SGD without replacement

Raghav Somani

March 24, 2019

This article focuses on one of the open questions put up by Léon Bottou [1] in 2009 where he observes a contrasting behavior in the performance of Stochastic Gradient Descent (SGD) when the strictly convex component functions of the objective function are chosen without replacement in comparison to when they are chosen with replacement. When sampled without replacement, observations suggest that the convergence rate in expectation is very close to t^{-2} where t is the number of iterations of SGD under the choice of sampling. From the light of the theoretical works associated with stochastic approximations [5], stochastic algorithms that converge faster than t^{-1} is very surprising.

1 Problem Setup

Consider the standard finite sum optimization problem

$$F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}; i) \tag{1.1}$$

where $f(\cdot; i) : \mathbb{R}^d \to \mathbb{R}$ is the i-th component function. The goal is to find the minimizer of F on a closed convex set $\mathcal{W} \subset \mathbb{R}^d$.

$$\min_{\mathbf{x}\in\mathcal{W}} F(\mathbf{x}) \tag{1.2}$$

The vanilla SGD update can be written as

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{W}}\left(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t; i_t)\right) \qquad \text{with } \mathbf{x}_0 = \mathbf{0} \tag{1.3}$$

where i_t is selected uniformly from $[n]$ with replacement yielding $\mathbb{E}_{i_t}\left[\nabla f(\mathbf{x}_t; i_t)\right] = \nabla F(\mathbf{x}_t)$, $\eta_t > 0$ is the step size at t-th iteration, and $\Pi_{\mathcal{W}}$ is the Euclidean projection operator on the set $\mathcal{W}$.
We can also consider a similar version of SGD where for every pass on the data, i.e., for every epoch, a random permutation $\sigma_k : [n] \to [n]$ is chosen uniformly, and the i-th update of the k-th epoch is performed along negative gradient of the $\sigma_k(i+1)$-th component of F. Concisely,

$$\mathbf{x}_{i+1}^k = \Pi_{\mathcal{W}}\left(\mathbf{x}_i^k - \eta_{k,i}\nabla f(\mathbf{x}_i^k; \sigma_k(i+1))\right) \qquad \forall\, i \in [n], k \in [K] \text{ and } \mathbf{x}_0^k := \mathbf{x}_n^{k-1}, \mathbf{x}_0^1 = \mathbf{0} \tag{1.4}$$

where $\eta_{k,i} > 0$ is the step size of the i-th iterate of k-th epoch. This algorithm is nothing but SGD without replacement where the algorithms passes over the data on a random permutations.

2 Related Works

There have been works which show that the convergence rates of SGD without replacement after K epochs behaves as $\mathcal{O}\left(1/K^2\right)$ [2], where the sub-optimality of SGD is known to be $\mathcal{O}\left(1/nK\right)$ and is known to be tight. [3] improves upon the results of [2] showing a sub-optimality bound of $\mathcal{O}\left(1/n^2K^2 + 1/K^3\right)$. However, these works require Hessian Lipschitz, gradient Lipschitz and strong convexity assumptions on F. In contrast, SGD's rate of $\mathcal{O}\left(1/nK\right)$ only requires the strong convexity assumption.

This article refers to a recent work [4] which try to answer the question - *Does SGD without replacement converge at a faster rate than SGD with replacement for general smooth, strongly convex functions without the Hessian Lipschitz condition?*.

The domain of interest is when the number of passes over the data is small. [6] consider a single pass of the data and show that for generalized linear models, the sub-optimality bounds are similar to that of SGD which is $\mathcal{O}\left(1/n\right)$ and $\mathcal{O}\left(1/\sqrt{n}\right)$ for convex functions with and without strong convexity respectively.

3 Discussion and Analysis

One of the major challenge is due to the fact that sampling without replacement leads to coupling between iterates and the gradients, and in expectation, the update does not follow Gradient Descent (GD), i.e.,

$$\mathbb{E}\left[\nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \mid \mathbf{x}_{i-1}^k, \sigma(1:i)\right] \neq \nabla F(\mathbf{x}_i^k) \qquad \forall\ i > 1\ \forall\ k \geq 1 \tag{3.1}$$

Let $\mathbf{x}^*$ be the minimizer of F over $\mathcal{W}$. We consider that the component functions are twice differentiable, uniformly G Lipschitz and L smooth over $\mathcal{W}$. Let us have following set of assumptions

1. Lipschitz continuity - $\exists\ G > 0 \ni \|\nabla f(\mathbf{x}; i)\|_2 \leq G\ \forall \mathbf{x} \in \mathcal{W}, i \in [n]$.
2. Smoothness Gradient Lipschitz - $\exists\ L > 0 \ni \|\nabla f(\mathbf{x}; i) - \nabla f(\mathbf{y}; i)\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2\ \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{W}, i \in [n]$.
3. Strongly convex - $\exists\ \mu > 0 \ni F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2\ \ \forall\ \mathbf{x}, \mathbf{y} \in \mathcal{W}$.

The condition number of the problem (1.2) is defined as $\kappa := {}^{L}/_{\mu}$. We also denote the distance of the initial point $\mathbf{x}_0^1$ from the optimum by $D := \left\|\mathbf{x}_0^1 - \mathbf{x}^*\right\|_2$.

3.1 Rates for GD and SGD

It has been shown that GD satisfies the below rates under the above discussed assumptions for K iterations

1. With assumption 1: $\mathcal{O}\left({}^{GD}/_{\sqrt{K}}\right)$.
2. With assumptions 1 and 2: $\mathcal{O}\left({}^{LD^2}/_{K}\right)$.
3. With assumptions 1, 2 and 3: $\mathcal{O}\left(LD^2 e^{-K/\kappa}\right)$.

