Convergence of mini batch SGD

1 Introduction

We consider the finite-sum optimization(FSOP) problem:

$$\min_{w\in\mathbb{R}^d}\left\{F(w) := \frac{1}{N}\sum_{i=1}^{N} f(w;i)\right\} \tag{1}$$

where certain assumptions are made for the **component** functions $f(\cdot;i), \forall i \in \{1,2,\dots,N\}$ and the **objective** function F. This standard problem arises in most machine learning tasks, including logistic regression, multi-kernel learning, and neural networks training([2]). Due to the generality, open ended nature of the problem, there has been a boom in research conducted around the finite-sum minimization problem([6], [8], [12]).

The main problem with trying to solve 1, arises from the fact that in practice, F is high dimensional, which means that N is usually big. The fact that a normal gradient descent approach which computes all the gradients of the component functions will require computational resources proportional to the dimensionality of the data, coupled with the understanding that today, 1 needs to be solved on devices that have limited resources, pushed the research community to find alternative algorithms for solving the FSOP. Therefore, the attention to solving 1 was shifted to the stochastic gradient descent algorithm ([10]), due exactly to its efficiency in dealing with large scale problems that have data of high dimensions. Stochastic gradient descent instead of calculating the full gradient of the objective function F like gradient descent would do, it calculates the gradient of only one component function in each iteration. That makes it more efficient. More specifically, the updates of SGD in each iteration are:

$$w_{t+1} = w_t - \eta_k \nabla f(w_k;i), i \in \{1,2,\dots,N\} \tag{2}$$

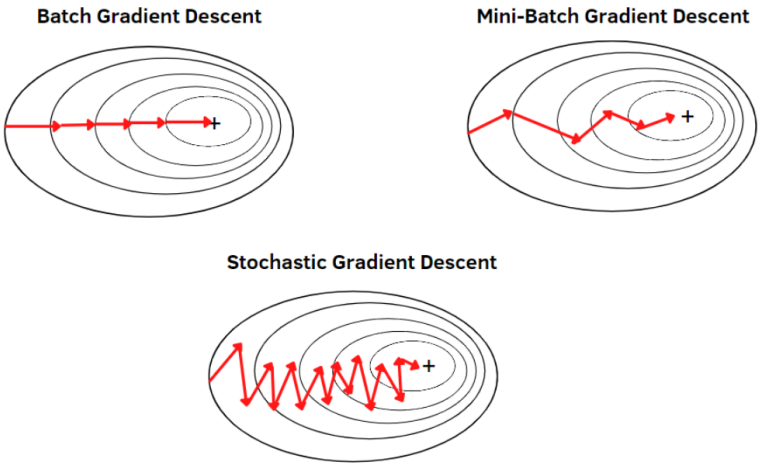
There are several ways i is chosen each iteration, that will be investigated later in section 1.1 where the updates of normal GD would be:

$$w_{t+1} = w_t - \eta_k \nabla F(w_k) = w_t - \frac{\eta_k}{N}\sum_{i=1}^{N}\nabla f(w_k;i) \tag{3}$$

Another reason why SGD attracts a lot of attention, is that it is widely used in practice in cases where the component functions can are non-convex and it converges to global minima with great success ([3]). This is something that GD is not able to achieve, as it gets stuck in local minima ([4]). The theoretical results on SGD are still being developed and are still not matching compared to what is achieved in practice, therefore pushing the community to investigate SGD's convergence gurantees.

On the downside, SGD can take longer than normal GD to converge to the stationary point. However, it requires fewer computational resources which makes it more useful in practice.

All the reasons we mentioned above lead the community to study the convergence rates of SGD. Going back to the initial discussion about GD and SGD, it seems that there is a fundamental trade-off: when one uses GD they have a better convergence bound, and computational complexity when the assumptions on F are nice but no gurantees when F is more general (for example non-convex), and when one uses SGD they get worse bounds but more computationally efficient. In this paper, we try to explicitly show this trade-off between GD and SGD. Instead of taking only the gradient of one component function (SGD) or the gradient of all the component functions (GD), we try to take the gradient of a "batch" B of the component functions. If we have $B = 1$ we get SGD and if we put $B = N$ we recover normal GD. The main question we are trying to answer is what happens when $1 < B < N$?. One can explicitly see the tradeoff in figure 1.



The algorithm for this mini-batch idea is shown below.

Algorithm 1 Mini-Batch SGD with generic permutation selection

Initialization: Choose an initial point $\tilde{w}_0 \in \mathbb{R}^d$; Mini-batch size $B \in [1, N]$; permutations $\sigma^{(i)}, i = 1, 2, \ldots, T, \sigma^{(i)} \in \mathcal{S}_N$
for $t = 1, 2, \cdots, T$ **do**
  Set $w_0^{(t)} := \tilde{w}_{t-1}$;
  for $i = 0, \cdots, \lfloor \frac{N}{B} \rfloor - 1$ **do**
    Update $w_{i+1}^{(t)} := w_i^{(t)} - \eta_k^{(i)(t)} \frac{1}{B} \sum_{j=1}^{B} \nabla f(w_i^{(t)}; \sigma^{(t)}(iB + j))$;
  end for
  Set $\tilde{w}_t := w_{\lfloor N/B \rfloor}^{(t)}$;
end for
Output: Choose $\tilde{w}_T$.

1.1 Different Algorithm Regimes

As mentioned previously, and as can be seen from algorithm 1, it remains to define how exactly we choose to permute our data $\{1, 2, \ldots, N\}$ in each epoch. In each epoch of algorithm 1, we have a permutation $\sigma^{(t)}$ that dictates in which order we are processing the data. Note that for GD we do not need this permutation because we compute the gradient of the whole objective function. There are different lines of work that study different ways in which these permutations are picked, according to the literature. The three main ones are the following:

1. **Random Reshuffling:** In each iteration of every epoch, sample a new random permutation from the family of all permutations. [5].
2. **Shuffle once:** In the beginning of the algorith, choose a permutation randomly and stick with it in all the iterations [11].
3. **Incremental gradient:** Similar to shuffle once, here a permutation is chosen (either randomly or deterministically) in the very beginning and is used for each iteration of every epoch. [7].

In this paper, we study the **generic shuffling** line of work, which does not assume anything about the selection of the permutation, and **Random Reshuffling**.

2 Our contributions

We start this section by outlining the different assumptions for which we have results on. We then specify the types of convergence we are searching for and finally We split our results into two sections: the generic shuffling regime results, and random reshuffling results.

2.1 Assumptions

First we start with an assumption of smoothness and the existence of stationary points we are interested in.

