1}-\tilde{x}). \quad \text{(S18)}$$

All terms in (S18) are actually quadratic forms, due to the state-space model:

$$\begin{bmatrix} z_{k+1}-x_\star \\ y_{k+1}-x_\star \\ \tilde{x}-x_\star \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \tau_1 & \tilde{\tau} & \tau_2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} z_k-x_\star \\ y_k-x_\star \\ \tilde{x}-x_\star \end{bmatrix} + \begin{bmatrix} -\alpha & -\alpha & 0 \\ -\zeta & 0 & -\zeta \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} v_k \\ g_k \\ h_k \end{bmatrix},$$
$$q_k - x_\star = \begin{bmatrix} \tau_1 & \tilde{\tau} & \tau_2 \end{bmatrix}\begin{bmatrix} z_k-x_\star \\ y_k-x_\star \\ \tilde{x}-x_\star \end{bmatrix},$$

where we recall the definition $\tilde{\tau} := 1-\tau_1-\tau_2$. For example, the term $v_k^{\mathsf{T}}(y_{k+1}-q_k)$ is equivalent to the quadratic form:

$$\begin{bmatrix} z_k-x_\star \\ y_k-x_\star \\ \tilde{x}-x_\star \\ v_k \\ g_k \\ h_k \end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\zeta & 0 & -\frac{\zeta}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{\zeta}{2} & 0 & 0 \end{bmatrix} \otimes I_p \right) \begin{bmatrix} z_k-x_\star \\ y_k-x_\star \\ \tilde{x}-x_\star \\ v_k \\ g_k \\ h_k \end{bmatrix}.$$

Summing all the these quadratic forms directly yields the desired supply rate.

E. Guidelines for Constructing and Choosing Supply Rates

In most cases, supply rates may be constructed by manipulating well-known quadratic inequalities. One can see this in the proof of Lemma 5 and Lemma 8. For momentum methods, the supply rate construction is more involved. One typically needs to regroup terms carefully after adding and subtracting $f(q_k)$, where q_k is the input to the stochastic gradient. See (S16) for such an example. We note that it is possible for different supply rate functions to yield the same iteration complexity bound. It is also possible to construct other supply rate functions that yield a constant-factor improvement for the convergence guarantees of Katyusha. In the present work, we only provide one supply rate for the analysis of Katyusha.

The selections of supply rate functions for a particular algorithm can be guided by the numerical solutions of the proposed LMIs. For example, one could include several candidate supply rates with associated multipliers λ_j in the LMI to identify which supply rate functions are needed to obtain the desired rate bound.

F. Telescoping Trick and Further Discussion on Katyusha

The telescoping trick in Allen-Zhu (2016, Section 3.2) provides a routine for converting the one-iteration analysis result into a complexity bound. We first fix $\zeta = \frac{1}{3L}$. Given $\frac{1}{5} \le \tau_2 < 1$, we choose $\tau_1 = \min\left\{\sqrt{\frac{(5\tau_2-1)m\sigma}{9\tau_2 L}}, 1-\tau_2\right\}$ and $\alpha = \frac{5\tau_2-1}{9\tau_1\tau_2 L}$. Then the telescoping argument in (Allen-Zhu, 2016, Section 3.2) leads to the following discussion of the resultant iteration complexity $\mathcal{O}\left(\left(\sqrt{\frac{Ln}{\sigma}}+n\right)\log(\frac{1}{\varepsilon})\right)$.

Case 1. Suppose $\frac{m\sigma}{L} \le \frac{9\tau_2(1-\tau_2)^2}{5\tau_2-1}$. We have $\alpha = \sqrt{\frac{5\tau_2-1}{9Lm\sigma\tau_2}}$, and $\tau_1 = m\sigma\alpha \le 1-\tau_2$. Hence $\alpha\sigma \le \frac{1-\tau_2}{m}$. This guarantees the following inequality,

$$(1+\sigma\alpha)^{m-1} \le 1 + \frac{1}{\tau_2}(m-1)\alpha\sigma.$$

Then the argument in Case 1 of (Allen-Zhu, 2016, Section 3.2) can be modified to show

$$\mathbb{E}[F(\tilde{x}^s) - F(x_\star)] \leq O\left(\left(1 + \sqrt{\frac{(5\tau_2 - 1)\sigma}{9\tau_2 L m}}\right)^{-sm}\right)(F(x_0) - F(x_\star))\,.$$

Case 2. Suppose $\frac{m\sigma}{L} > \frac{9\tau_2(1-\tau_2)^2}{5\tau_2-1}$. We have $\tau_1 = 1 - \tau_2$ and $\alpha = \frac{5\tau_2-1}{9(1-\tau_2)\tau_2 L}$. Tailoring the argument in Case 2 of (Allen-Zhu, 2016, Section 3.2), we can easily show

$$\mathbb{E}[F(\tilde{x}^s) - F(x_\star)] \leq O\left(\min\{1/\tau_2, 2-\tau_2\}^{-s}\right)(F(x_0) - F(x_\star)) = O\left((2-\tau_2)^{-s}\right)(F(x_0) - F(x_\star))\,.$$