Similar tight rates have been shown for SGD as well. Here K passes over the data implies $T = nK$ IFO calls.

1. With assumption 1: $\mathcal{O}\left({}^{GD}/_{\sqrt{nK}}\right)$.
2. With assumption 1 and 3: $\mathcal{O}\left({}^{G^2}/_{\mu nK}\right)$.
3. With assumptions 1, 2 and 3: Variance reduction methods of SGD like SVRG, SAGA, SAG and SDCA achieve faster rates of convergence.

None of these results apply to SGD without replacement due to the dependencies between iterates and the gradients. [7] show that for a small enough step size, the distribution of iterates of SGD without replacement converge closer to the optimum than the iterates of SGD.

3.2 Coupling and Wasserstein distance

The main problem with the classical analysis tools for SGD with replacement when applied to SGD without replacement is that because the iterates are dependent, $\mathbb{E}\left[f(\mathbf{x}_i^k; \sigma_k(i+1))\right] \neq \mathbb{E}\left[F(\mathbf{x}_i^k)\right]$. However the two can be shown to be comparable. It can also be shown that in expectation, SGD without replacement over one epoch approximates one step of GD applied on F. Therefore K epochs of SGD without replacement would approximate GD after K iterations.

$$\mathbf{x}_0^{k+1} = \mathbf{x}_0^k - \eta_k \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \tag{3.2}$$

If $\mathbf{x}_i^k \simeq \mathbf{x}_0^k$, the above equation implies

$$\mathbf{x}_0^{k+1} \simeq \mathbf{x}_0^k - \eta_k \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_0^k; \sigma_k(i+1)) = \mathbf{x}_0^k - n\eta_k \nabla F(\mathbf{x}_0^k) \tag{3.3}$$

To show both the claims, consider two independent permutations σ_k and σ_k' after k epochs of SGD without replacement. Starting from $\mathbf{x}_0^k$, denote the iterates of k-th epoch with σ_k as $(\mathbf{x}_i(\sigma_k))_{i=1}^n$ and with σ_k' as $(\mathbf{x}_i(\sigma_k'))_{i=1}^n$. Therefore $(\mathbf{x}_i(\sigma_k))_{i=1}^n$ and $(\mathbf{x}_i(\sigma_k'))_{i=1}^n$ are independent of each other and identically distributed. Which implies

$$\mathbb{E}\left[f(\mathbf{x}_i(\sigma_k'); \sigma_k(i+1))\right] = \mathbb{E}\left[F(\mathbf{x}_i^k)\right] \tag{3.4}$$

Therefore we need to show that $\mathbb{E}\left[f(\mathbf{x}_i(\sigma'_k);\sigma_k(i+1))\right] - \mathbb{E}\left[f(\mathbf{x}_i(\sigma_k);\sigma_k(i+1))\right] \simeq 0$. Since $f(\cdot;j)$ is Lipschitz $\forall\ j \in [n]$, so a bound on distance between $\mathbf{x}_i(\sigma_k)$ and $\mathbf{x}_i(\sigma'_k)$ is required.
To prove the above claim, we need to set up some more notation and definitions. Let $\mathcal{D}_{i,k} := \mathcal{L}(\mathbf{x}_i(\sigma_k))$ and $\mathcal{D}^{(r)}_{i,k} := \mathcal{L}(\mathbf{x}_i(\sigma_k) \mid \sigma_k(i+1) = r)$. Here $\mathcal{L}(X)$ denotes the distribution of the random variable X. Let $\mathsf{Lip}_d(\beta)$ denote the set of all β Lipschitz functions from $\mathbb{R}^d \to \mathbb{R}$.

Definition 3.1. Let P and Q be two probability measures over $\mathbb{R}^d$ s.t. $\mathbb{E}_{X\sim P}\left[\|X\|_2^2\right] < \infty$ and $\mathbb{E}_{Y\sim Q}\left[\|Y\|_2^2\right] < \infty$. Let $X \sim P$ and $Y \sim Q$ be random vectors defined on a common measure space, then the Wasserstein-1 and Wasserstein-2 distance between P and Q are defined as

$$\mathcal{W}_1(P,Q) := \inf_{\substack{(X,Y):\\ X\sim P, Y\sim Q}} \mathbb{E}\left[\|X-Y\|_2\right], \text{ and} \tag{3.5}$$

$$\mathcal{W}_2(P,Q) := \inf_{\substack{(X,Y):\\ X\sim P, Y\sim Q}} \sqrt{\mathbb{E}\left[\|X-Y\|_2^2\right]} \tag{3.6}$$

respectively. Here the infimum is over all joint distributions over (X,Y) with prescribed marginals.

From Jensen's inequality, we have $\mathcal{W}_1(P,Q) \le \mathcal{W}_2(P,Q)$. There is a fundamental characterization of Wasserstein-1 distance from Kantorovich's duality as

$$\mathcal{W}_1(P,Q) := \sup_{g\in\mathsf{Lip}_d(1)} \mathbb{E}\left[g(X)\right] - \mathbb{E}\left[g(Y)\right] \tag{3.7}$$

We can now use the above definitions and characterizations to show that the approximation error $\left|\mathbb{E}\left[F(\mathbf{x}_i^k)\right] - \mathbb{E}\left[f(\mathbf{x}_i(\sigma_k);\sigma_k(i+1))\right]\right|$ is bounded in terms of the average Wasserstein distance between $\mathcal{D}_{i,k}$ and $\mathcal{D}^{(r)}_{i,k}$.