Assumption 1. *Assume that the following are satisfied:*

a) $dom(F) \stackrel{def}{=} \{x \in \mathbb{R}^d | F(x) < \infty\} \neq \emptyset$ *and* $F_* = \inf_{w \in \mathbb{R}^d} F(w) > -\infty$

b) $f(\cdot; i)$ *is L-smooth for all* $i \in [N] \stackrel{def}{=} \{1, 2, \ldots, N\}$, $u, v \in \mathbb{R}^d$, *which means that*

$$\Big\|\nabla f(u; i) - \nabla f(v; i)\Big\| \leq L\Big\|u - v\Big\|$$

Note that Assumption 1a) is establishing the well definiteness of our problem, while Assumption 1b), the so called L-Smoothness of the component functions' gradients $f(u; \cdot)$, is a standard assumption in literature. Both assumptions are used for all of our results, and contain an interesting family of functions for which we are interested in.

We continue with the so-called mean bounded variance assumption.

Assumption 2 (Mean Bounded Variance)**.** *There exists two constants* $\Theta, \sigma \in [0, \infty)$ *for which* $\forall w \in dom(F)$*:*

$$\frac{1}{N}\sum_{i=1}^{N}\Big\|\nabla f(w; i) - F(w)\Big\|^2 \leq \Theta\Big\|\nabla F(w)\Big\|^2 + \sigma^2$$

Setting $\Theta = 0$, we get exactly the bounded variance assumption, that $\frac{1}{N}\sum_{i=1}^{N}\Big\|\nabla f(w; i) - F(w)\Big\|^2 \leq \sigma^2$, which is usually assumed for non-convex objective functions ([3]). Our assumption is stronger than that and can be applied to a bigger family of functions.

Finally, in our results we use the Strong convexity assumption, which is the following.

Assumption 3 (Strong-Convexity)**.** *The objective function* F *is* $\mu-$*strongly convex which means that:*

$$F(v) \geq F(u) + \langle \nabla F(u), v - u\rangle + \frac{\mu}{2}\Big\|v - u\Big\|^2$$

It is also implied by the above condition that there is a unique solution w_* *to* F *for which holds*

$$\frac{\mu}{2}\Big\|w - w_*\Big\|^2 \leq F(w) - F(w_*) \leq \frac{2}{\mu}\Big\|\nabla F(w)\Big\|^2, \forall w \in \mathbb{R}^d$$

Substituting $\mu = 0$ and we can use this assumption for when F is just convex. When $\mu > 0$ we have stronger guarantees, for F, and usually looking for stronger convergence results, therefore we make the separation between convex and strongly convex cases. One more thing to note is that, even though F has to be convex, there might exist some component functions that are non-convex.

One more property of strongly convex functions, is that they have a unique optimal solution. This motivated us to use the bounded variance of the optimal solution and use it in our analysis for strongly convex functions.

Remark. *Because of the uniqueness of* w_* *we can denote with* σ_* *the following*

$$\sigma_*^2 \stackrel{def}{=} \frac{1}{N}\sum_{i=1}^{N}\Big\|\nabla f(w_*; i)\Big\|^2$$

2.2 Convergence types

In this section, we provide our results for the convergence of 1. Convergence to a minimum point w^* that solves 1 can be equivalent to proving any of the following:

$$\Big\|\nabla F(w_T)\Big\|^2 \leq \epsilon \qquad \Big\|w_T - w^*\Big\|^2 \leq \epsilon \qquad F(w_T) - F(w^*) \leq \epsilon$$

Note that w_T is the point we end up after T epochs, and in the end we try to reach w^*, the point that minimizes 1. All of the below results we show prove a convergence type of the three aforementioned ones. Convergence results

can further be categorized in **last-iterate**, **average-iterate** and **min-iterate**. For the purposes of this paper, we only use last and average iterate convergence. Average iterate convergence for the above types will look like this:

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(w_t)\right\|^2 \leq \epsilon \qquad \frac{1}{T}\sum_{t=1}^{T}\left\|w_t - w^*\right\|^2 \leq \epsilon \qquad \frac{1}{T}\sum_{t=1}^{T}[F(w_t) - F(w^*)] \leq \epsilon$$

From our generic shuffling results, Theorems 5 and 6 are of average-iterate convergence, and theorem 4 is of last iterate convergence.

2.3 Generic Shuffling results

We have the following results for the strongly convex, convex and non-convex cases.

Theorem 4 (Strongly-convex)**.** *Under assumptions 1 and 3, after T epochs for sufficiently big enough T, and $\eta_t = \frac{6\log(T)}{\mu T}\frac{B}{N}$ we have that*

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2\left(2\mu^2 + 3L^2\right)\log^2(T)}{T^2\mu^3} \tag{4}$$

Theorem 5 (Convex Case)**.** *Under assumptions 1 and 3, after T epochs for sufficiently big enough T, where $\eta_t \leq \min\{\frac{B}{N}\sqrt[3]{\frac{3BN\|\tilde{w}_0 - w_*\|^2}{T\sigma_*^2}}, \frac{B}{N}\frac{B+2}{2L}\}$ we have that*

$$\frac{1}{T}\sum_{t=0}^{T}[F(\tilde{w}_{t-1}) - F(w_*)] \leq \sqrt[3]{\frac{\|\tilde{w}_0 - w_*\|^4\sigma_*^2}{24T^2BN}} \tag{5}$$

Theorem 6 (Non-Convex Case)**.** *We have that under assumptions 1 and 2, after T epochs for sufficiently large enough T with $\eta_t \leq \frac{B}{NL}\sqrt{\frac{1}{3}}$:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\left\|\nabla F(\tilde{w}_t)\right\|^2] \leq \frac{4}{\eta T}(F(\tilde{w}_0) - F^*) + 6\eta^2L^2\sigma^2 \tag{6}$$

They are heavily inspired by [8]. The full proofs of these results are in the appendix, however we provide a sketch for the non-convex and strongly convex proofs here.

Both Theorem 4 and Theorem 6, start by bounding the following quantity:

$$I = \sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2$$

Then they use the following standard inequality that is Implied from the L-smoothness of F ([1]):

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \langle\nabla F(\tilde{w}_t), \tilde{w}_{t+1} - \tilde{w}_t\rangle + \frac{L}{2}\left\|\tilde{w}_{t+1} - \tilde{w}_t\right\|^2$$

2.4 Random Reshuffling results

In this section, we present two theorems that are for the Random Reshuffling Regime. These results are **novel**, and have high probability convergence bounds, i.e. they hold with probability $1-\epsilon$, where ϵ is a small constant. Let us start the discussion with the non-convex result:

Theorem 7 (Non-convex)**.** *We have that under assumptions 1 and 2, after T epochs for sufficiently big enough T, with $\eta_t = \frac{\eta}{N/B}$, and $\eta \leq \frac{1}{2L}$, we have with probability $1-\epsilon$ that:*

$$\frac{1}{T}\sum_{i=0}^{T}\left\|\nabla F(\tilde{w}_t)\right\|^2 \leq \frac{(F(\tilde{w}_0) - F^*)^{2/3}L^{2/3}\sigma^{2/3}(\ln(T) + \ln(1/\epsilon))^{1/3}}{N^{1/3}T^{2/3}} \tag{7}$$

