Random Reshuffling with Variance Reduction: New Analysis and Better Rates (Supplementary material)

Grigory Malinovsky[1] **Alibek Sailanbayev**[1] **Peter Richtárik**[1]

[1]AI Initiative, King Abdullah University of Science and Technology, Saudi Arabia

CONTENTS

A Basic Facts **2**

A.1 Elementary Inequalities 2

A.2 Convexity and smoothness 2

A.3 From convergence rate to iteration complexity 2

B Proof of Proposition 1 **4**

C Proof of Lemma 1 **5**

D Analysis of Rand-Shuffle and Rand-Reshuffle **6**

D.1 Proof of Theorems 1 and 2 6

D.2 Proof of Theorem 3 9

D.3 Proof of Theorem 4 10

D.4 Proof of Theorem 5 and Theorem 6 13

E Analysis of Det-Shuffle **15**

E.1 Proof of Theorem 7 15

E.2 Proof of Theorem 8 16

F One More Algorithm: RR-VR **18**

F.1 New Algorithm: RR-VR 18

F.2 Convergence Theory 18

F.3 Proof of Theorem 9 18

F.4 Proof of Theorem 10 20

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Appendix

A BASIC FACTS

A.1 ELEMENTARY INEQUALITIES

Proposition 1. *For all $a, b \in \mathbb{R}^d$ and $t > 0$ the following inequalities hold*

$$\begin{aligned} \langle a, b\rangle &\leq \frac{\|a\|^2}{2t} + \frac{t\|b\|^2}{2}, \\ \|a+b\|^2 &\leq 2\|a\|^2 + 2\|b\|^2, \\ \frac{1}{2}\|a\|^2 - \|b\|^2 &\leq \|a+b\|^2. \end{aligned} \tag{1}$$

A.2 CONVEXITY AND SMOOTHNESS

Proposition 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and let $L \geq 0$. Then the following statements are equivalent:*

- *f is L-smooth,*
- *$2D_f(x, y) \leq L\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$,*
- *$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.*

Proposition 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and let $\mu \geq 0$. Then the following statements are equivalent:*

- *f is μ-strongly convex,*
- *$2D_f(x, y) \geq \mu\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$,*
- *$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.*

Note that the $\mu = 0$ case reduces to convexity.

Proposition 4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $L > 0$. Then the following statements are equivalent:*

- *f is convex and L-smooth*
- *$0 \leq 2D_f(x, y) \leq L\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$,*
- *$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq 2D_f(x, y)$ for all $x, y \in \mathbb{R}^d$,*
- *$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$ for all $x, y \in \mathbb{R}^d$.*

Proposition 5 (Jensen's inequality)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, $x_1, \ldots, x_m \in \mathbb{R}^d$, and $\lambda_1, \ldots, \lambda_m$ be nonnegative real numbers adding up to 1. Then*

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \leq \sum_{i=1}^m \lambda_i f\left(x_i\right).$$

A.3 FROM CONVERGENCE RATE TO ITERATION COMPLEXITY

We implicitly use the following standard result to derive iteration complexity results in our theorems. We include the statement and proof, for completeness.

Lemma 1. *Consider a randomized algorithm producing a sequence of random iterates $\{x_t\}_{t\geq 0}$. Let S_t be some nonnegative function of x_t (example: $S_t = \|x_t - x_*\|^2$). Assume that there exists $q \in (0, 1)$ such that the following inequality holds for all $t \geq 0$:*

$$\mathbb{E}\left[S_t\right] \leq (1 - q)^t S_0. \tag{2}$$

Fix any $\varepsilon > 0$. Then as long as

$$T \geq \frac{1}{q} \ln \left(\frac{1}{\varepsilon} \right),$$

we have

$$\mathbb{E}\left[S_T\right] \leq \varepsilon S_0.$$

Proof. Since $e^q \geq 1 + q$ for all $q \in \mathbb{R}$, we have $e^{-q} \geq 1 - q$ for all $q \in (0, 1)$. Since logarithm is an increasing over $\mathbb{R}_+$, it follows that $-q \geq \ln(1 - q)$ for all $q \in (0, 1)$. Therefore, the inequality

$$-tq \geq t \ln (1 - q)$$

holds for all $t \geq 0$ and all $q \in (0, 1)$. Now if we have $T \geq \frac{1}{q} \ln \left(\frac{1}{\varepsilon} \right)$, which is equivalent to $-T \cdot q \leq \ln(\varepsilon)$, we obtain $T \ln (1 - q) \leq \ln(\varepsilon)$. Taking exponential on both sides, we get

$$0 < (1 - q)^T \leq \varepsilon. \tag{3}$$

Finally, we have

$$\mathbb{E}\left[S_T\right] \overset{(2)}{\leq} (1 - q)^T S_0 \overset{(3)}{\leq} \varepsilon S_0.$$

□

Lemma 2. *Consider a randomized algorithm producing a sequence of random iterates x_t. Let S_t be some nonnegative function of x_t (example: $S_t = \|x_t - x_*\|^2$). Assume that there exists $q \in (0, 1)$ such that the following inequality holds for all $t \geq 0$:*

$$\mathbb{E}\left[S_t\right] \leq (1 - q)^{\beta t} S_0.$$

Fix any $\varepsilon > 0$. Then as long as

$$T \geq \frac{1}{q\beta} \ln \left(\frac{1}{\varepsilon} \right)$$

we have

$$\mathbb{E}\left[S_T\right] \leq \varepsilon.$$

Proof. : Since $e^q \geq 1 + q$ for all $q \in \mathbb{R}$, we have $e^{-q} \geq 1 - q$ for all $q \in (0, 1)$. Since logarithm is an increasing function over $\mathbb{R}_+$, it follows that $-q \geq \ln(1 - q)$ for all $q \in (0, 1)$. Therefore, the inequality $-\beta tq \geq \beta t \ln(1 - q)$ holds for all $t \geq 0$ and all $q \in (0, 1)$. Now, if we have $T \geq \frac{1}{\beta q} \ln \left(\frac{1}{\varepsilon} \right)$, which is equivalent to $-T\beta \cdot q \leq \ln(\varepsilon)$, we obtain $\beta T \ln(1 - q) \leq \ln(\varepsilon)$. Taking exponential on both sides, we get

$$0 < (1 - q)^{\beta T} \leq \varepsilon.$$

Finally, we have

$$\mathbb{E}\left[\Psi_T\right] \leq (1 - q)^{\beta T} \Psi_0 \leq \varepsilon \Psi_0.$$

□

B PROOF OF PROPOSITION 1

Assume that each f_i is μ-strongly convex (resp. convex) and L-smooth. Then the function

$$f^t \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i^t,$$

and

$$f_i^t(x) \coloneqq f_i(x) + \left\langle a_i^t, x \right\rangle, \tag{4}$$

are μ-strongly convex (resp. convex) and L-smooth.

Proof. Let us compute Bregman divergence with respect to the new function $f_i^t(x)$:

$$D_{f_i^t}(x,y) = f_i^t(x) - f_i^t(y) - \langle \nabla f_i^t(y), x - y\rangle.$$

Note that $\nabla f_i^t(y) = \nabla f_i(y) + a_i^t$. Now we have

$$\begin{aligned}
D_{f_i^t}(x,y) &= f_i^t(x) - f_i^t(y) - \langle \nabla f_i^t(y), x - y\rangle \\
&= f_i(x) + \left\langle a_i^t, x\right\rangle - \left(f_i(y) + \left\langle a_i^t, y\right\rangle\right) - \langle \nabla f_i(y) + a_i^t, x - y\rangle \\
&= f_i(x) + \left\langle a_i^t, x\right\rangle - f_i(y) - \left\langle a_i^t, y\right\rangle - \langle \nabla f_i(y), x - y\rangle - \langle a_i^t, x - y\rangle \\
&= f_i(x) + \left\langle a_i^t, x\right\rangle - f_i(y) - \left\langle a_i^t, y\right\rangle - \langle \nabla f_i(y), x - y\rangle - \langle a_i^t, x\rangle + \langle a_i^t, y\rangle \\
&= f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle \\
&= D_{f_i}(x,y).
\end{aligned}$$