$$\begin{aligned}
\left|\mathbb{E}\left[F(\mathbf{x}_i^k)\right] - \mathbb{E}\left[f(\mathbf{x}_i(\sigma_k);\sigma_k(i+1))\right]\right| &= \left|\mathbb{E}\left[\frac{1}{n}\sum_{r=1}^{n} f(\mathbf{x}_i(\sigma'_k);r)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{r=1}^{n} f(\mathbf{x}_i(\sigma_k);r) \mid \sigma_k(i+1)=r\right]\right| \\
&\le \frac{1}{n}\sum_{r=1}^{n} \left|\mathbb{E}\left[f(\mathbf{x}_i(\sigma'_k);r)\right] - \mathbb{E}\left[f(\mathbf{x}_i(\sigma_k);r) \mid \sigma_k(i+1)=r\right]\right| \\
&\le \frac{1}{n}\sum_{r=1}^{n} \sup_{g\in\mathsf{Lip}_d(G)} \left(\mathbb{E}\left[g(\mathbf{x}_i(\sigma'_k))\right] - \mathbb{E}\left[g(\mathbf{x}_i(\sigma_k)) \mid \sigma_k(i+1)=r\right]\right) \\
&= \frac{G}{n}\sum_{r=1}^{n} \mathcal{W}_1(\mathcal{D}_{i,k}, \mathcal{D}^{(r)}_{i,k}) \\
&\le \frac{G}{n}\sum_{r=1}^{n} \mathcal{W}_2(\mathcal{D}_{i,k}, \mathcal{D}^{(r)}_{i,k})
\end{aligned} \tag{3.8}$$

We are now left to bound $\mathcal{W}_2(\mathcal{D}_{i,k}, \mathcal{D}^{(r)}_{i,k})$. From the definition of $\mathcal{W}_2$, we have

$$\mathcal{W}_2(\mathcal{D}_{i,k}, \mathcal{D}^{(r)}_{i,k}) \le \sqrt{\mathbb{E}\left[\|\mathbf{x}_i(\sigma_k) - \mathbf{x}_i(\sigma'_k)\|_2^2\right]} \tag{3.9}$$

such that $\mathbf{x}_i(\sigma_k) \sim \mathcal{D}_{i,k}$ and σ'_k is such that $\sigma'_k(i+1) = r$. Since (3.9) holds for all σ'_k with this property, consider the permutation which is obtained from σ_k by swapping at most one pair such that $\sigma'_k(i+1) = r$ holds true. Let $j < i$, and first assume $\sigma_k(j+1) \neq \sigma'_k(j+1)$, then

$$\begin{aligned}
\|\mathbf{x}_{j+1}(\sigma_k) - \mathbf{x}_{j+1}(\sigma'_k)\|_2 &= \|\Pi_{\mathcal{W}}\left(\mathbf{x}_j(\sigma_k) - \eta_{k,i}\nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1))\right) - \Pi_{\mathcal{W}}\left(\mathbf{x}_j(\sigma'_k) - \eta_{k,i}\nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1))\right)\|_2 \\
&\le \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k) - \eta_{k,i}\left(\nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1)) - \nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1))\right)\|_2 \\
&\le \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2 + 2G\eta_{k,i} \\
&\le \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2 + 2G\eta_{k,0}
\end{aligned}$$

If $\sigma_k(j+1) = \sigma'_k(j+1)$,

$$\begin{aligned}
\|\mathbf{x}_{j+1}(\sigma_k) - \mathbf{x}_{j+1}(\sigma'_k)\|_2^2 &= \|\Pi_{\mathcal{W}}\left(\mathbf{x}_j(\sigma_k) - \eta_{k,i}\nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1))\right) - \Pi_{\mathcal{W}}\left(\mathbf{x}_j(\sigma'_k) - \eta_{k,i}\nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1))\right)\|_2^2 \\
&\leq \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k) - \eta_{k,i}\left(\nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1)) - \nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1))\right)\|_2^2 \\
&= \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2^2 - 2\eta_{k,i}\left\langle \nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1)) - \nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1)), \mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\right\rangle \\
&\qquad + \eta_{k,i}^2\|\nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1)) - \nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1))\|_2^2 \\
&\leq \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2^2 \\
&\qquad - (2\eta_{k,i} - L\eta_{k,i}^2)\left\langle \nabla f(\mathbf{x}_j(\sigma_k);\sigma_k(j+1)) - \nabla f(\mathbf{x}_j(\sigma'_k);\sigma'_k(j+1)), \mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\right\rangle \\
&\leq \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2^2 \qquad \text{(if } \eta_{k,0} \leq 2/L\text{, then } (2\eta_{k,i} - L\eta_{k,i}^2) \geq 0) \\
\implies \|\mathbf{x}_{j+1}(\sigma_k) - \mathbf{x}_{j+1}(\sigma'_k)\|_2 &\leq \|\mathbf{x}_j(\sigma_k) - \mathbf{x}_j(\sigma'_k)\|_2 \qquad (3.10)
\end{aligned}$$

Since $|\{j < i \mid \sigma_k(j+1) \neq \sigma'_k(j+1)\}| \leq 1$, we have

$$\|\mathbf{x}_i(\sigma_k) - \mathbf{x}_i(\sigma'_k)\|_2 \leq 2G\eta_{k,0} \qquad (3.11)$$

Using (3.11) in (3.9) we have

$$\mathcal{W}_2(\mathcal{D}_{i,k}, \mathcal{D}_{i,k}^{(r)}) \leq 2G\eta_{k,0} \qquad (3.12)$$