During the proof of Theorem 6, we encountered a term of the form:

$$S_1 = \left\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left(\nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) - \nabla F(w_0^{(t)})\right)\right\|^2$$

which was bounded as:

$$\left\| \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \left(\nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) - \nabla F(w_0^{(t)}) \right) \right\|^2 \leq \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \left\| \nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) - \nabla F(w_0^{(t)}) \right\|^2$$

$$\leq \frac{N}{iB} \left(\Theta \left\| \nabla F(w_0^{(t)}) \right\|^2 + \sigma^2 \right)$$

The first inequality is an application of Cauchy Schwartz's inequality (10) and the second one comes from Assumption 3. We noticed that the second inequality is weak. Inspired by [12], we used a similar way to bound S with probability $1-\delta$, as follows:

$$S_1 \leq \frac{4 \ln(2/\delta) N \left(\Theta \left\| \nabla F(w_0^{(t)}) \right\|^2 + \sigma^2 \right)}{(iB)^2}$$

which is an improvement from the previous bound for S. Following the same path of the proof from Theorem 6 we get the results for theorem 7.

Below we present our final result, the novel result for the strongly convex case. This time, the term we optimize is

$$S_2 = \sum_{k=i}^{\frac{N}{B}} \sum_{j=1}^{B} \left\| \nabla f(w^*; \sigma(iB+j)) \right\|^2$$

which was bounded as

$$\sum_{k=i}^{\frac{N}{B}} \sum_{j=1}^{B} \left\| \nabla f(w^*; \sigma(iB+j)) \right\|^2 \leq N\sigma_*^2$$

in Theorem 4, but with our high probability lemma we have w.p.$1-\delta$:

$$\sum_{k=i}^{\frac{N}{B}} \sum_{j=1}^{B} \left\| \nabla f(w^*; \sigma(iB+j)) \right\|^2 \leq \frac{4 \ln(2/\delta) N}{iB} \sigma_*^2$$

which is also an improvement.

Theorem 8 (Strongly-convex)**.** *We have that under assumptions 1 and 3, after T epochs for sufficiently big enough T, and $\eta_t = \frac{6 \log(T)}{\mu T} \frac{B}{N}$ we have that with probability $1-\epsilon$*

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2 \left(2\mu^2 + 12L^2 \ln(\frac{2NT}{B\epsilon}) \right) \log^2(T)}{NT^2\mu^3} \tag{8}$$

3 Conclusion

In this paper we studied the convergence guarantees of mini-batch stochastic gradient descent, in the generic shuffling and random reshuffling regime. We have two novel results for the random reshuffling technique, however, they are not state of the art.

While our project started with trying to investigate the effect of the batch size on the convergence of the algorithm, we did not succeed in mathematically showing such a trade-off. The fundamental reason as to why we did not achieve that, is the fact that the quantity:

$$S_1 = \left\| \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \left(\nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) - \nabla F(w_0^{(t)}) \right) \right\|^2$$

is extremely hard to bound if we know nothing about the permutation $\sigma^{(t)}$. This is why we were able to achieve something for Random Reshuffling, where each epoch we assume we get a fresh/random permutation, but in the generic regime, we are not able to make such assumptions.

References

[1] Leonid D Akulenko and Sergei V Nesterov. *High-precision methods in eigenvalue problems and their applications*. CRC Press, 2004.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[3] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[4] M Gori and A Tesi. Some examples of local minima during learning with backpropagation. In *Third Italian Workshop on Parallel Architectures and Neural Networks*, pages 87–94. World Scientific, 1990.

[5] Hiroyuki Kasai. Sgdlibrary: A matlab library for stochastic optimization algorithms. *The Journal of Machine Learning Research*, 18(1):7942–7946, 2017.

[6] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[7] Angelia Nedic and Dimitri P Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[8] Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten Van Dijk. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

[9] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[10] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[11] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

[12] Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local sgd with shuffling: Tight convergence bounds and beyond. *arXiv preprint arXiv:2110.10342*, 2021.

4 Appendix

Here we include proofs of Theorems 4,5,6,7,8.

4.1 Useful Lemmas

Lemma 9. *We have that for vectors $u, v \in \mathbb{R}^d$:*

$$u^T v = \frac{1}{2}\left(\left\|u\right\|^2 + \left\|v\right\|^2 - \left\|u - v\right\|^2\right) \tag{9}$$

Lemma 10 (Cauchy-Schwartz)**.** *We have that for vectors $u \in \mathbb{R}^d$:*

$$u^T v \leq \left\|u\right\|\left\|v\right\| \tag{10}$$

Lemma 11 (Lemma 1 from [8])**.** *Let $\{Y_t\}_{t\geq 1}$ be a non-negative sequence in $\mathbb{R}$ and q be a positive integer number. Let $\rho > 0$ and $D > 0$ be two given constants and $0 < \eta t \leq \frac{1}{\rho}$ be given for all $t \geq 1$ and assume that, for all $t \geq 1$, we have*

$$Y_{t+1} \leq (1 - \rho\eta_t)Y_t + D\eta^{q+1}$$

then choosing $\eta_t \in (0, \frac{1}{\rho}), \forall t \geq 1$ we get:

$$Y_{t+1} \leq (1-\rho\eta)^t Y_1 + \frac{D\eta^q[1-(1-\rho\eta)^t)]}{\rho} \leq Y_1 exp(-\rho\eta t) + \frac{D\eta^q}{\rho}$$

4.2 Strongly-Convex Results (Theorem 4)

Lemma 12. *We have that*

$$I = \sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \leq 8\eta_t^2\frac{N^2}{B^2}L^2\left(\left\|w_0^{(t)} - w^*\right\|^2 + \eta_t^2\frac{N^2}{B^2}\sigma_*^2\right) + \eta_t^2\frac{N^3}{B^3}\sigma_*^2 \tag{11}$$

$$\tag{12}$$

Proof. First we try to bound the quantity $\left\|w_i^{(t)} - w_0^{(t)}\right\|^2$. We have that:

$$\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 = \left\|\frac{\eta_t}{B}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(iB+j))\right\|^2 \tag{13}$$

$$\overset{(a)}{=} \frac{\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left(\nabla f(w_k^{(t)}; \sigma(iB+j)) - \nabla f(w^*; \sigma(iB+j))\right) - \sum_{k=i}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w^*; \sigma(iB+j))\right\|^2 \tag{14}$$

$$\overset{(b)}{\leq} \frac{2\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left(\nabla f(w_k^{(t)}; \sigma(iB+j)) - \nabla f(w^*; \sigma(iB+j))\right)\right\|^2 + \tag{15}$$

$$+ \frac{2\eta_t^2}{B^2}\left\|\sum_{k=i}^{\frac{N}{B}}\sum_{j=1}^{B}\nabla f(w^*; \sigma(iB+j))\right\|^2 \tag{16}$$

$$\overset{(c)}{\leq} \frac{2\eta_t^2 iB}{B^2}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)}; \sigma(iB+j)) - \nabla f(w^*; \sigma(iB+j))\right\|^2 + \tag{17}$$

$$+ \frac{2\eta_t^2(N-iB)}{B^2}\sum_{k=i}^{\frac{N}{B}}\sum_{j=1}^{B}\left\|\nabla f(w^*; \sigma(iB+j))\right\|^2 \tag{18}$$

$$\overset{(d)}{\leq} \frac{2\eta_t^2 iB^2L^2}{B^2}\sum_{k=0}^{i-1}\left\|w_k^{(t)} - w^*\right\|^2 + \frac{2\eta_t^2(N-iB)}{B^2}N\sigma_*^2 \tag{19}$$

$$= 2\eta_t^2\left(iL^2\sum_{k=0}^{i-1}\left\|w_k^{(t)} - w^*\right\|^2 + \frac{(N-iB)N}{B^2}\sigma_*^2\right) \tag{20}$$

where above we used the following inequalities:

1. (a) We use the fact that $\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w^*;\sigma(iB+j))=0$
2. (b) $\|a+b\|\leq 2\|a\|^2+2\|b\|^2$
3. (c) Jensen's inequality
4. (d) The Lipschitz condition plus the fact that:

$$\sum_{k=i}^{\frac{N}{B}}\sum_{j=1}^{B}\left\|\nabla f(w^*;\sigma(iB+j))\right\|^2\leq\sum_{k=0}^{\frac{N}{B}}\sum_{j=1}^{B}\left\|\nabla f(w^*;\sigma(iB+j))\right\|^2=\sigma_*^2$$

Now, note that because of 20:

$$\left\|w_i^{(t)}-w^*\right\|^2=\left\|\left(w_i^{(t)}-w_0^{(t)}\right)+\left(w_0^{(t)}-w^*\right)\right\|^2\leq 2\left\|w_i^{(t)}-w_0^{(t)}\right\|^2+2\left\|w_0^{(t)}-w^*\right\|^2 \tag{21}$$

$$\leq 4\eta_t^2\left(iL^2\sum_{k=0}^{i-1}\left\|w_k^{(t)}-w^*\right\|^2+\frac{(N-iB)N}{B^2}\sigma_*^2\right)+2\left\|w_0^{(t)}-w^*\right\|^2 \tag{22}$$

Summing up the above we get:

$$\sum_{j=0}^{i-1}\left\|w_j^{(t)}-w^*\right\|^2\leq 2i\left\|w_0^{(t)}-w^*\right\|^2+4\eta_t^2\sum_{j=0}^{i-1}\left(jL^2\sum_{k=0}^{j-1}\left\|w_k^{(t)}-w^*\right\|^2+\frac{(N-jB)N}{B^2}\sigma_*^2\right) \tag{23}$$

$$=2i\left\|w_0^{(t)}-w^*\right\|^2+4\eta_t^2L^2\sum_{j=0}^{i-1}j\sum_{k=0}^{j-1}\left\|w_k^{(t)}-w^*\right\|^2+4\eta_t^2\frac{N}{B^2}\sigma_*^2\left(iN-\frac{Bi^2}{2}\right) \tag{24}$$

$$=2i\left\|w_0^{(t)}-w^*\right\|^2+2\eta_t^2L^2\sum_{j=0}^{i-1}(i^2-j^2)\left\|w_j^{(t)}-w^*\right\|^2+4\eta_t^2\frac{N}{B^2}\sigma_*^2\left(iN-\frac{Bi^2}{2}\right) \tag{25}$$

$$\leq 2i\left\|w_0^{(t)}-w^*\right\|^2+2\eta_t^2L^2\frac{N^2}{B^2}\sum_{j=0}^{i-1}\left\|w_j^{(t)}-w^*\right\|^2+2\eta_t^2\frac{N^2i}{B^2}\sigma_*^2 \tag{26}$$

$$=2i\left\|w_0^{(t)}-w^*\right\|^2+2\eta_t^2\frac{N^2}{B^2}L^2\sum_{j=0}^{i-1}\left\|w_j(t)-w^*\right\|^2+2\eta_t^2\frac{N^2}{B^2}i\sigma_*^2 \tag{27}$$

Note that above we used that we set η_t such that $1-2\eta_t^2L^2\frac{N^2}{B^2}\geq\frac{1}{2}$, so we get that:

$$\sum_{j=0}^{i-1}\left\|w_j^{(t)}-w^*\right\|^2\leq 4i\left\|w_0^{(t)}-w^*\right\|^2+4i\eta_t^2\sigma_*^2\frac{N^2}{B^2} \tag{28}$$

Finally we have:

$$\sum_{i=1}^{\frac{N}{B}}\left\|w_i^{(t)}-w_0^{(t)}\right\|^2\leq 2\eta_t^2\left(L^2\sum_{i=1}^{\frac{N}{B}}\sum_{k=0}^{i-1}\left\|w_k^{(t)}-w^*\right\|^2+\frac{N}{B^2}\sigma_*^2\sum_{i=1}^{\frac{N}{B}}(N-iB)\right) \tag{29}$$

$$\leq 2\eta_t^2\left(L^2\sum_{i=1}^{\frac{N}{B}}\sum_{k=0}^{i-1}\left\|w_k^{(t)}-w^*\right\|^2+\frac{N}{B^2}\sigma_*^2\left(\frac{N^2}{B}-B\frac{N^2}{2B^2}\right)\right) \tag{30}$$

$$=2\eta_t^2\left(L^2\sum_{i=1}^{\frac{N}{B}}\sum_{k=0}^{i-1}\left\|w_k^{(t)}-w^*\right\|^2+\frac{N^3}{2B^3}\sigma_*^2\right) \tag{31}$$

$$\leq 2\eta_t^2\left(L^2\sum_{i=1}^{\frac{N}{B}}\left(4i\left\|w_0^{(t)}-w^*\right\|^2+4i\eta_t^2\frac{N^2}{B^2}\sigma_*^2\right)+\frac{N^3}{2B^3}\sigma_*^2\right) \tag{32}$$

$$\leq 4\eta_t^2L^2\frac{N^2}{B^2}\left(\left\|w_0^{(t)}-w^*\right\|^2+\eta_t^2\frac{N^2}{B^2}\sigma_*^2\right)+\eta_t^2\frac{N^3}{B^3}\sigma_*^2 \tag{33}$$

$$\tag{34}$$

□

Lemma 13. *We have that*

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\hat{\eta}_t BL^2}{2N}\sum_{i=0}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \tag{35}$$

where $\hat{\eta}_t = \frac{B}{N}\eta_t$

Proof. We have

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \langle \nabla F(\tilde{w}_t), \tilde{w}_{t+1} - \tilde{w}_t\rangle + \frac{L}{2}||\tilde{w}_{t+1} - \tilde{w}_t||^2 \tag{36}$$