Since the Bregman divergence is not changed, the new function $f_i^t(x)$ has the same properties (μ-strong convexity or convexity and L-smoothness) as the initial function $f_i(x)$. □

C PROOF OF LEMMA 1

Proof. We start from definition of $\left(\sigma_*^t\right)^2$ and a_t^i:

$$\left(\sigma_*^t\right)^2 := \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i^t\left(x_*\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i\left(x_*\right) - \nabla f_i\left(y_t\right) + \nabla f\left(y_t\right)\right\|^2.$$

Using the fact that $\nabla f\left(x_*\right) = 0$ we have

$$\left(\sigma_*^t\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i\left(x_*\right) - \nabla f_i\left(y_t\right) + \nabla f\left(y_t\right) - \nabla f\left(x_*\right)\right\|^2.$$

Applying Young's inequality (12) we obtain

$$\left(\sigma_*^t\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(2\left\|\nabla f_i\left(y_t\right) - \nabla f_i\left(x_*\right)\right\|^2 + 2\left\|\nabla f\left(y_t\right) - \nabla f\left(x_*\right)\right\|^2\right).$$

Now we apply Proposition 5 for the squared norms of gradient differences:

$$\left(\sigma_*^t\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n} 4LD_{f_i}\left(y_t, x_*\right) + \frac{1}{n}\sum_{i=1}^{n} 4LD_f\left(y_t, x_*\right).$$

We need to use the fact that $\frac{1}{n}\sum_{i=1}^{n} D_{f_i}\left(y_t, x_*\right) = D_f\left(y_t, x_*\right)$. It is true since $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. So,

$$\left(\sigma_*^t\right)^2 \leq 4LD_f\left(y_t, x_*\right) + 4LD_f\left(y_t, x_*\right) = 8LD_f\left(y_t, x_*\right).$$

Finally, we apply the L-smoothness property from Proposition 2:

$$\left(\sigma_*^t\right)^2 \leq 4L^2\left\|y_t - x_*\right\|^2.$$

□

D ANALYSIS OF RAND-SHUFFLE AND RAND-RESHUFFLE

D.1 PROOF OF THEOREMS 1 AND 2

Proof. We start from Lemma 3 in paper of Mishchenko et al. [2020].

Lemma 3. *Assume that functions $f_1, \ldots, f_n$ are convex and that Assumption 1 is satisfied. If Random Reshuffling or Shuffle-Once is run with a stepsize satisfying $\gamma \leq \frac{1}{\sqrt{2}Ln}$, then*

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] \leq \mathbb{E}\left[\|x_t - x_*\|^2\right] - 2\gamma n \mathbb{E}\left[f\left(x_{t+1}\right) - f(x_*)\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}.$$

The proof of the analogous inequality from Mishchenko et al. [2020] but with condition expectation is identical with very minor changes. We provide such proof below:

We denote by $\mathcal{F}_t$ the σ-algebra generated by the collection of $(\mathcal{X} \times \mathcal{Y})$-valued random variables $(x_0, y_0), \ldots, (x_t, y_t)$, for every $t \geq 0$. In this work, we consider unbiased random estimates: for every $t \geq 0$. If the method does not depend on y_τ we can still use such notation because of the independence property for conditional expectations. We denote by $\mathcal{F}_t$ the σ-algebra generated by the collection of $(\mathcal{X} \times \mathcal{Y})$-valued random variables $(x_0, y_0), \ldots, (x_t, y_t)$, for every $t \geq 0$. In this work, we consider unbiased random estimates: for every $t \geq 0$. We define the forward per-epoch deviation over the t-th epoch $\mathcal{V}_i$ as

$$\mathcal{V}_t = \sum_{i=0}^{n-1} \left\|x_t^i - x_{t+1}\right\|^2$$

Lemma 2. Consider the iterates of Random Reshuffling or Shuffle-Once. If the functions $f_1, \ldots, f_n$ are convex and Assumption 1 is satisfied, then

$$\mathbb{E}\left[\mathcal{V}_t \mid \mathcal{F}_t\right] \leq 4\gamma^2 n^2 L \sum_{i=0}^{n-1} \mathbb{E}\left[D_{f_i}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] + \frac{1}{2}\gamma^2 n^2 \sigma_*^2$$

where $\mathcal{V}_i$ is defined above, and σ_*^2 is the variance at the optimum given by $\sigma_*^2 = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i\left(x_*\right)\right\|^2$ We will follow the steps from Mishchenko et al. [2020].

Proof. For any fixed $k \in 0, \ldots, n-1$, by definition of x_t^k and x_{t+1} (According to Algorithm 1 or 2 in Mishchenko et al. [2020]) we get the decomposition

$$x_t^k - x_{t+1} = \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_t^i\right) = \gamma \sum_{i=k}^{n-1} \left(\nabla f_{\pi_i}\left(x_t^i\right) - \nabla f_{\pi_i}\left(x_*\right)\right) + \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)$$

Applying Young's inequality to the sums above yields

$$\left\|x_t^k - x_{t+1}\right\|^2 \leq 2\gamma^2 \left\|\sum_{i=k}^{n-1} \left(\nabla f_{\pi_i}\left(x_t^i\right) - \nabla f_{\pi_i}\left(x_*\right)\right)\right\|^2 + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2$$

Using Jensen's inequality we have

$$\left\|x_t^k - x_{t+1}\right\|^2 \leq 2\gamma^2 n \sum_{i=k}^{n-1} \left\|\nabla f_{\pi_i}\left(x_t^i\right) - \nabla f_{\pi_i}\left(x_*\right)\right\|^2 + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2$$

Using L-smoothness property from Proposition 3 we have

$$\left\|x_t^k - x_{t+1}\right\|^2 \leq 4\gamma^2 L n \sum_{i=k}^{n-1} D_{f_{\pi_i}}\left(x_*, x_t^i\right) + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2$$

Further, we have

$$\left\|x_t^k - x_{t+1}\right\|^2 \leq 4\gamma^2 Ln \sum_{i=0}^{n-1} D_{f_{\pi_i}}\left(x_*, x_t^i\right) + 2\gamma^2 \left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2$$

Summing up and taking conditional expectations leads to

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\left\|x_t^k - x_{t+1}\right\|^2 \mid \mathcal{F}_t\right] \leq 4\gamma^2 Ln^2 \sum_{i=0}^{n-1} \mathbb{E}\left[D_{f_i}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] + 2\gamma^2 \sum_{k=0}^{n-1} \mathbb{E}\left[\left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2 \mid \mathcal{F}_t\right]$$

Since $\sum_{k=0}^{n-1} \mathbb{E}\left[\left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2 \mid \mathcal{F}_t\right]$ does not depend on $\mathcal{F}_t$ but only on permutations we have

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\left\|x_t^k - x_{t+1}\right\|^2 \mid \mathcal{F}_t\right] \leq 4\gamma^2 Ln^2 \sum_{i=0}^{n-1} \mathbb{E}\left[D_{f_{\pi_i}}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] + 2\gamma^2 \sum_{k=0}^{n-1} \mathbb{E}\left[\left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2\right]$$

We now bound the second term in the right-hand side. First, using Lemma 1 from Mishchenko et al. [2020], we get $\mathbb{E}\left[\left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2\right] = (n-k)^2 \mathbb{E}\left[\left\|\frac{1}{n-k}\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2\right] = (n-k)^2 \frac{k}{(n-k)(n-1)}\sigma_*^2 = \frac{k(n-k)}{n-1}\sigma_*^2$. Next, by summing this for k from 0 to $n-1$, we obtain

$$\sum_{k=0}^{n-1} \mathbb{E}\left[\left\|\sum_{i=k}^{n-1} \nabla f_{\pi_i}\left(x_*\right)\right\|^2\right] = \sum_{k=0}^{n-1} \frac{k(n-k)}{n-1}\sigma_*^2 = \frac{1}{6}n(n+1)\sigma_*^2 \leq \frac{n^2\sigma_*^2}{4}$$

where in the last step we also used $n \geq 2$. The result follows. □

Let us provide analogue for Lemma 3 from Mishchenko et al. [2020].