Using (3.12) in (3.8) we have

$$\left|\mathbb{E}\left[F(\mathbf{x}_i^k)\right] - \mathbb{E}\left[f(\mathbf{x}_i(\sigma_k);\sigma_k(i+1))\right]\right| \leq 2G^2\eta_{k,0} \qquad (3.13)$$

thus proving the claim.
We have now seen that the iterates of SGD with and without replacement are close in sub-optimality. SGD without replacement also has a property of variance reduction in some sense. We will now see that the iterates $\mathbf{x}_i^k$ do not move much when they are close to the optimum. Let $\hat{\mathbf{x}} \in \left\{\mathbf{x}_0^k, \mathbf{x}^*\right\}$, which implies that it is independent of σ_k

$$\begin{aligned}
\left\|\mathbf{x}_{i+1}^k - \hat{\mathbf{x}}\right\|_2^2 &\leq \left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2 - 2\eta_{k,i}\left\langle \nabla f(\mathbf{x}_i^k;\sigma_k(i+1)), \mathbf{x}_i^k - \hat{\mathbf{x}}\right\rangle + \eta_{k,i}^2 G^2 \\
&\leq \left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2 + 2\eta_{k,i}\left(f(\hat{\mathbf{x}};\sigma_k(i+1)) - f(\mathbf{x}_i^k;\sigma_k(i+1))\right) + \eta_{k,i}^2 G^2 \\
\implies \mathbb{E}\left[\left\|\mathbf{x}_{i+1}^k - \hat{\mathbf{x}}\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + \eta_{k,i}^2 G^2 + 2\eta_{k,i}\mathbb{E}\left[f(\hat{\mathbf{x}};\sigma_k(i+1)) - f(\mathbf{x}_i^k;\sigma_k(i+1))\right] \\
&= \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + 2\eta_{k,i}\mathbb{E}\left[F(\hat{\mathbf{x}}) - f(\mathbf{x}_i^k;\sigma_k(i+1))\right] + \eta_{k,i}^2 G^2 \\
&= \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + 2\eta_{k,i}\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}_i^k)\right] + 2\eta_{k,i}\mathbb{E}\left[F(\mathbf{x}_i^k) - f(\mathbf{x}_i^k;\sigma_k(i+1))\right] + \eta_{k,i}^2 G^2 \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + 2\eta_{k,i}\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}_i^k)\right] + 4\eta_{k,0}\eta_{k,i}G^2 + \eta_{k,i}^2 G^2 \qquad \text{(Using (3.13))} \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + 2\eta_{k,0}\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}_i^k)\right] + 5\eta_{k,0}^2 G^2 \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \hat{\mathbf{x}}\right\|_2^2\right] + 2\eta_{k,0}\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}^*)\right] + 5\eta_{k,0}^2 G^2 \qquad (3.14)
\end{aligned}$$

For $\hat{\mathbf{x}} = \mathbf{x}_0^k$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_{i-1}^k - \mathbf{x}_0^k\right\|_2^2\right] + 2\eta_{k,0}\mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right] + 5\eta_{k,0}^2 G^2 \\
&\leq 5i\eta_{k,0}^2 G^2 + 2i\eta_{k,0}\mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right] \qquad (3.15)
\end{aligned}$$

And for $\hat{\mathbf{x}} = \mathbf{x}^*$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_{i-1}^k - \mathbf{x}^*\right\|_2^2\right] + 5\eta_{k,0}^2 G^2 \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2^2\right] + 5i\eta_{k,0}^2 G^2 \qquad (3.16)
\end{aligned}$$

4 Convergence analysis

Theorem 4.1. *Suppose F satisfies assumptions 1-3. Fix $l > 0$, and let the number of epochs K be such that $K \geq 32l\kappa^2 \log(nK)$. Let $\eta_{k,i} = \eta := 4l\frac{\log(nK)}{\mu nK}$. The the following holds for the tail average $\hat{\mathbf{x}} := \frac{1}{K - \lceil K/2 \rceil + 1} \sum_{k=\lceil K/2 \rceil}^{K} \mathbf{x}_0^k$ of the iterates:*

$$\mathbb{E}\left[F(\hat{\mathbf{x}})\right] - F(\mathbf{x}^*) \leq \mathcal{O}\left(\frac{\mu D^2}{(nK)^l}\right) + \mathcal{O}\left(\frac{\kappa^2 G^2}{\mu}\frac{(\log(nK))^2}{nK^2}\right) \tag{4.1}$$

Proof. The update of SGD without replacement for the k-th epochs can be written as

$$\mathbf{x}_0^{k+1} = \mathbf{x}_0^k - \eta \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \tag{4.2}$$

Subtracting $\mathbf{x}^*$ on both sides and taking the squared Euclidean norm, we get

$$\begin{aligned}
\left\|\mathbf{x}_0^{k+1} - \mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2^2 - 2\eta \sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle + \eta^2 \left\| \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \right\|_2^2 \\
&= \left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2^2 - 2n\eta \left\langle \nabla F(\mathbf{x}_0^k), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle - 2\eta \sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla F(\mathbf{x}_0^k), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle \\
&\quad + \eta^2 \left\| \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \right\|_2^2 \\
&\leq (1 - n\eta\mu) \left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2^2 - 2n\eta \left(F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right) - 2\eta \sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla F(\mathbf{x}_0^k), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle \\
&\quad + \eta^2 \left\| \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \right\|_2^2
\end{aligned} \tag{4.3}$$

We shall analyze terms of Equation (4.3) individually.