$$= F(\tilde{w}_t) - \frac{\eta_t N}{B}\left\langle \nabla F(\tilde{w}_t), \frac{1}{N}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\rangle + \frac{L}{2}||\tilde{w}_{t+1} - \tilde{w}_t||^2 \tag{37}$$

$$= F(\tilde{w}_t) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 - \frac{\eta_t N}{2B}\left\|\frac{1}{N}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 + \tag{38}$$

$$+ \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 + \frac{L\eta_t^2}{2B^2}\left\|\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 \tag{39}$$

$$= F(\tilde{w}_t) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 - \tag{40}$$

$$- \frac{\eta_t}{2NB}\left(1 - \frac{N}{B}L\eta_t\right)\left\|\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 \tag{41}$$

$$\leq F(\tilde{w}_t) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\eta_t N}{2BN^2}\left\|\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left(\nabla f(\tilde{w}_t; \sigma(iB+j)) - \nabla f(w_i^{(t)}; \sigma(iB+j))\right)\right\|^2 \tag{42}$$

$$\leq F(\tilde{w}_t) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\eta_t}{2B}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(\tilde{w}_t; \sigma(iB+j)) - \nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 \tag{43}$$

$$= F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\hat{\eta}_t}{2N}\sum_{i=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(\tilde{w}_t; \sigma(iB+j)) - \nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 \tag{44}$$

$$\leq F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\hat{\eta}_t BL^2}{2N}\sum_{i=0}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \tag{45}$$

$$\tag{46}$$

□

Proof of Theorem 4. We use Lemma 8 and Lemma 7 and we have:

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\hat{\eta}_t BL^2}{2N}\sum_{i=0}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \tag{47}$$

$$\leq F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{\hat{\eta}_t BL^2}{2N}\left(8\hat{\eta}_t^2 L^2\left(\left\|w_0^{(t)} - w^*\right\|^2 + \hat{\eta}_t^2\sigma_*^2\right) + \hat{\eta}_t^2\frac{N}{B}\sigma_*^2\right) \tag{48}$$

$$= F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{4\hat{\eta}_t^3 BL^4}{N}\left\|w_0^{(t)} - w^*\right\|^2 + \frac{4\hat{\eta}_t^5\sigma_*^2 BL^3}{N} + \frac{\hat{\eta}_t^3 L^2\sigma_*^2}{2} \tag{49}$$

$$\leq F(\tilde{w}_t) - \frac{\hat{\eta}_t}{2}2\mu\left[F(\tilde{w}_t) - F(w_*)\right] + \frac{4\hat{\eta}_t^3 BL^4}{N}\frac{2}{\mu}\left[F(\tilde{w}_t) - F(w_*)\right] + \frac{4\hat{\eta}_t^5\sigma_*^2 BL^4}{N} + \frac{\hat{\eta}_t^3 L^2\sigma_*^2}{2} \tag{50}$$

$$\tag{51}$$

Subtracting $F(w_*)$ from both sides we have:

$$F(\tilde{w}_{t+1}) - F(w_*) \leq \left(1 - \hat{\eta}_t\left(\mu - \frac{8\hat{\eta}_t^2 BL^4}{N\mu}\right)\right)\left[F(\tilde{w}_t) - F(w_*)\right] + \hat{\eta}_t^3\left(\frac{4\hat{\eta}_t^2\sigma_*^2 BL^4}{N} + \frac{L^2\sigma_*^2}{2}\right) \tag{52}$$

$$\tag{53}$$

Note that we already assume $\hat{\eta}_t \leq \frac{1}{2L}$. We want to have

$$\mu - \frac{8\hat{\eta}_t^2 BL^4}{N\mu} \geq \frac{\mu}{3} \implies \frac{2\mu}{3} \geq \frac{8\hat{\eta}_t^2 BL^4}{N\mu} \implies \hat{\eta}_t \leq \frac{\mu}{L^2}\sqrt{\frac{N}{12B}} \tag{54}$$

The above also implies

$$\frac{4\hat{\eta}_t^2\sigma_*^2 BL^4}{N} \leq \frac{\mu^2\sigma_*^2}{3} \tag{55}$$

Combining the above gives:

$$F(\tilde{w}_{t+1}) - F(w_*) \leq \left(1 - \frac{\mu\hat{\eta}_t}{3}\right)[F(\tilde{w}_t) - F(w_*)] + \frac{1}{6}\hat{\eta}_t^3\sigma_*^2\left(2\mu^2 + 3L^2\right) \tag{56}$$

$$\tag{57}$$

The above can be written as follows:

$$Y_{T+1} \leq (1 - \rho\hat{\eta}_t)\, Y_T + \hat{\eta}_t^3 D \tag{58}$$

Using Lemma 11 we have that if $\hat{\eta}_t \leq \frac{1}{\rho}$:

$$Y_{T+1} \leq (1 - \rho\hat{\eta}_t)\, Y_T + \hat{\eta}_t^3 D \leq Y_1 exp(-\rho\hat{\eta}_t T) + \frac{D\hat{\eta}_t^2}{\rho} \tag{59}$$

Choosing $\hat{\eta}_t = \frac{2\log(T)}{\rho T}$, we have:

$$Y_{T+1} \leq \frac{Y_1}{T^2} + \frac{4D\log^2(T)}{T^2\rho^3} = \frac{Y_1}{T^2} + \frac{4\frac{1}{6}\sigma_*^2\left(2\mu^2 + 3L^2\right)\log^2(T)}{T^2\frac{\mu^3}{3^3}} \tag{60}$$

Which finally gives:

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2\left(2\mu^2 + 3L^2\right)\log^2(T)}{T^2\mu^3} \tag{61}$$

We choose T such that

$$\hat{\eta}_t = \frac{6\log T}{\mu T} \leq \min\{\frac{1}{2L}, \frac{\mu}{L^2}\sqrt{\frac{N}{12B}}\} \implies \frac{\log(T)}{T} \leq \frac{\mu^2}{6L^2}\sqrt{\frac{N}{12B}} \tag{62}$$

□

Lemma 14 (Similar to 12)**.** *If we apply our theorem with probability $1-\delta$ we have*

$$\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \leq 2\eta_t^2\left(iL^2\sum_{k=0}^{i-1}\left\|w_k^{(t)} - w^*\right\|^2 + \frac{\ln(2/\delta)N}{i^2B^2}\sigma_*^2\right) \tag{63}$$

4.2.1 Results with high probability bounds (Theorem 8)

Lemma 15.

$$I = \sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \leq 8\eta_t^2\frac{N^2}{B^2}L^2\left(\left\|w_0^{(t)} - w^*\right\|^2 + \eta_t^2\frac{N^2}{B^2}\sigma_*^2\right) + \eta_t^2\frac{8N^2\ln(2/\delta)}{B^3}\sigma_*^2 \tag{64}$$

$$\tag{65}$$

$$F(\tilde{w}_{t+1}) - F(w_*) \leq \left(1 - \hat{\eta}_t\left(\mu - \frac{8\hat{\eta}_t^2 BL^4}{N\mu}\right)\right)[F(\tilde{w}_t) - F(w_*)] + \hat{\eta}_t^3\left(\frac{4\hat{\eta}_t^2\sigma_*^2 BL^4}{N} + \frac{4\ln(2/\delta)L^2\sigma_*^2}{N}\right) \tag{66}$$

$$\tag{67}$$

Previous bound:

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2\left(2\mu^2 + 3L^2\right)\log^2(T)}{T^2\mu^3} \tag{68}$$

New bound with probability $(1-\delta)^{NT/B}$

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2\left(2\mu^2 + 12L^2\ln(2/\delta)\right)\log^2(T)}{NT^2\mu^3} \tag{69}$$

If indeed we want probability at least $1-\epsilon$, we want: $\delta = \frac{\epsilon B}{NT}$. We have with probability at least $1-\epsilon$:

$$F(\tilde{w}_T) - F(w_*) \leq \frac{F(\tilde{w}_0) - F(w_*)}{T^2} + \frac{18\sigma_*^2\left(2\mu^2 + 12L^2\ln(\frac{2NT}{B\epsilon})\right)\log^2(T)}{NT^2\mu^3} \tag{70}$$

4.3 Non Strongly-Convex Results (Theorem 5)

Lemma 16 (Lemma 3 from [8]). *Let $X_1, \dots, X_n$ be n given vectors in $\mathbb{R}^d$, $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ be their average, and $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \|X_i - \bar{X}\|^2$ be their population variance. Fix any $k \in \{1, \dots, n\}$, let $X_{\pi_1}, \dots, X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, \dots, X_n\}$ and $\bar{X}_\pi := \frac{1}{k}\sum_{i=1}^k X_{\pi_k}$ be their average. Then, we have:*

$$\mathbb{E}[\bar{X}_\pi] = \bar{X}, \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-k}{k(n-1)}\sigma^2$$

Lemma 17 (Similar to Lemma 7 in [8]). *Under Assumption 1, the following holds:*

$$\left\|\tilde{w}_t - w_*\right\|^2 \leq \left\|\tilde{w}_{t-1} - w_*\right\|^2 - 2\eta\Big[F(\tilde{w}_{t-1}) - F(w_*)\Big] + \frac{L\eta^3}{3BN}\sigma_*^2 \tag{71}$$

Proof. We have that:

$$\begin{aligned}
\left\|w_i^{(t)} - \tilde{w}_{t-1}\right\|^2 &= \left\|\frac{\eta_t}{B}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(kB+j))\right\|^2 \\
&= \frac{\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\Big(\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\Big) + \nabla f(w_*; \sigma(kB+j))\right\|^2 \\
&\leq \frac{2\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\Big(\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\Big)\right\|^2 + \frac{2\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*; \sigma(kB+j))\right\|^2 \\
&\leq \frac{2\eta_t^2 iB}{B^2}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\right\|^2 + \frac{2\eta_t^2}{B^2}\left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*; \sigma(kB+j))\right\|^2
\end{aligned}$$

Let us denote with $B_i^* := \left\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*; \sigma(kB+j))\right\|^2$. Summing the above from $i = 1$ to $\frac{N}{B} - 1$:

$$\begin{aligned}
\sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)} - \tilde{w}_{t-1}\right\|^2 &\leq \frac{2\eta_t^2}{B}\sum_{i=1}^{\frac{N}{B}-1} i\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\right\|^2 + \frac{2\eta_t^2}{B^2}\sum_{i=1}^{\frac{N}{B}-1} B_i^* \qquad (72) \\
&\leq \frac{2\eta_t^2 N^2}{B^3}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\right\|^2 + \frac{2\eta_t^2}{B^2}B_* \qquad (73)
\end{aligned}$$

Where $B_* = \sum_{i=1}^{\frac{N}{B}-1} B_i^*$. Now we have:

$$\tilde{w}_t - w_* = \tilde{w}_{t-1} - w_* - \frac{\eta_t}{B}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(kB+j))$$

$$\begin{aligned}
\left\|\tilde{w}_t - w_*\right\|^2 &= \left\|\tilde{w}_{t-1} - w_*\right\|^2 + \frac{\eta_t^2}{B^2}\left\|\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(kB+j))\right\|^2 - \frac{2\eta_t}{B}\left\langle\tilde{w}_{t-1} - w_*, \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(kB+j))\right\rangle \\
&= \left\|\tilde{w}_{t-1} - w_*\right\|^2 + \frac{\eta_t^2}{B^2}\left\|\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big(\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\Big)\right\|^2 - \\
&- \frac{2\eta_t}{B}\left\langle\tilde{w}_{t-1} - w_*, \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)}; \sigma(kB+j))\right\rangle \\
&\leq \left\|\tilde{w}_{t-1} - w_*\right\|^2 + \frac{\eta_t^2 N}{B^2}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)}; \sigma(kB+j)) - \nabla f(w_*; \sigma(kB+j))\right\|^2 - \\
&- \frac{2\eta_t}{B}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\langle\tilde{w}_{t-1} - w_*, \nabla f(w_k^{(t)}; \sigma(kB+j))\right\rangle
\end{aligned}$$

Now let us take a better look at the term:

$$
\begin{aligned}
T &= \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\langle w_* - \tilde{w}_{t-1}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle \\
&= \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\langle w_* - w_k^{(t)}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle + \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\langle w_k^{(t)} - \tilde{w}_{t-1}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle \\
&\overset{(a)}{\leq} \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big[f(w_k^{(t)};\sigma(kB+j)) - f(\tilde{w}_{t-1};\sigma(kB+j))\Big] + \frac{L}{2}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|w_k^{(t)} - \tilde{w}_{t-1}\Big\|^2 + \\
&+ \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\langle w_* - w_k^{(t)}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle \\
&= -\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big[f(w_*;\sigma(kB+j)) - f(w_k^{(t)};\sigma(kB+j)) - \Big\langle w_* - w_k^{(t)}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle\Big] \\
&+ \sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big[f(w_*;\sigma(kB+j)) - f(\tilde{w}_{t-1};\sigma(kB+j))\Big] + \frac{L}{2}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|w_k^{(t)} - \tilde{w}_{t-1}\Big\|^2 \\
&\overset{(b)}{\leq} -\frac{1}{2L}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|\nabla f(w_*;\sigma(kB+j) - \nabla f(w_k^{(t)};\sigma(kB+j)\Big\|^2 + N\Big[F(w_*) - F(\tilde{w}_{t-1})\Big] + \frac{L}{2}\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|w_k^{(t)} - \tilde{w}_{t-1}\Big\|^2 \\
&\overset{(c)}{\leq} \left(\frac{\eta_t^2 N^2 L}{B^2} - \frac{1}{2L}\right)\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|\nabla f(w_*;\sigma(kB+j) - \nabla f(w_k^{(t)};\sigma(kB+j)\Big\|^2 + N\Big[F(w_*) - F(\tilde{w}_{t-1})\Big] + \frac{L\eta_t^2}{B}B_*
\end{aligned}
$$