Lemma 3*. Assume that functions $f_1, \ldots, f_n$ are convex and that Assumption 1 is satisfied. If Random Reshuffling (Algorithm 1) or Shuffle-Once (Algorithm 2) is run with a stepsize satisfying $\gamma \leq \frac{1}{\sqrt{2}Ln}$, then

$$\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] \leq \left\|x_t - x_*\right\|^2 - 2\gamma n \mathbb{E}\left[f\left(x_{t+1}\right) - f_* \mid \mathcal{F}_t\right] + \frac{\gamma^3 Ln^2\sigma_*^2}{2}$$

Proof. Define the sum of gradients used in the t-th epoch as $g_t = \sum_{i=0}^{n-1} \nabla f_{\pi_i}\left(x_t^i\right)$. We will use g_t to relate the iterates x_t and x_{t+1}. By definition of x_{t+1}, we can write

$$x_{t+1} = x_t^n = x_t^{n-1} - \gamma \nabla f_{\pi_{n-1}}\left(x_t^{n-1}\right) = \ldots = x_t^0 - \gamma \sum_{i=0}^{n-1} \nabla f_{\pi_i}\left(x_t^i\right)$$

Further, since $x_t^0 = x_t$, we see that $x_{t+1} = x_t - \gamma g_t$, which leads to

$$\left\|x_t - x_*\right\|^2 = \left\|x_{t+1} + \gamma g_t - x_*\right\|^2 = \left\|x_{t+1} - x_*\right\|^2 + 2\gamma\left\langle g_t, x_{t+1} - x_*\right\rangle + \gamma^2\left\|g_t\right\|^2$$

Since $\gamma^2\left\|g_t\right\|^2 \geq 0$ we have

$$\left\|x_t - x_*\right\|^2 \geq \left\|x_{t+1} - x_*\right\|^2 + 2\gamma\left\langle g_t, x_{t+1} - x_*\right\rangle = \left\|x_{t+1} - x_*\right\|^2 + 2\gamma \sum_{i=0}^{n-1} \left\langle \nabla f_{\pi_i}\left(x_t^i\right), x_{t+1} - x_*\right\rangle$$

Observe that for any i, we have the following decomposition

$$\left\langle \nabla f_{\pi_i}\left(x_t^i\right), x_{t+1} - x_*\right\rangle = \left(f_{\pi_i}\left(x_{t+1}\right) - f_{\pi_i}\left(x_*\right)\right) + D_{f_{\pi_i}}\left(x_*, x_t^i\right) - D_{f_{\pi_i}}\left(x_{t+1}, x_t^i\right)$$

Summing the first quantity over i from 0 to $n-1$ gives

$$\sum_{i=0}^{n-1}\left(f_{\pi_i}\left(x_{t+1}\right) - f_{\pi_i}\left(x_*\right)\right) = n\left(f\left(x_{t+1}\right) - f_*\right)$$

Now, we can bound the third term in the decomposition (33) using L-smoothness as follows:

$$D_{f_{\pi_i}}\left(x_{t+1}, x_t^i\right) \leq \frac{L}{2}\left\|x_{t+1} - x_t^i\right\|^2$$

By summing the right-hand side over i from 0 to $n-1$ we get the forward deviation over an epoch $\mathcal{V}_t$, which we bound by analogue of Lemma 2 to get

$$\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}\left(x_{t+1}, x_t^i\right) \mid \mathcal{F}_t\right] \leq \frac{L}{2}\mathbb{E}\left[\mathcal{V}_t \mid \mathcal{F}_t\right] \leq 2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] + \frac{\gamma^2 L n^2 \sigma_*^2}{4}$$

Therefore, we can lower-bound the sum of the second and the third term as

$$\begin{aligned}\sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}\left(x_*, x_t^i\right) - D_{f_{\pi_i}}\left(x_{t+1}, x_t^i\right) \mid \mathcal{F}_t\right] &\geq \sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] \\ - 2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1}\mathbb{E}\left[D_{f_{\pi_i}}\left(x_*, x_t^i\right) \mid \mathcal{F}_t\right] &- \frac{\gamma^2 L n^2 \sigma_*^2}{4}.\end{aligned}$$

□

Proof. We start from analogue of Lemma 3 in paper of Mishchenko et al. [2020], which we proved above.

$$\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] \leq \left\|x_t - x_*\right\|^2 - 2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right) \mid \mathcal{F}_t\right] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}$$

Now we can apply this inequality to the reformulated problem (2). Using strong convexity, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \left\|x_t - x_*\right\|^2 - 2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right) \mid \mathcal{F}_t\right] + \frac{\gamma^3 L n^2 \left(\sigma_*^t\right)^2}{2} \\ \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \left\|x_t - x_*\right\|^2 - \gamma n\mu\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] + \frac{\gamma^3 L^2 \left(\sigma_*^t\right)^2}{2}\end{aligned}$$

Since we update $y_t = x_t$ after each epoch, this leads to

$$\begin{aligned}\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \frac{1}{1+\gamma\mu n}\left(\left\|x_t - x_*\right\|^2 + \frac{\gamma^3 L n^2 \left(\sigma_*^t\right)^2}{2}\right) \\ \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \frac{1}{1+\gamma\mu n}\left(\left\|x_t - x_*\right\|^2 + \frac{\gamma^3 L n^2 \cdot 4L^2\left\|y_t - x_*\right\|^2}{2}\right) \\ \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \frac{1}{1+\gamma\mu n}\left(\left\|x_t - x_*\right\|^2 + 2\gamma^3 n^2 L^3\left\|x_t - x_*\right\|^2\right) \\ \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] &\leq \frac{1}{1+\gamma\mu n}\left(1 + 2\gamma^3 n^2 L^3\right)\left\|x_t - x_*\right\|^2\end{aligned}$$

We can use the tower property of conditional expectation to obtain

$$\mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right] \leq \frac{1 + 2\gamma^3 L^3 n^2}{1+\gamma\mu n}\mathbb{E}\left[\left\|x_t - x_*\right\|^2\right]$$

Since $\gamma \leq \frac{1}{2\sqrt{2}Ln}\sqrt{\frac{\mu}{L}}$, $n \geq 1$ and $\mu \leq L$ we have

$$\frac{1}{4n} + \frac{1}{4\sqrt{2}}\frac{\mu}{L}\sqrt{\frac{\mu}{L}} \leq \frac{1}{2}$$

From this inequality we obtain

$$\mathbb{E}\left[\|x_{t+1}-x_*\|^2\right] \leq \frac{1+2\gamma^3L^3n^2}{1+\gamma\mu n}\mathbb{E}\left[\|x_t-x_*\|^2\right]$$

Since $\gamma \leq \dfrac{1}{2\sqrt{2}Ln}\sqrt{\dfrac{\mu}{L}}, n \geq 1$ and $\mu \leq L$ we have

$$\frac{1}{4n}+\frac{1}{4\sqrt{2}}\frac{\mu}{L}\sqrt{\frac{\mu}{L}} \leq \frac{1}{2}$$

From this inequality we obtain $2 \cdot \frac{1}{8L^2n^2} \cdot \frac{\mu}{L}L^3n + \frac{1}{2\sqrt{2}Ln}\sqrt{\frac{\mu}{L}} \cdot \frac{n\mu^2}{2} \leq \frac{\mu}{2}$

$$\frac{1}{4n}\mu+\frac{1}{4\sqrt{2}}\frac{\mu}{L}\sqrt{\frac{\mu}{L}}\mu \leq \frac{\mu}{2}$$

We continue to derive inequalities:

$$\begin{aligned}
2\gamma^2L^3n+\frac{\gamma n\mu^2}{2} &\leq \frac{\mu}{2}\\
2\gamma^2L^3n &\leq \frac{\mu}{2}-\frac{\gamma n\mu^2}{2}\\
2\gamma^2L^3n^2 &\leq \frac{n\mu}{2}-\frac{\gamma n^2\mu^2}{2}\\
1+2\gamma^3L^3n^2 &\leq 1+\frac{\gamma n\mu}{2}-\frac{\gamma^2n^2\mu^2}{2}
\end{aligned}$$