$$\begin{aligned}
T_1 &:= -2\eta \sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla F(\mathbf{x}_0^k), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle \\
&= -2\eta \sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla f(\mathbf{x}_0^k; \sigma_k(i+1)), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle \\
\implies \mathbb{E}\left[T_1\right] &= -2\eta \mathbb{E}\left[\sum_{i=0}^{n-1} \left\langle \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla f(\mathbf{x}_0^k; \sigma_k(i+1)), \mathbf{x}_0^k - \mathbf{x}^* \right\rangle \right] \\
&\leq 2\eta L \sum_{i=0}^{n-1} \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2 \left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2 \right] \\
&\leq 2\eta L \sum_{i=0}^{n-1} \sqrt{\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2^2\right]} \sqrt{\mathbb{E}\left[\left\|\mathbf{x}_0^k - \mathbf{x}^*\right\|_2^2\right]} \\
&\leq 2\eta L n \sqrt{\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2^2\right]} \sqrt{5n\eta^2 G^2 + 2n\eta \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right]} \\
&\leq \eta L n \left[\frac{\mu \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2^2\right]}{4L} + \frac{4L(5n\eta^2 G^2 + 2n\eta \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right])}{\mu} \right] \\
&= \frac{\eta\mu n}{4} \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}_0^k\right\|_2^2\right] + 20\frac{L^2\eta^3 n^2 G^2}{\mu} + 8\frac{\eta^2 L^2 n^2}{n} \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*)\right]
\end{aligned} \tag{4.4}$$

Now consider

$$
\begin{aligned}
T_2 &:= \eta^2 \left\| \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) \right\|_2^2 \\
&= \eta^2 \left\| \sum_{i=0}^{n-1} \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla f(\mathbf{x}^*; \sigma_k(i+1)) \right\|_2^2 \\
&\leq \eta^2 \left[\sum_{i=0}^{n-1} \left\| \nabla f(\mathbf{x}_i^k; \sigma_k(i+1)) - \nabla f(\mathbf{x}^*; \sigma_k(i+1)) \right\|_2 \right]^2 \\
&\leq \eta^2 L^2 \left[\sum_{i=0}^{n-1} \left\| \mathbf{x}_i^k - \mathbf{x}^* \right\|_2 \right]^2 \\
&= \eta^2 L^2 \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \left\| \mathbf{x}_i^k - \mathbf{x}^* \right\|_2 \left\| \mathbf{x}_j^k - \mathbf{x}^* \right\|_2 \\
\implies \mathbb{E}\left[T_2\right] &\leq \eta^2 L^2 \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \mathbb{E}\left[\left\| \mathbf{x}_i^k - \mathbf{x}^* \right\|_2 \left\| \mathbf{x}_j^k - \mathbf{x}^* \right\|_2 \right] \\
&\leq \eta^2 L^2 \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sqrt{\mathbb{E}\left[\left\| \mathbf{x}_i^k - \mathbf{x}^* \right\|_2^2 \right]} \sqrt{\mathbb{E}\left[\left\| \mathbf{x}_j^k - \mathbf{x}^* \right\|_2 \right]} \\
&\leq \eta^2 L^2 n^2 \left[\mathbb{E}\left[\left\| \mathbf{x}_0^k - \mathbf{x}^* \right\|_2^2 \right] + 5n\eta^2 G^2 \right] \qquad (4.5)
\end{aligned}
$$

Plugging in (4.4) and Equation (4.5) in Equation (4.3) we get

$$
\begin{aligned}
\mathbb{E}\left[\left\| \mathbf{x}_0^{k+1} - \mathbf{x}^* \right\|_2^2 \right] \leq \left(1 - \frac{3n\eta\mu}{4} + \eta^2 n^2 L^2 \right) \mathbb{E}\left[\left\| \mathbf{x}_0^k - \mathbf{x}^* \right\|_2^2 \right] - 2n\eta \left(1 - \frac{4\eta n L^2}{\mu} \right) \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*) \right] \\
+ \frac{20 L^2 \eta^3 n^2 G^2}{\mu} + 5\eta^4 L^2 G^2 n^3 \qquad (4.6)
\end{aligned}
$$

Since $\eta = 4l\frac{\log(nK)}{\mu nK}$ and $K \geq 32l\kappa^2 \log(nK)$, we have that $\left(1 - \frac{3n\eta\mu}{4} + \eta^2 n^2 L^2\right) \leq \left(1 - \frac{n\eta\mu}{2}\right)$ and $\left(1 - \frac{4\eta n L^2}{\mu}\right) \geq 0$. Plugging these inequalities in (4.6) we have

$$
\begin{aligned}
\mathbb{E}\left[\left\| \mathbf{x}_0^{k+1} - \mathbf{x}^* \right\|_2^2 \right] &\leq \left(1 - \frac{n\eta\mu}{2} \right) \mathbb{E}\left[\left\| \mathbf{x}_0^k - \mathbf{x}^* \right\|_2^2 \right] + \frac{20 L^2 \eta^3 n^2 G^2}{\mu} + 5\eta^4 L^2 G^2 n^3 \\
&\leq \left(1 - \frac{n\eta\mu}{2} \right)^k \mathbb{E}\left[\left\| \mathbf{x}_0^1 - \mathbf{x}^* \right\|_2^2 \right] + \sum_{t=0}^{\infty} \left(1 - \frac{n\eta\mu}{2} \right)^t \left[\frac{20 L^2 \eta^3 n^2 G^2}{\mu} + 5\eta^4 L^2 G^2 n^3 \right] \\
&\leq \exp\left(-\frac{nk\eta\mu}{2} \right) \mathbb{E}\left[\left\| \mathbf{x}_0^1 - \mathbf{x}^* \right\|_2^2 \right] + \frac{40 L^2 \eta^2 n G^2}{\mu^2} + \frac{10\eta^3 L^2 G^2 n^2}{\mu} \qquad (4.7)
\end{aligned}
$$