The above hold because:

1. (a) Due to the convexity assumption

$$\Big\langle w_k^{(t)} - \tilde{w}_{t-1}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle \leq f(w_k^{(t)};\sigma(kB+j)) - f(\tilde{w}_{t-1};\sigma(kB+j)) + \frac{L}{2}\Big\|w_k^{(t)} - \tilde{w}_{t-1}\Big\|^2$$

2. (b) Due to the convexity assumption

$$f(w_*;\sigma(kB+j)) - f(w_k^{(t)};\sigma(kB+j)) - \Big\langle w_* - w_k^{(t)}, \nabla f(w_k^{(t)};\sigma(kB+j))\Big\rangle \geq$$

$$\geq \frac{1}{2L}\Big\|\nabla f(w_*;\sigma(kB+j) - \nabla f(w_k^{(t)};\sigma(kB+j)\Big\|^2$$

3. (c) This is due to 72

Substituting back to 74 we have:

$$\Big\|\tilde{w}_t - w_*\Big\|^2 \leq \Big\|\tilde{w}_{t-1} - w_*\Big\|^2 + \frac{2\eta_t}{B}\left(N\Big[F(w_*) - F(\tilde{w}_{t-1})\Big] + \frac{L\eta_t^2}{B}B_*\right) + \tag{74}$$

$$+ \left(-\frac{\eta_t}{BL} + \frac{2\eta_t^3 N^2 L}{B^3} + \frac{\eta_t^2 N}{B^3}\right)\sum_{k=0}^{\frac{N}{B}-1}\sum_{j=1}^{B}\Big\|\nabla f(w_*;\sigma(kB+j) - \nabla f(w_k^{(t)};\sigma(kB+j)\Big\|^2 \tag{75}$$

Choosing $\eta_t = \frac{\eta}{\frac{N}{B}}$ such that

$$-\frac{\eta_t}{BL} + \frac{2\eta_t^3 N^2 L}{B^3} + \frac{\eta_t^2 N}{B^2} \leq 0 \implies -\frac{\eta}{LB} + \frac{2\eta^3 L}{N} + \frac{\eta^2}{N} \leq 0$$

$$\eta^2 + \frac{\eta}{2L} - \frac{N}{2BL^2} \leq 0 \implies (\eta - \frac{1}{4L})^2 \leq \frac{\frac{8N}{B}+1}{16L^2} \implies \eta \leq \frac{1+\sqrt{8N/B+1}}{4L}$$

then the above becomes:

$$\left\|\tilde{w}_t - w_*\right\|^2 \leq \left\|\tilde{w}_{t-1} - w_*\right\|^2 + \frac{2\eta_t N}{B}\Big[F(w_*) - F(\tilde{w}_{t-1})\Big] + \frac{2L\eta_t^3}{B^2}B_*$$

Finally, to bound B_* we use 16 in the following way:

$$\mathbb{E}[B_*] = \mathbb{E}\left[\sum_{i=1}^{\frac{N}{B}-1} B_i^*\right] = \mathbb{E}\left[\sum_{i=1}^{\frac{N}{B}-1}\Big\|\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*;\sigma(kB+j))\Big\|^2\right] \tag{76}$$

$$= \sum_{i=1}^{\frac{N}{B}-1}(iB)^2\mathbb{E}\left[\Big\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*;\sigma(kB+j))\Big\|^2\right] \tag{77}$$

$$= \sum_{i=1}^{\frac{N}{B}-1}(iB)^2\mathbb{E}\left[\Big\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_*;\sigma(kB+j)) - \nabla F(w_*)\Big\|^2\right] \tag{78}$$

$$= \sum_{i=1}^{\frac{N}{B}-1}(iB)^2\frac{N-iB}{(iB)(N-1)}\sigma_*^2 \leq \frac{\sigma_*^2}{N}\sum_{i=1}^{\frac{N}{B}-1} iB(N-iB) \tag{79}$$

$$\leq \frac{\sigma_*^2 B}{N}\left(\frac{N\frac{N^2}{B^2}}{2} - B\frac{\frac{N^3}{B^3}}{3}\right) = \frac{\sigma_*^2 B}{N}\left(\frac{N^3}{2B^2} - \frac{N^3}{3B^2}\right) = \frac{N^2}{6B}\sigma_*^2 \tag{80}$$

Finally we have:

$$\left\|\tilde{w}_t - w_*\right\|^2 \leq \left\|\tilde{w}_{t-1} - w_*\right\|^2 + \frac{2\eta_t N}{B}\Big[F(w_*) - F(\tilde{w}_{t-1})\Big] + \frac{2L\eta_t^3}{B^2}\frac{N^2}{6B}\sigma_*^2 \tag{81}$$

$$= \left\|\tilde{w}_{t-1} - w_*\right\|^2 - 2\eta\Big[F(\tilde{w}_{t-1}) - F(w_*)\Big] + \frac{L\eta^3}{3N}\sigma_*^2 \tag{82}$$

□

Theorem 5. We have from 17 that:

$$F(\tilde{w}_{t-1}) - F(w_*) \leq \frac{1}{2\eta}\left(\left\|\tilde{w}_{t-1} - w_*\right\|^2 - \left\|\tilde{w}_t - w_*\right\|^2\right) + \frac{L\eta^2}{6N}\sigma_*^2 \tag{83}$$

Summing up gives:

$$\sum_{t=0}^{T}[F(\tilde{w}_t) - F(w_*)] \leq \frac{1}{2\eta}\left(\left\|\tilde{w}_0 - w_*\right\|^2 - \left\|\tilde{w}_{T+1} - w_*\right\|^2\right) + \frac{LT\eta^2}{6N}\sigma_*^2 \implies$$

$$\frac{1}{T}\sum_{t=0}^{T}[F(\tilde{w}_{t-1}) - F(w_*)] \leq \frac{1}{2\eta T}\left\|\tilde{w}_0 - w_*\right\|^2 + \frac{L\eta^2}{6N}\sigma_*^2$$

□

4.4 Non-Convex Results (Theorem 6)

Lemma 18. *We have that*

$$I = \sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)} - w_0^{(t)}\right\|^2 \leq \frac{N^2\eta_t^2}{B^2}\Big(\Big(3\Theta + 2\Big)\Big\|\nabla F(w_0^{(t)})\Big\|^2 + \sigma^2\Big) \tag{84}$$