Finally, we obtain

$$\frac{1+2\gamma^3L^3n^2}{1+\gamma\mu n} \leq 1-\frac{\gamma n\mu}{2}$$

Plugging this inequality into $\mathbb{E}\left[\|x_{t+1}-x_*\|^2\right] \leq \frac{1+2\gamma^3L^3n^2}{1+\gamma\mu n}\mathbb{E}\left[\|x_t-x_*\|^2\right]$, we unroll the recursion and obtain the final result:

$$\mathbb{E}\left[\|x_T-x_*\|^2\right] \leq \left(1-\frac{\gamma n\mu}{2}\right)^T\|x_0-x_*\|^2$$

□

D.2 PROOF OF THEOREM 3

We start from conditional analogue of Theorem 1 in [Mishchenko et al., 2020] (similarly to Section D.1), which states that

$$\mathbb{E}\left[\|x_{t+1}-x_*\|^2 \mid \mathcal{F}_t\right] \leq (1-\gamma\mu)^n\|x_t-x_*\|^2 + 2\gamma^2\sigma^2_{\text{Shuffle}}\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right).$$

Using Proposition 1 from [Mishchenko et al., 2020], which says that

$$\frac{\gamma\mu n}{8}\sigma_*^2 \leq \sigma^2_{\text{Shuffle}} \leq \frac{\gamma Ln}{4}\sigma_*^2,$$

we get

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1}-x_*\|^2 \mid \mathcal{F}_t\right] &\leq (1-\gamma\mu)^n\|x_t-x_*\|^2 + \frac{\gamma^3Ln}{2}\sigma_*^2\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right)\\
&\leq (1-\gamma\mu)^n\|x_t-x_*\|^2 + \frac{\gamma^2Ln}{2\mu}\sigma_*^2.
\end{aligned}$$

Now we can apply Lemma 1 and Reformulation. Using $y_t = x_t$ we have the following inequality:

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t\right] \leq (1-\gamma\mu)^n \|x_t - x_*\|^2 + \frac{2\gamma^2 L^3 n}{\mu}\|x_t - x_*\|^2$$
$$\leq \left((1-\gamma\mu)^n + \frac{2\gamma^2 L^3 n}{\mu}\right)\|x_t - x_*\|^2.$$

Applying the tower property, we get

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] \leq \left((1-\gamma\mu)^n + \frac{2\gamma^2 L^3 n}{\mu}\right)\mathbb{E}\left[\|x_t - x_*\|^2\right],$$

and after unrolling this recursion, we get

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \left((1-\gamma\mu)^n + \frac{2\gamma^2 L^3 n}{\mu}\right)^T \mathbb{E}\left[\|x_0 - x_*\|^2\right]$$
$$\leq \left((1-\gamma\mu)^n + \frac{\delta^2}{L^2}\frac{\mu}{2nL}\frac{2L^3 n}{\mu}\right)^T \mathbb{E}\left[\|x_0 - x_*\|^2\right]$$
$$\leq \left((1-\gamma\mu)^n + \delta^2\right)^T \mathbb{E}\left[\|x_0 - x_*\|^2\right],$$

where we used the stepsize restriction $\gamma \leq \frac{\delta}{L}\sqrt{\frac{\mu}{2nL}}$. In order for this to lead to convergence, we need to assume that $(1-\gamma\mu)^n + \delta^2 < 1$. This is satisfied, for example, if n is large enough. In particular, this holds when

$$n > \log\left(\frac{1}{1-\delta^2}\right)\cdot\left(\log\left(\frac{1}{1-\gamma\mu}\right)\right)^{-1}.$$

Finally, using the additional assumption $\delta^2 \leq (1-\gamma\mu)^{\frac{n}{2}}\left(1-(1-\gamma\mu)^{\frac{n}{2}}\right)$, we get

$$\delta^2 + (1-\gamma\mu)^n \leq (1-\gamma\mu)^{\frac{n}{2}}.$$

Now we can apply Theorem 3 and get

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1-\gamma\mu)^{\frac{nT}{2}}\|x_0 - x_*\|^2.$$

Finally, we apply Lemma 1 with $\gamma = \frac{\delta}{L}\sqrt{\frac{\mu}{2nL}}$ and get iteration complexity $T = \mathcal{O}\left(\kappa\sqrt{\frac{\kappa}{n}}\log\left(\frac{1}{\varepsilon}\right)\right)$.

D.3 PROOF OF THEOREM 4

Suppose the functions $f_1, f_2, \ldots, f_n$ are convex and Assumption 1 holds. Then for Rand-Reshuffle or Rand-Shuffle with stepsize $\gamma \leq \frac{1}{\sqrt{2}Ln}$, the average iterate $\hat{x}_T := \frac{1}{T}\sum_{t=1}^T x_t$ satisfies

$$\mathbb{E}\left[f\left(\hat{x}_T\right) - f\left(x_*\right)\right] \leq \frac{3\left\|x_0 - x_*\right\|^2}{2\gamma nT}.$$

Proof. We start with conditional analogue of Lemma 3 from Mishchenko et al. [2020] (similarly to Section D.1), which says that

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid \mathcal{F}_t\right] \leq \|x_t - x_*\|^2 - 2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right) \mid \mathcal{F}_t\right] + \frac{\gamma^3 Ln^2\sigma_*^2}{2}.$$

Apply this inequality to the reformulated problem, we get

$$2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right) \mid \mathcal{F}_t\right] \leq \|x_t - x_*\|^2 - \mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid \mathcal{F}_t\right] + \frac{\gamma^3 Ln^2\left(\sigma_*^t\right)^2}{2}. \tag{5}$$

Using Lemma 1 and the fact that $y_t = x_t$ and $f = f^t$, we get

$$\left(\sigma_*^t\right)^2 \leq 8LD_{f^t}(x_t, x_*) = 8LD_f(x_t, x_*) = 8L(f(x_t) - f(x_*)), \tag{6}$$

where the last identity follows from Proposition 1.

Plugging (6) into (5), we obtain

$$2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right) \mid \mathcal{F}_t\right] \leq \left\|x_t - x_*\right\|^2 - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2 \mid \mathcal{F}_t\right] + \frac{\gamma^3 Ln^2}{2} \cdot 8L(f(x_t) - f(x_*)),$$

which after using the tower property turns into

$$2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] \leq \mathbb{E}\left[\left\|x_t - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right] + 4\gamma^3 L^2 n^2 \mathbb{E}\left[f(x_t) - f(x_*)\right].$$

Now we subtract from both sides:

$$\begin{aligned}
2\gamma n\mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] - 4\gamma^3 L^2 n^2 \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] &\leq \mathbb{E}\left[\left\|x_t - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right] \\
&\quad + 4\gamma^3 L^2 n^2 \mathbb{E}\left[f(x_t) - f(x_*)\right] \\
&\quad - 4\gamma^3 L^2 n^2 \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] \\
\left(2\gamma n - 4\gamma^3 L^2 n^2\right) \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] &\leq \mathbb{E}\left[\left\|x_t - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right] \\
&\quad + 4\gamma^3 L^2 n^2 \left(\mathbb{E}\left[f(x_t) - f(x_*)\right] - \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right]\right) \\
2\gamma n\left(1 - 2\gamma^2 L^2 n\right) \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] &\leq \mathbb{E}\left[\left\|x_t - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right] \\
&\quad + 4\gamma^3 L^2 n^2 \left(\mathbb{E}\left[f(x_t) - f(x_*)\right] - \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right]\right).
\end{aligned}$$