With $\eta = 4l\frac{\log(nK)}{\mu nK}$ and $k = K/2$ Equation (4.7) becomes

$$
\mathbb{E}\left[\left\| \mathbf{x}_0^{K/2} - \mathbf{x}^* \right\|_2^2 \right] \leq \frac{1}{(nK)^l} \mathbb{E}\left[\left\| \mathbf{x}_0^1 - \mathbf{x}^* \right\|_2^2 \right] + \frac{40 L^2 \eta^2 n G^2}{\mu^2} + \frac{10\eta^3 L^2 G^2 n^2}{\mu} \qquad (4.8)
$$

From Equation (4.6) we also get

$$
\mathbb{E}\left[\left\| \mathbf{x}_0^{k+1} - \mathbf{x}^* \right\|_2^2 \right] \leq \mathbb{E}\left[\left\| \mathbf{x}_0^k - \mathbf{x}^* \right\|_2^2 \right] - n\eta \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*) \right] + \frac{20 L^2 \eta^3 n^2 G^2}{\mu} + 5\eta^4 L^2 G^2 n^3 \qquad (4.9)
$$

Summing Equation (4.9) from $k = K/2$ to K, we get

$$
n\eta \frac{\sum_{k=\lceil K/2 \rceil}^{K} \mathbb{E}\left[F(\mathbf{x}_0^k) - F(\mathbf{x}^*) \right]}{K - \lceil K/2 \rceil + 1} \leq \frac{\mathbb{E}\left[\left\| \mathbf{x}_0^{K/2} - \mathbf{x}^* \right\|_2^2 \right]}{K - \lceil K/2 \rceil + 1} + \frac{20 L^2 \eta^3 n^2 G^2}{\mu} + 5\eta^4 L^2 G^2 n^3 \qquad (4.10)
$$

From convexity of F, and using (4.8) in (4.10) we have

$$\begin{aligned}\mathbb{E}\left[F(\hat{\mathbf{x}})-F(\mathbf{x}^*)\right] &\leq \frac{2}{nK\eta}\frac{\mathbb{E}\left[\left\|\mathbf{x}_0^1-\mathbf{x}^*\right\|_2^2\right]}{(nK)^l}+\frac{80L^2\eta G^2}{\mu^2 K}+\frac{20\eta^2L^2G^2n}{\mu K}+\frac{20L^2\eta^2nG^2}{\mu}+5\eta^3L^2G^2n^2\\ &= \mathcal{O}\left(\frac{\mu D^2}{(nK)^l}\right)+\mathcal{O}\left(\frac{\kappa^2G^2}{\mu}\frac{(\log(nK))^2}{nK^2}\right)\end{aligned} \tag{4.11}$$

□

From the above proof, we can trace back the source of the improvement to see that the variance reduction claim is indeed crucial.
We see that ones $K \in \Omega\left(\kappa^2\right)$, the convergence rate for SGD without replacement gets strictly better than that of SGD with replacement. Theorem 4.1 requires $K \in \Omega\left(\kappa^2\right)$, but the rather interesting regime is when the number of epochs is relatively smaller. It can be shown that SGD without replacement is at least as good as SGD with replacement for all $K \in \mathbb{N}$.

Theorem 4.2. *Suppose F satisfies Assumptions 1-3, and let $\eta_{k,i}=\eta:=\min\left(\frac{2}{L},4l\frac{\log(nK)}{\mu nK}\right)$ for a fixed $l>0$. Then the tail average $\hat{\mathbf{x}}:=\frac{1}{n(K-\lceil K/2\rceil+1)}\sum_{k=\lceil K/2\rceil}^{K}\sum_{i=0}^{n-1}\mathbf{x}_i^k$ satisfies*

$$\mathbb{E}\left[F(\hat{\mathbf{x}})\right]-F(\mathbf{x}^*)=\mathcal{O}\left(\frac{\mu D^2}{(nK)^l}+\frac{LD^2}{(nK)^{l+1}}\right)+\mathcal{O}\left(\frac{G^2}{\mu nK}\log(nK)+\frac{L^2G^2}{\mu^3n^2K^2}(\log nK)^2\right) \tag{4.12}$$

Proof. Writing the SGD without replacement update and taking Euclidean squared norm on both sides, we have

$$\begin{aligned}\left\|\mathbf{x}_{i+1}^k-\mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_i^k-\mathbf{x}^*\right\|_2^2-2\eta\left\langle\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle+\eta^2\left\|\nabla f(\mathbf{x}_i^k;\sigma_k(i+1))\right\|_2^2\\ &\leq \left\|\mathbf{x}_i^k-\mathbf{x}^*\right\|_2^2-2\eta\left\langle\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle+\eta^2G^2\\ &\leq \left\|\mathbf{x}_i^k-\mathbf{x}^*\right\|_2^2-2\eta\left\langle\nabla F(\mathbf{x}_i^k),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle+2\eta\left\langle\nabla F(\mathbf{x}_i^k)-\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle+\eta^2G^2\\ &\leq (1-\eta\mu)\left\|\mathbf{x}_i^k-\mathbf{x}^*\right\|_2^2-2\eta(F(\mathbf{x}_i^k)-F(\mathbf{x}^*))+2\eta\left\langle\nabla F(\mathbf{x}_i^k)-\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle+\eta^2G^2\end{aligned} \tag{4.13}$$

Define $R_{i,k}:=\left\langle\nabla F(\mathbf{x}_i^k)-\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle$.

