# 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 $\sigma$-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 $\sigma$-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 $\sigma$-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 $\sigma$-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 $\sigma$-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 $\psi$. We will use the properties of $f$ and $\psi$ 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 $\psi$ 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 $\lambda_j$ 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

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

## References

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