Summing these inequalities for $t = 0, 1, \ldots, T-1$ gives

$$\begin{aligned}
2\gamma n\left(1 - 2\gamma^2 L^2 n\right) \sum_{t=0}^{T-1} \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] &\leq \sum_{t=0}^{T-1} \left(\mathbb{E}\left[\left\|x_t - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_{t+1} - x_*\right\|^2\right]\right) \\
&\quad + 4\gamma^3 L^2 n^2 \sum_{t=0}^{T-1} \left(\mathbb{E}\left[f(x_t) - f(x_*)\right] - \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right]\right) \\
&= \mathbb{E}\left[\left\|x_0 - x_*\right\|^2\right] - \mathbb{E}\left[\left\|x_T - x_*\right\|^2\right] \\
&\quad + 4\gamma^3 L^2 n^2 \mathbb{E}\left[f\left(x_0\right) - f\left(x_*\right)\right] - 4\gamma^3 L^2 n^2 \mathbb{E}\left[f\left(x_T\right) - f\left(x_*\right)\right] \\
&\leq \mathbb{E}\left[\left\|x_0 - x_*\right\|^2\right] + 4\gamma^3 L^2 n^2 \mathbb{E}\left[f\left(x_0\right) - f\left(x_*\right)\right] \\
&\leq \mathbb{E}\left[\left\|x_0 - x_*\right\|^2\right] + 2\gamma^3 L^3 n^2 \mathbb{E}\left[\left\|x_0 - x_*\right\|^2\right] \\
&= (1 + 2\gamma^3 L^3 n^2)\mathbb{E}\left[\left\|x_0 - x_*\right\|^2\right],
\end{aligned}$$

and dividing both sides by $2\gamma n\left(1 - 2\gamma^2 L^2 n\right) T$, we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[f\left(x_{t+1}\right) - f\left(x_*\right)\right] \leq \frac{1 + 2\gamma^3 L^3 n^2}{1 - 2\gamma^2 L^2 n} \frac{\left\|x_0 - x_*\right\|^2}{2\gamma nT}.$$

Using the convexity of f, the average iterate $\hat{x}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} x_t$ satisfies

$$\mathbb{E}\left[f\left(\hat{x}_T\right) - f\left(x_*\right)\right] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[f\left(x_t\right) - f\left(x_*\right)\right] \leq \frac{1 + 2\gamma^3 L^3 n^2}{1 - 2\gamma^2 L^2 n} \frac{\left\|x_0 - x_*\right\|^2}{2\gamma nT}.$$

Let us show that

$$\frac{1 + 2\gamma^3 L^3 n^2}{1 - 2\gamma^2 L^2 n} \leq 3.$$

Applying $\gamma \leq \frac{1}{\sqrt{2}Ln}$ we have

$$\frac{1 + 2\frac{1}{2\sqrt{2}L^3n^3}L^3n^2}{1 - 2\frac{1}{2L^2n^2}L^2n} = \frac{1 + \frac{1}{\sqrt{2}n}}{1 - \frac{1}{n}} \leq 3.$$

This leads to $4n > 6 + \sqrt{2}$ and since $n \in \mathbb{N} : n > 1$, this inequality holds. Finally, we have

$$\mathbb{E}\left[f\left(\hat{x}_T\right) - f\left(x_*\right)\right] \leq \frac{3\left\|x_0 - x_*\right\|^2}{2\gamma nT}.$$

□

D.4 PROOF OF THEOREM 5 AND THEOREM 6

We provide analysis for non-convex settings.

Let us remind you our reformulation:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n}\left(f_i(x) + \left\langle a_t^i, x\right\rangle\right) := \frac{1}{n}\sum_{i=1}^{n} f_i^t(x),$$

where $f_i^t(x) := f_i(x) + \left\langle a_t^i, x\right\rangle$ and $\sum_{i=1}^n a_t^i = 0$. Note that

$$\nabla f_i^t(x) = \nabla f_i(x) + a_t^I.$$

In particular, we choose

$$a_t^i := -\nabla f_{\pi_i}(y_t) + \nabla f(y_t).$$

Finally, we have

$$\nabla f_{\pi_i}^t(x) = \nabla f_{\pi_i}(x) - \nabla f_{\pi_i}(y_t) + \nabla f(y_t).$$

Now we need to establish an analogue of Lemma 1 for gradient variance. Let us define

$$\sigma^2(x_t) = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x_t) - \nabla f(x_t)\right\|^2.$$

Lemma 4. *If we apply the linear perturbation reformulation, then the gradient variance of the reformulated problem $\left(\sigma_t^2\right)$ is equal to zero.*

Proof.

$$\sigma_t^2(x_t) = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i^t(x_t) - \nabla f(x_t)\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x_t) - \nabla f_i(y_t) + \nabla f(y_t) - \nabla f(x_t)\right\|^2$$

In Algorithm **??** (Rand-Reshuffle) we set $x_t = y_t$, and hence we have

$$\sigma_t^2(x_t) = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x_t) - \nabla f_i(x_t) + \nabla f(x_t) - \nabla f(x_t)\right\|^2 = 0.$$

□

Suppose that Assumption 1 holds. Then for Algorithm Rand-Reshuffle run for T epochs with a stepsize $\gamma \leq \frac{1}{2Ln}$ we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x_t)\right\|^2\right] \leq \frac{4(f(x_0)-f_*)}{\gamma n T}.$$

Choose $\gamma = \frac{1}{2nL}$. Then the mean of gradient norms satisfies $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x_t)\right\|^2\right] \leq \varepsilon^2$ provided the number of iterations satisfies $T = \mathcal{O}\left(\frac{8\delta_0 L}{\varepsilon^2}\right)$.

Suppose that Assumption **??** holds and f satisfies the Polyak-Łojasiewicz inequality with $\mu > 0$, i.e., $\left\|\nabla f(x)\right\|^2 \geq 2\mu(f(x) - f_*)$ for any $x \in \mathbb{R}^d$. Then for Algorithm Rand-Reshuffle run for T epochs with a stepsize $\gamma \leq \frac{1}{2Ln}$ we have

$$\mathbb{E}\left[f(x_T) - f_*\right] \leq \left(1 - \frac{\gamma\mu n}{2}\right)^T \left(f(x_0) - f_*\right),$$

then the relative error satisfies $\frac{\mathbb{E}[f(x_T)-f_*]}{f(x_0)-f_*} \leq \varepsilon$ provided the number of iterations satisfies $T = \mathcal{O}(\kappa \log \frac{1}{\varepsilon})$.

Proof. We start from conditional analogues of Lemmas 4 and 5 from Mishchenko et al. [2020] (similarly to Section D.1)

$$\mathbb{E}\left[f(x_{t+1})|\mathcal{F}_t\right] \leq f(x_t) - \frac{\gamma n}{2}\left\|\nabla f(x_t)\right\|^2 + \frac{\gamma L^2}{2}\left(\gamma^2 n^3\left\|\nabla f(x_t)\right\|^2 + \gamma^2 n^2 \sigma^2(x_t)\right)$$

This lemma works for the reformulated problem. Since we do not change initial function f(x) the gradient $\nabla f(x_t)$ remains the same. The only thing that changes is the variance of the gradient. According to the lemma proved above, this variance is equal to zero. Now we have the following inequality:

$$\begin{aligned}\mathbb{E}\left[f(x_{t+1})|\mathcal{F}_t\right] &\leq f(x_t) - \frac{\gamma n}{2}\left\|\nabla f(x_t)\right\|^2 + \frac{\gamma L^2}{2}\gamma^2 n^3\left\|\nabla f(x_t)\right\|^2 \\ &\leq f\left(x_t\right) - \frac{\gamma n}{2}\left(1 - \gamma^2 L^2 n^2\right)\left\|\nabla f\left(x_t\right)\right\|^2\end{aligned}$$

Let $\delta_t = f(x_t) - f_*$. Adding $-f_*$ to both sides,

$$\mathbb{E}\left[\delta_{t+1}|\mathcal{F}_t\right] \leq \delta_t - \frac{\gamma n}{2}\left(1 - \gamma^2 L^2 n^2\right)\left\|\nabla f\left(x_t\right)\right\|^2$$

Taking unconditional expectations and using that $\gamma \leq \frac{1}{2Ln}$ we have $1 - \gamma^2 L^2 n^2 \geq \frac{1}{2}$, we get

$$\mathbb{E}\left[\delta_{t+1}\right] \leq \mathbb{E}\left[\delta_t\right] - \frac{\gamma n}{4}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right].$$

It leads to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leq \frac{4}{\gamma n}\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\delta_{t+1}\right] - \mathbb{E}\left[\delta_t\right]\right) \leq \frac{4\delta_0}{\gamma n T}$$

If we have PL condition, then we start from

$$\mathbb{E}\left[\delta_{t+1}\right] \leq \mathbb{E}\left[\delta_t\right] - \frac{\gamma n}{4}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right].$$