$$\begin{aligned}R_{i,k} &= \frac{1}{n}\sum_{r=0}^{n-1}\left\langle\nabla f(\mathbf{x}_i^k;r),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle-\left\langle\nabla f(\mathbf{x}_i^k;\sigma_k(i+1)),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle\\ \implies \mathbb{E}\left[R_{i,k}\right] &= \frac{1}{n}\sum_{r=0}^{n-1}\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}_i^k;r),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle\right]-\frac{1}{n}\sum_{r=0}^{n-1}\mathbb{E}\left[\left\langle\nabla f(\mathbf{x}_i^k;r),\mathbf{x}_i^k-\mathbf{x}^*\right\rangle\mid\sigma_k(i+1)=r\right]\end{aligned} \tag{4.14}$$

The above equality not just holds for $(\mathbf{x}_i^k,\mathbf{x}_i^k\mid\sigma_k(i+1))$ but also for any other pair of random variables (Y,Z_r) which follow marginal distributions $\mathcal{D}_{i,k}$ and $\mathcal{D}_{i,k}^{(r)}$ respectively. The equality doesn't depend of their coupling so we can take an advantage of it.

$$\begin{aligned}\therefore \mathbb{E}\left[R_{i,k}\right] &= \frac{1}{n}\sum_{r=0}^{n-1}\mathbb{E}\left[\langle\nabla f(Y;r),Y-\mathbf{x}^*\rangle-\langle\nabla f(Z_r;r),Z_r-\mathbf{x}^*\rangle\right]\\ &= \frac{1}{n}\sum_{r=0}^{n-1}\mathbb{E}\left[\langle\nabla f(Y;r)-\nabla f(Z_r;r),Y-\mathbf{x}^*\rangle+\langle\nabla f(Z_r;r),Y-Z_r\rangle\right]\\ &\leq \frac{1}{n}\sum_{r=0}^{n-1}\mathbb{E}\left[L\left\|Y-x^*\right\|_2\left\|Z_r-Y\right\|_2+G\left\|Z_r-Y\right\|_2\right]\\ &\leq \frac{1}{n}\sum_{r=0}^{n-1}L\sqrt{\mathbb{E}\left[\left\|Y-x^*\right\|_2^2\right]}\sqrt{\mathbb{E}\left[\left\|Z_r-Y\right\|_2^2\right]}+G\mathbb{E}\left[\left\|Z_r-Y\right\|_2\right]\end{aligned} \tag{4.15}$$

Inequality (4.15) holds for all couplings between Y and Z_r, so we can take an infimum on both sides to have

$$\begin{aligned}
\mathbb{E}\left[R_{i,k}\right] &\leq \frac{1}{n}\sum_{r=0}^{n-1} L\mathcal{W}_2\left(\mathcal{D}_{i,k}, \mathcal{D}_{i,k}^{(r)}\right)\sqrt{\mathbb{E}\left[\left\|Z_r - Y\right\|_2^2\right]} + G\mathcal{W}_2\left(\mathcal{D}_{i,k}, \mathcal{D}_{i,k}^{(r)}\right) \\
&\leq \frac{1}{n}\sum_{r=0}^{n-1} \frac{L^2}{\mu}\left[\mathcal{W}_2\left(\mathcal{D}_{i,k}, \mathcal{D}_{i,k}^{(r)}\right)\right]^2 + \frac{\mu}{4}\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] + G\mathcal{W}_2\left(\mathcal{D}_{i,k}, \mathcal{D}_{i,k}^{(r)}\right) \qquad \text{(AM-GM inequality)} \\
&\leq \frac{4L^2G^2\eta^2}{\mu} + \frac{\mu}{4}\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] + 2G^2\eta \qquad (4.16)
\end{aligned}$$

Using Equation (4.16) in Equation (4.13), we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_{i+1}^k - \mathbf{x}^*\right\|_2^2\right] &\leq \left(1 - \frac{\eta\mu}{2}\right)\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] - 2\eta\mathbb{E}\left[F(\mathbf{x}_i^k) - F(\mathbf{x}^*)\right] + \frac{8L^2G^2\eta^3}{\mu} + 4G^2\eta^2 \\
&\leq \left(1 - \frac{\eta\mu}{2}\right)\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] + \frac{8L^2G^2\eta^3}{\mu} + 4G^2\eta^2 \\
&\leq \left(1 - \frac{\eta\mu}{2}\right)^{nk}\left\|\mathbf{x}_0^1 - \mathbf{x}^*\right\|_2^2 + \sum_{t=0}^{\infty}\left(1 - \frac{\eta\mu}{2}\right)^t\left[\frac{8L^2G^2\eta^3}{\mu} + 4G^2\eta^2\right] \\
&\leq \exp\left(-\frac{n\eta k\mu}{2}\right)D^2 + \frac{16L^2G^2\eta^2}{\mu^2} + \frac{8G^2\eta}{\mu} \qquad (4.17)
\end{aligned}$$

For $k \geq \frac{K}{2}$, (4.17) becomes

$$\mathbb{E}\left[\left\|\mathbf{x}_{i+1}^k - \mathbf{x}^*\right\|_2^2\right] \leq \frac{D^2}{(nK)^l} + \frac{16L^2G^2\eta^2}{\mu^2} + \frac{8G^2\eta}{\mu} \qquad (4.18)$$