Applying $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_*)$ leads to

$$\mathbb{E}\left[\delta_{t+1}\right] \leq \mathbb{E}\left[\delta_t\right] - \frac{\gamma\mu n}{2}\mathbb{E}\left[f\left(x_t\right) - f_*\right].$$

Unrolling this recursion, we get

$$\mathbb{E}\left[\delta_T\right] \leq \left(1 - \frac{\gamma\mu n}{2}\right)^T \delta_0.$$

Suppose that Assumption 1 holds. Choose the stepsize γ as $\frac{1}{2nL}$. Then the mean of gradient norms satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f\left(x_t\right)\right\|^2\right] \leq \varepsilon^2$$

provided the number of iterations satisfies

$$T \geq \frac{8\delta_0 L}{\varepsilon^2}.$$

If f satisfies the Polyak-Łojasiewicz inequality, then the relative error satisfies

$$\frac{\mathbb{E}\left[f\left(x_T\right) - f_*\right]}{\left(f\left(x_0\right) - f_*\right)} \leq \varepsilon$$

provided the number of iterations satisfies

$$T = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right).$$

□

E ANALYSIS OF DET-SHUFFLE

E.1 PROOF OF THEOREM 7

We start from Lemma 8 in Mishchenko et al. [2020]

$$\|x_{t+1} - x_*\|^2 \leq \|x_t - x_*\|^2 - 2\gamma n \left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) + \gamma^3 L n^3 \sigma_*^2. \tag{7}$$

Now we can apply to the reformulated problem (**??**). Using strong convexity we obtain

$$\begin{aligned}\|x_{t+1} - x_*\|^2 &\leq \|x_t - x_*\|^2 - 2\gamma n \left(f\left(x_{t+1}\right) - f(x_*)\right) + \gamma^3 L n^2 \left(\sigma_*^t\right)^2 \\ &\leq \|x_t - x_*\|^2 - \gamma n \mu \left(\|x_{t+1} - x_*\|^2\right) + \gamma^3 L n^3 \left(\sigma_*^t\right)^2.\end{aligned}$$

Since we update $y_t = x_t$ after each epoch, this leads to

$$\begin{aligned}\|x_{t+1} - x_*\|^2 &\leq \frac{1}{1+\gamma\mu n}\left(\|x_t - x_*\|^2 + \gamma^3 L n^3 \left(\sigma_*^t\right)^2\right) \\ &\leq \frac{1}{1+\gamma\mu n}\left(\|x_t - x_*\|^2 + \gamma^3 L n^3 \cdot 4L^2 \|y_t - x_*\|^2\right) \\ &= \frac{1}{1+\gamma\mu n}\left(\|x_t - x_*\|^2 + 4\gamma^3 n^3 L^3 \|x_t - x_*\|^2\right) \\ &= \frac{1}{1+\gamma\mu n}\left(1 + 4\gamma^3 n^3 L^3\right)\|x_t - x_*\|^2.\end{aligned}$$

We obtain

$$\|x_{t+1} - x_*\|^2 \leq \frac{1 + 4\gamma^3 L^3 n^3}{1+\gamma\mu n}\|x_t - x_*\|^2.$$

Since we have $\mu \leq L$ we obtain

$$\begin{aligned}\frac{1}{8} + \frac{1}{8}\frac{\mu}{L}\sqrt{\frac{\mu}{L}} &\leq \frac{1}{2} \\ \frac{1}{8}\mu + \frac{1}{8}\frac{\mu}{L}\sqrt{\frac{\mu}{L}}\mu &\leq \frac{\mu}{2} \\ 2 \cdot \frac{1}{16L^2n^2} \cdot \frac{\mu}{L} L^3 n^2 + \frac{1}{4Ln}\sqrt{\frac{\mu}{L}} \cdot \frac{n\mu^2}{2} &\leq \frac{\mu}{2}.\end{aligned}$$

Now as $\gamma \leq \frac{1}{4Ln}\sqrt{\frac{\mu}{L}}$, we have

$$\begin{aligned}4\gamma^2 L^3 n^2 + \frac{\gamma n \mu^2}{2} &\leq \frac{\mu}{2} \\ 4\gamma^2 L^3 n^2 &\leq \frac{\mu}{2} - \frac{\gamma n \mu^2}{2} \\ 4\gamma^2 L^3 n^3 &\leq \frac{n\mu}{2} - \frac{\gamma n^2 \mu^2}{2} \\ 1 + 4\gamma^3 L^3 n^3 &\leq 1 + \frac{\gamma n \mu}{2} - \frac{\gamma^2 n^2 \mu^2}{2}.\end{aligned}$$

Let us simplify it:

$$\frac{1 + 4\gamma^3 L^3 n^3}{1+\gamma\mu n} \leq 1 - \frac{\gamma n \mu}{2}.$$

We can unroll the recursion and obtain

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq \left(1 - \frac{\gamma n \mu}{2}\right)^T \|x_0 - x_*\|^2.$$

E.2 PROOF OF THEOREM 8

Suppose the functions $f_1, f_2, \ldots, f_n$ are convex and Assumption 1 holds. Then for Algorithm 1 (Det-Shuffle) with a stepsize $\gamma \leq \frac{1}{2\sqrt{2}Ln}$, the average iterate $\hat{x}_T := \frac{1}{T}\sum_{j=1}^{T} x_j$ satisfies

$$f\left(\hat{x}_T\right) - f\left(x_*\right) \leq \frac{2\left\|x_0 - x_*\right\|^2}{\gamma n T}.$$

We start with Lemma 8 from Mishchenko et al. [2020]:

$$\begin{aligned}\left\|x_{t+1} - x_*\right\|^2 &\leq \left\|x_t - x_*\right\|^2 - 2\gamma n\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) + \gamma^3 L n^3 \sigma_*^2\\ 2\gamma n\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) &\leq \left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2 + \gamma^3 L n^3 \sigma_*^2.\end{aligned}$$

Using Lemma **??** and considering $y_t = x_t$, we have

$$\left(\sigma_*^t\right)^2 \leq 8LD_{f^t}(x_t, x_*).$$

Applying Proposition **??** we get

$$\left(\sigma_*^t\right)^2 \leq 8LD_f(x_t, x_*) = 8L(f(x_t) - f(x_*)).$$

Next, we utilize the inner product reformulation and get

$$2\gamma n\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) \leq \left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2 + \gamma^3 L n^3 \cdot 8L(f(x_t) - f(x_*)).$$

Using tower property we have

$$2\gamma n\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) \leq \left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2 + 8\gamma^3 L^2 n^3\left(\left(f(x_t) - f(x_*)\right)\right).$$

Now we subtract from both sides:

$$\begin{aligned}2\gamma n\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) - 8\gamma^3 L^2 n^3\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) &\leq \left(\left\|x_t - x_*\right\|^2\right) - \left\|x_{t+1} - x_*\right\|^2\\ &\quad + 8\gamma^3 L^2 n^3\left(\left(f(x_t) - f(x_*)\right)\right)\\ &\quad - 8\gamma^3 L^2 n^3\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right)\\ \left(2\gamma n - 8\gamma^3 L^2 n^3\right)\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) &\leq \left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2\\ &\quad + 8\gamma^3 L^2 n^3\left(\left(f(x_t) - f(x_*)\right) - \left(f\left(x_{t+1}\right) - f\left(x_*\right)\right)\right)\\ 2\gamma n\left(1 - 4\gamma^2 L^2 n^2\right)\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) &\leq \left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2\\ +8\gamma^3 L^2 n^3\left(\left(f(x_t) - f(x_*)\right) - \left(f\left(x_{t+1}\right) - f\left(x_*\right)\right)\right).\end{aligned}$$

Summing these inequalities for $t = 0, 1, \ldots, T-1$ gives

$$\begin{aligned}2\gamma n\left(1 - 4\gamma^2 L^2 n^2\right)\sum_{t=0}^{T-1}\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) &\leq \sum_{t=0}^{T-1}\left(\left\|x_t - x_*\right\|^2 - \left\|x_{t+1} - x_*\right\|^2\right)\\ &\quad + 8\gamma^3 L^2 n^3 \sum_{t=0}^{T-1}\left(\left(f(x_t) - f(x_*)\right) - \left(f\left(x_{t+1}\right) - f\left(x_*\right)\right)\right)\\ &= \left\|x_0 - x_*\right\|^2 - \left\|x_T - x_*\right\|^2\\ &\quad + 8\gamma^3 L^2 n^3\left(f\left(x_0\right) - f\left(x_*\right)\right) - 8\gamma^3 L^2 n^3\left(f\left(x_T\right) - f\left(x_*\right)\right)\\ &\leq \left\|x_0 - x_*\right\|^2 + 8\gamma^3 L^2 n^3\left(f\left(x_0\right) - f\left(x_*\right)\right)\\ &\leq \left\|x_0 - x_*\right\|^2 + 4\gamma^3 L^3 n^3\left\|x_0 - x_*\right\|^2\\ &= \left(1 + 4\gamma^3 L^3 n^3\right)\left\|x_0 - x_*\right\|^2,\end{aligned}$$

and dividing both sides by $2\gamma n\left(1 - 4\gamma^2 L^2 n^2\right) T$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(f\left(x_{t+1}\right) - f\left(x_*\right)\right) \leq \frac{1 + 4\gamma^3 L^3 n^3}{1 - 4\gamma^2 L^2 n^2}\frac{\left\|x_0 - x_*\right\|^2}{2\gamma n T}.$$

Using the convexity of f, the average iterate $\hat{x}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} x_t$ satisfies

$$\left(f\left(\hat{x}_T\right) - f\left(x_*\right)\right) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f\left(x_t\right) - f\left(x_*\right)\right) \leq \frac{1 + 4\gamma^3 L^3 n^3}{1 - 4\gamma^2 L^2 n^2}\frac{\left\|x_0 - x_*\right\|^2}{2\gamma n T}.$$

Let us show that

$$\frac{1 + 4\gamma^3 L^3 n^3}{1 - 4\gamma^2 L^2 n^2} \leq 4.$$

Applying $\gamma \leq \frac{1}{2\sqrt{2}Ln}$ we have

$$\frac{1 + 4\frac{1}{16\sqrt{2}L^3 n^3}L^3 n^3}{1 - 4\frac{1}{8L^2 n^2}L^2 n^2} = \frac{1 + \frac{1}{4\sqrt{2}}}{1 - \frac{1}{2}} \leq 4.$$

Finally, we have

$$f\left(\hat{x}_T\right) - f\left(x_*\right) \leq \frac{2\left\|x_0 - x_*\right\|^2}{\gamma n T}.$$

This ends the proof.

F ONE MORE ALGORITHM: RR-VR

F.1 NEW ALGORITHM: RR-VR

Algorithm 1 Random Reshuffling with Variance Reduction

Input: Stepsize $\gamma > 0$, probability p, $x_0 = x_0^0 \in \mathbb{R}^d, y_0 \in \mathbb{R}^d$, number of epochs T.
for $t = 0, 1, \ldots T-1$ **do**
  Choose a random permutation $\{\pi_0, \ldots, \pi_{n-1}\}$ of $\{1, \ldots, n\}$
  $x_t^0 = x_t$
  for $i = 0, \ldots, n-1$ **do**
    $g_t^i(x_t^i, y_t) = \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(y_t) + \nabla f(y_t)$
    $x_t^{i+1} = x_t^i - \gamma g_t^i(x_t^i, y_t)$
  end for
  $x_{t+1} = x_t^n$
  $y_{t+1} = \begin{cases} y_t & \text{with probability } 1-p \\ x_t & \text{with probability } p \end{cases}$
end for

In this section we formulate convergence results for a generalized version of SVRG under random reshuffling. Analysis of RR-VR (Algorithm 1) is more complicated.

F.2 CONVERGENCE THEORY

To analyze this method, we introduce Lyapunov functions.

Suppose that each f_i is convex, f is μ-strongly convex, and Assumption 1 holds. Then provided the parameters satisfy $n > \kappa$, $\frac{\kappa}{n} < p < 1$ and $\gamma \leq \frac{1}{2\sqrt{2}Ln}$, the final iterate generated by RR-VR (Algorithm 1) satisfies $V_T \leq \max(q_1, q_2)^T V_0$, where $q_1 = 1 - \frac{\gamma\mu n}{4}\left(1 - \frac{p}{2}\right)$, $q_2 = 1 - p + \frac{8}{\mu}\gamma^2 L^3 n$, and the Lyapunov function is defined via

$$V_t := \mathbb{E}\left[\|x_t - x_*\|^2\right] + \left(\frac{4}{\gamma\mu n}\right)^{-1} \mathbb{E}\left[\|y_t - x_*\|^2\right].$$

This means that the iteration complexity of Algorithm 1 is $T = \mathcal{O}\left(\kappa \log\left(\frac{1}{\varepsilon}\right)\right)$.

Note that the probability p should not be too small. We obtain the same complexity as that of of Rand-Reshuffle.

Suppose that the functions $f_1, \ldots, f_n$ are μ-strongly convex, and that Assumption **??** holds. Then for RR-VR (Algorithm 1) with parameters that satisfy $\gamma \leq \frac{1}{2L}\sqrt{\frac{\mu}{2nL}}$, $\frac{1}{2} < \delta < \frac{1}{\sqrt{2}}$, $0 < p < 1$, and for a sufficiently large number of functions, $n > \log\left(\frac{1}{1-\delta^2}\right) \cdot \left(\log\left(\frac{1}{1-\gamma\mu}\right)\right)^{-1}$, the iterates generated by the RR-VR algorithm satisfy $V_T \leq \max(q_1, q_2)^T V_0$, where $q_1 = (1-\gamma\mu)^n + \delta^2$, $q_2 = 1 - p\left(1 - \frac{2\gamma^2 L^3 n}{\mu\delta^2}\right)$, and

$$V_t := \mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{\delta^2}{p}\mathbb{E}\left[\|y_t - x_*\|^2\right].$$

This means that the iteration complexity of Algorithm 1 is $T = \mathcal{O}\left(\max\left(\kappa\sqrt{\frac{\kappa}{n}}, \frac{1}{2\log(2\delta)}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.

We get almost the same rate as the rate of Rand-Reshuffle, but there is one difference. Complexity depends on δ term. However, the first term dominates in most cases.

F.3 PROOF OF THEOREM 9

Suppose that each f_i is convex, f is μ-strongly convex, and Assumption 1 holds. Then provided the parameters satisfy $n > \kappa$, $\frac{\kappa}{n} < p < 1$ and $\gamma \leq \frac{1}{2\sqrt{2}Ln}$, the final iterate generated by RR-VR (Algorithm 1) satisfies

$$V_T \leq \max(q_1, q_2)^T V_0,$$

where

$$q_1 = 1 - \frac{\gamma\mu n}{4}\left(1 - \frac{p}{2}\right), \quad q_2 = 1 - p + \frac{8}{\mu}\gamma^2 L^3 n,$$

and the Lyapunov function is defined via

$$V_t \coloneqq \mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{4}{\gamma\mu n}\mathbb{E}\left[\|y_t - x_*\|^2\right].$$

Proof. For the problem $\frac{1}{n}\sum_{i=1}^n f_i^t(x)$ we will use an inequality from Mishchenko et al. [2020]:

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t\right] &\leq \frac{1}{1+\gamma\mu n}\left(\|x_t - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\right) \\ &= \frac{1}{1+\gamma\mu n}\|x_t - x_*\|^2 + \frac{1}{1+\gamma\mu n}\frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\leq \left(1 - \frac{\gamma\mu n}{2}\right)\|x_t - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2}.\end{aligned}$$

Now we apply inequality

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t, y_t\right] &\leq \left(1 - \frac{\gamma\mu n}{2}\right)\|x_t - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\leq \left(1 - \frac{\gamma\mu n}{2}\right)\|x_t - x_*\|^2 + 2\gamma^3 L^3 n^2 \|y_t - x_*\|^2.\end{aligned}$$