Separately we also have

$$\begin{aligned}
\left\|\mathbf{x}_{i+1}^k - \mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2 - 2\eta\left\langle\nabla f(\mathbf{x}_i^k; \sigma_k(i+1)), \mathbf{x}_i^k - \mathbf{x}^*\right\rangle + \eta^2\left\|\nabla f(\mathbf{x}_i^k; \sigma_k(i+1))\right\|_2^2 \\
&\leq \left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2 - 2\eta\left\langle\nabla f(\mathbf{x}_i^k; \sigma_k(i+1)), \mathbf{x}_i^k - \mathbf{x}^*\right\rangle + \eta^2G^2 \\
\mathbb{E}\left[\left\|\mathbf{x}_{i+1}^k - \mathbf{x}^*\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] + \eta^2G^2 - 2\eta\mathbb{E}\left[F(\mathbf{x}_i^k) - f(\mathbf{x}^*; \sigma_k(i+1))\right] + 2\eta\mathbb{E}\left[F(\mathbf{x}_i^k) - f(\mathbf{x}_i^k; \sigma_k(i+1))\right] \\
&\leq \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|_2^2\right] - 2\eta\mathbb{E}\left[F(\mathbf{x}_i^k) - F(\mathbf{x}^*)\right] + 5\eta^2G^2 \qquad \text{(Using (3.13))} \qquad (4.19)
\end{aligned}$$

Summing Equation (4.19) for $0 \leq i \leq n-1, \left\lceil\frac{K}{2}\right\rceil \leq k \leq K$, we get

$$\begin{aligned}
\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}^*)\right] &\leq \frac{1}{n\left(K - \lceil K/2\rceil + 1\right)}\sum_{k=\lceil K/2\rceil}^{K}\sum_{i=0}^{n-1}\mathbb{E}\left[F(\mathbf{x}_i^k) - F(\mathbf{x}^*)\right] \\
&\leq \frac{1}{2n\eta\left(K - \lceil K/2\rceil + 1\right)}\mathbb{E}\left[\left\|\mathbf{x}_0^{\lceil K/2\rceil} - \mathbf{x}^*\right\|_2^2\right] + \frac{5}{2}\eta G^2 \qquad (4.20)
\end{aligned}$$

Using (4.20) in (4.18) we get

$$\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}^*)\right] \leq \frac{1}{n\eta K}\left[\frac{D^2}{(nK)^l} + \frac{16L^2G^2\eta^2}{\mu^2} + \frac{8G^2\eta}{\mu}\right] + \frac{5}{2}\eta G^2 \qquad (4.21)$$

Using $\frac{1}{\eta} \leq \frac{L}{2} + \frac{nK\mu}{4l\log(nK)}$ and $\eta = \frac{4l\log(nK)}{\mu nK}$, we finally get

$$\mathbb{E}\left[F(\hat{\mathbf{x}}) - F(\mathbf{x}^*)\right] \leq \mathcal{O}\left(\frac{LD^2}{(nK)^{l+1}} + \frac{\mu D^2}{(nK)^l}\right) + \mathcal{O}\left(\frac{G^2\log(nK)}{\mu nK} + \frac{L^2G^2\log(nK)}{\mu^3n^2K^2}\right) \qquad (4.22)$$

□

It is to note that the above theorem is even true for small K. In the regime where $nK > \kappa^2$, the rate essentially boils down to $\mathcal{O}\left(\frac{G^2 \log(nK)}{\mu nK}\right)$ which matches the rate of SGD with replacement up to log factors.

When there is no strong convexity, it can be again shown that SGD without replacement is at least as good as SGD with replacement.

Theorem 4.3. *If F satisfies Assumptions 1-2, the step size $\eta = \min\left(\frac{2}{L}, \frac{D}{G\sqrt{nK}}\right)$, the average iterate of SGD without replacement $\hat{x} := \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=0}^{n-1}\mathbf{x}_i^k$ satisfies*

$$\mathbb{E}\left[F(\hat{x}) - F(\mathbf{x}^*)\right] \leq \frac{D^2 L}{4nK} + \frac{3GD}{\sqrt{nK}} \tag{4.23}$$

Proof. Summing Equation (4.19) from $k = 1$ to K and $i = 0$ to $n-1$, we have

$$\begin{aligned}
\mathbb{E}\left[F(\hat{x}) - F(\mathbf{x}^*)\right] &\leq \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=0}^{n-1}\left(F(\mathbf{x}_i^k) - F(\mathbf{x}^*)\right) \\
&\leq \mathbb{E}\left[\frac{D^2}{2\eta nK}\right] + \frac{5}{2}\eta G^2 \\
&\leq \frac{D^2}{2nK}\max\left(\frac{L}{2}, \frac{G\sqrt{nK}}{D}\right) + \frac{5G^2}{2}\min\left(\frac{2}{L}, \frac{D}{G\sqrt{nK}}\right) \\
&\leq \frac{D^2}{2nK}\left(\frac{L}{2} + \frac{G\sqrt{nK}}{D}\right) + \frac{5G^2}{2}\frac{D}{G\sqrt{nK}} \\
&\leq \frac{D^2 L}{4nK} + \frac{3GD}{\sqrt{nK}}
\end{aligned} \tag{4.24}$$

□

References

[1] Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. 2009.

[2] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Why Random Reshuffling Beats Stochastic Gradient Descent. *arXiv e-prints*, page arXiv:1510.08560, Oct 2015.

[3] Jeffery Z. HaoChen and Suvrit Sra. Random Shuffling Beats SGD after Finite Epochs. *arXiv e-prints*, page arXiv:1806.10077, Jun 2018.

[4] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. *arXiv e-prints*, page arXiv:1903.01463, Mar 2019.

[5] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv e-prints*, page arXiv:1212.2002, Dec 2012.

[6] Ohad Shamir. Without-Replacement Sampling for Stochastic Gradient Methods: Convergence Results and Application to Distributed Optimization. *arXiv e-prints*, page arXiv:1603.00570, Mar 2016.

[7] Bicheng Ying, Kun Yuan, Stefan Vlaski, and Ali H. Sayed. Stochastic Learning Under Random Reshuffling With Constant Step-Sizes. *IEEE Transactions on Signal Processing*, 67:474–489, Jan 2019.