Using tower property we have

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t, y_t\right]\right] \\ &\leq \left(1 - \frac{\gamma\mu n}{2}\right)\mathbb{E}\left[\|x_t - x_*\|^2\right] + 2\gamma^3 L^3 n^2 \mathbb{E}\left[\|y_t - x_*\|^2\right].\end{aligned}$$

Now we look at

$$y_{t+1} = \begin{cases} y_t & \text{with probability } 1-p \\ x_t & \text{with probability } p \end{cases}.$$

We get

$$\mathbb{E}\left[\|y_{t+1} - x_*\|^2 \mid x_t, y_t\right] = (1-p)\|y_t - x_*\|^2 + p\|x_t - x_*\|^2.$$

Using tower property

$$\begin{aligned}\mathbb{E}\left[\|y_{t+1} - x_*\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|y_{t+1} - x_*\|^2 \mid x_t, y_t\right]\right] \\ &= (1-p)\mathbb{E}\left[\|y_t - x_*\|^2\right] + p\mathbb{E}\left[\|x_t - x_*\|^2\right].\end{aligned}$$

Finally, we have

$$\begin{aligned}\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] + M\mathbb{E}\left[\|y_{t+1} - x_*\|^2\right] \leq \left(1 - \frac{\gamma\mu n}{2}\right)\|x_t - x_*\|^2 + 2\gamma^3 L^3 n^2 \mathbb{E}\left[\|y_t - x_*\|^2\right] \\ + (1-p)M\mathbb{E}\|y_t - x_*\|^2 + pM\mathbb{E}\|x_t - x_*\|^2.\end{aligned}$$

Denote $V_t = \mathbb{E}\left[\|x_t - x_*\|^2\right] + M\mathbb{E}\left[\|y_t - x_*\|^2\right]$. Using this we obtain

$$\begin{aligned}V_{t+1} \leq \left(1 - \frac{\gamma\mu n}{2}\right)\mathbb{E}\left[\|x_t - x_*\|^2\right] + 2\gamma^3 L^3 n^2 \mathbb{E}\left[\|y_t - x_*\|^2\right] \\ + (1-p)M\mathbb{E}\left[\|y_t - x_*\|^2\right] + pM\mathbb{E}\left[\|x_t - x_*\|^2\right].\end{aligned}$$

Thus,

$$V_{t+1} \leq \left(1 - \frac{\gamma\mu n}{2} + pM\right) \mathbb{E}\left[\|x_t - x_*\|^2\right] + \left(1 - p + \frac{1}{M} 2\gamma^3 L^3 n^2\right) M\mathbb{E}\left[\|y_t - x_*\|^2\right].$$

To have contraction we use

$$M = \frac{\gamma\mu n}{4}, \qquad \gamma = \frac{1}{2\sqrt{2}Ln}.$$

We have the final rate

$$V_{t+1} \leq \max\left(1 - \frac{\gamma\mu n}{4}\left(1 - \frac{p}{2}\right), 1 - p + \frac{8}{\mu}\gamma^2 L^3 n\right) V_t$$
$$V_T \leq \max\left(1 - \frac{\gamma\mu n}{4}\left(1 - \frac{p}{2}\right), 1 - p + \frac{8}{\mu}\gamma^2 L^3 n\right)^T V_0.$$

□

F.4 PROOF OF THEOREM 10

Suppose that the functions $f_1, \ldots, f_n$ are μ-strongly convex, and that Assumption 1 holds. Then for RR-VR (Algorithm 1) with parameters that satisfy $\gamma \leq \frac{1}{2L}\sqrt{\frac{\mu}{2nL}}$, $\frac{1}{2} < \delta < \frac{1}{\sqrt{2}}$, $0 < p < 1$, and for a sufficiently large number of functions, $n > \log\left(\frac{1}{1-\delta^2}\right) \cdot \left(\log\left(\frac{1}{1-\gamma\mu}\right)\right)^{-1}$, the iterates generated by the RR-VR algorithm satisfy

$$V_T \leq \max\left(q_1, q_2\right)^T V_0,$$

where

$$q_1 = (1 - \gamma\mu)^n + \delta^2, \quad q_2 = 1 - p\left(1 - \frac{2\gamma^2 L^3 n}{\mu\delta^2}\right),$$

and

$$V_t \coloneqq \mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{\delta^2}{p}\mathbb{E}\left[\|y_t - x_*\|^2\right].$$

Proof. For the problem $\frac{1}{n}\sum_{i=1}^n f_i^t(x)$ we will use two inequalities from Mishchenko et al. [2020]:

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t\right] \leq (1 - \gamma\mu)^n \|x_t - x_*\|^2 + 2\gamma^2 \sigma^2_{\text{Shuffle}} \left(\sum_{i=0}^{n-1}(1 - \gamma\mu)^i\right)$$
$$\sigma^2_{\text{Shuffle}} \leq \frac{\gamma L n}{4}\sigma_*^2.$$

Using this result, we have

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t, y_t\right] \leq (1 - \gamma\mu)^n \|x_t - x_*\|^2 + \frac{1}{2}\gamma^3 L n \sigma_*^2 \left(\sum_{i=0}^{n-1}(1 - \gamma\mu)^i\right)$$
$$\leq (1 - \gamma\mu)^n \|x_t - x_*\|^2 + \frac{1}{\mu} 2\gamma^2 L^2 n L \|y_t - x_*\|^2.$$

Using tower property

$$\mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|x_{t+1} - x_*\|^2 \mid x_t, y_t\right]\right]$$
$$\leq (1 - \gamma\mu)^n \mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{1}{\mu} 2\gamma^2 L n L^2 \mathbb{E}\left[\|y_t - x_*\|^2\right].$$

Now we look at

$$y_{t+1} = \begin{cases} y_t & \text{with probability } 1-p \\ x_t & \text{with probability } p \end{cases}.$$

Thus, $\mathbb{E}\left[\|y_{t+1} - x_*\|^2 \mid x_t, y_t\right] = (1-p)\|y_t - x_*\|^2 + p\|x_t - x_*\|^2$. Using tower property

$$\begin{aligned}\mathbb{E}\left[\|y_{t+1} - x_*\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|y_{t+1} - x_*\|^2 \mid x_t, y_t\right]\right] \\ &= (1-p)\mathbb{E}\left[\|y_t - x_*\|^2\right] + p\mathbb{E}\left[\|x_t - x_*\|^2\right].\end{aligned}$$

Denote $V_t = \mathbb{E}\left[\|x_t - x_*\|^2\right] + M\mathbb{E}\left[\|y_t - x_*\|^2\right]$ and we have

$$\begin{aligned}V_{t+1} &= \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] + M\mathbb{E}\left[\|y_{t+1} - x_*\|^2\right] \\ &\leq (1-\gamma\mu)^n \mathbb{E}\left[\|x_t - x_*\|^2\right] + \frac{2}{\mu}\gamma^2 L^3 n\mathbb{E}\left[\|y_t - x_*\|^2\right] + (1-p)M\mathbb{E}\left[\|y_t - x_*\|^2\right] + pM\mathbb{E}\left[\|x_t - x_*\|^2\right] \\ &\leq \left((1-\gamma\mu)^n + pM\right)\mathbb{E}\left[\|x_t - x_*\|^2\right] + \left((1-p) + \frac{2\gamma^2 L^3 n}{\mu M}\right)M\mathbb{E}\left[\|x_t - x_*\|^2\right] \\ &\leq \max\left(\left((1-\gamma\mu)^n + pM\right), \left((1-p) + \frac{2\gamma^2 L^3 n}{\mu M}\right)\right)V_t.\end{aligned}$$

Unrolling the recusrion we have

$$V_T \leq \max\left(\left((1-\gamma\mu)^n + pM\right), \left(1 - p + \tfrac{2\gamma^2 L^3 n}{\mu M}\right)\right)^T V_0.$$

Applying $M = \frac{\delta^2}{p}$ and $\gamma \leq \frac{1}{2L}\sqrt{\frac{\mu}{2nL}}$ we get

$$V_T \leq \max\left((1-\gamma\mu)^n + \delta^2, 1 - p\left(1 - \frac{2\gamma^2 L^3 n}{\mu\delta^2}\right)\right)^T V_0.$$

□

References

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.