Proof. We have:

$$\begin{aligned}
\Big\|w_i^{(t)} - w_0^{(t)}\Big\|^2 &= \eta_t^2 \Big\| \frac{1}{B} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \nabla f(w_k^{(t)}; \sigma^{(t)}(kB+j)) \Big\|^2 \\
&= i^2 \cdot \eta_t^2 \Big\| \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \nabla f(w_k^{(t)}; \sigma^{(t)}(kB+j)) \Big\|^2 \\
&\overset{(a)}{\leq} 3i^2\eta_t^2 \Big\| \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \Big(\nabla f(w_k^{(t)}; \sigma^{(t)}(kB+j)) - \nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) \Big) \Big\|^2 + \\
&+ 3i^2\eta_t^2 \Big(\Big\| \frac{1}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{b} \Big(\nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) - \nabla F(w_0^{(t)}) \Big) \Big\|^2 + \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big) \\
&\overset{(b)}{\leq} \frac{3i^2\eta_t^2}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} \Big\| \nabla f(w_k^{(t)}; \sigma^{(t)}(kB+j)) - \nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) \Big\|^2 + \\
&+ 3i^2\eta_t^2 \Big(\frac{N}{iB} \Big(\Theta \|\nabla F(w_0^{(t)})\|^2 + \sigma^2 \Big) + \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big) \\
&\overset{(c)}{\leq} \frac{3i^2\eta_t^2}{iB} \sum_{k=0}^{i-1} \sum_{j=1}^{B} L^2 \Big\| w_k^{(t)} - w_0^{(t)} \Big\|^2 + 3i^2\eta_t^2 \Big(\frac{N}{iB} \Big(\Theta \|\nabla F(w_0^{(t)})\|^2 + \sigma^2 \Big) + \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big) \\
&= \frac{3i^2\eta_t^2 L^2}{i} \sum_{k=0}^{i-1} \Big\| w_k^{(t)} - w_0^{(t)} \Big\|^2 + 3i^2\eta_t^2 \Big(\frac{N}{iB} \Big(\Theta \|\nabla F(w_0^{(t)})\|^2 + \sigma^2 \Big) + \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big) \\
&\overset{(d)}{\leq} 3i\eta_t^2 L^2 I + 3\eta_t^2 \Big(\frac{Ni}{B} \Big(\Theta \|\nabla F(w_0^{(t)})\|^2 + \sigma^2 \Big) + i^2 \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big)
\end{aligned}$$

The above holds because of the following:

1. (a) $(a+b+c)^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$
2. (b) If we denote by $x_j = \nabla f(w_0^{(t)}; \sigma(j)) - \nabla F(w_0^{(t)})$ we have that:

$$\Big\| \frac{1}{iB} \sum_{j=1}^{iB} x_j \Big\|^2 \leq \frac{1}{iB} \sum_{j=0}^{iB-1} \Big\| x_j \Big\|^2 < \frac{1}{iB} \sum_{j=1}^{N} \Big\| x_j \Big\|^2 \leq \frac{N}{iB} \left(\Theta \Big\| \nabla F(w_0^{(t)}) \Big\|^2 + \sigma^2 \right)$$

3. (c)

$$\Big\| \nabla f(w_k^{(t)}; \sigma^{(t)}(kB+j)) - \nabla f(w_0^{(t)}; \sigma^{(t)}(kB+j)) \Big\| \leq L \Big\| w_k^{(t)} - w_0^{(t)} \Big\|$$

4. (d)

$$I = \sum_{i=1}^{\frac{N}{B}-1} \Big\| w_i^{(t)} - w_0^{(t)} \Big\|^2 \geq \sum_{i=1}^{i-1} \Big\| w_i^{(t)} - w_0^{(t)} \Big\|^2, \forall i \leq \frac{N}{B}$$

We can now finish off in the following way:

$$\begin{aligned}
I = \sum_{i=1}^{\frac{N}{B}-1} \Big\| w_i^{(t)} - w_0^{(t)} \Big\|^2 &\leq \sum_{i=1}^{\frac{N}{B}-1} \left(3i\eta_t^2 L^2 I + 3\eta_t^2 \Big(\frac{Ni}{B} \Big(\Theta \Big\| \nabla F(w_0^{(t)}) \Big\|^2 + \sigma^2 \Big) + i^2 \Big\| \nabla F(w_0^{(t)}) \Big\|^2 \Big) \right) \\
&\leq \frac{3\eta_t^2 L^2 I N^2}{2B^2} + \frac{3N^3\eta_t^2}{2B^3} \Big(\Theta \Big\| \nabla F(w_0^{(t)}) \Big\|^2 + \sigma^2 \Big) + \frac{N^3\eta_t^2}{B^3} \Big\| \nabla F(w_0^{(t)}) \Big\|^2
\end{aligned}$$

Now choosing η_t such that:

$$\frac{3\eta_t^2 L^2 I N^2}{B^2} \leq \frac{I}{2} \iff \eta_t \leq \frac{B}{NL} \sqrt{\frac{1}{3}} \tag{85}$$

The above becomes

$$I \leq \frac{N^2\eta_t^2}{B^2} \Big(\Big(3\Theta + 2 \Big) \Big\| \nabla F(w_0^{(t)}) \Big\|^2 + \sigma^2 \Big)$$

□

Lemma 19. *We have that:*

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{3}{2}\eta^3 L^2\sigma^2 \tag{86}$$

Proof. Here we assume that $\tilde{w}_t = w_0^{(t)}$. We have:

$$\begin{aligned}
F(\tilde{w}_{t+1}) &\leq F(\tilde{w}_t) - \langle \nabla F(\tilde{w}_t), \tilde{w}_{t+1} - \tilde{w}_t\rangle + \frac{L}{2}\left\|\tilde{w}_{t+1} - \tilde{w}_t\right\|^2 \\
&\leq F(\tilde{w}_t) - \left\langle \nabla F(\tilde{w}_t), \frac{\eta_t}{B}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\rangle + \frac{L\eta_t^2}{2B^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&= F(\tilde{w}_t) + \frac{L\eta_t^2}{2B^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&- \frac{\eta_t N}{2B}\left\|\frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 + \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&= F(\tilde{w}_t) + \left(\frac{L\eta_t^2}{2B^2} - \frac{\eta_t}{2BN}\right)\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&+ \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2
\end{aligned}$$

Now we take $\eta_t = \frac{\eta}{\frac{N}{B}}$, and $\eta \leq \frac{1}{L}$. We also have:

$$\begin{aligned}
S &= \left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&= \left\|\frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_0^{(t)};\sigma(iB+j)) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&= \frac{1}{N^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left(\nabla f(w_0^{(t)};\sigma(iB+j)) - \nabla f(w_i^{(t)};\sigma(iB+j))\right)\right\|^2 \\
&\leq \frac{N}{N^2}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(w_0^{(t)};\sigma(iB+j)) - \nabla f(w_i^{(t)};\sigma(iB+j))\right\|^2 \\
&\leq \frac{BL^2}{N}I
\end{aligned}$$

This gives:

$$\begin{aligned}
F(\tilde{w}_{t+1}) &\leq F(\tilde{w}_t) + \frac{\eta_t N}{2B}\frac{L^2 I}{\frac{N}{B}} - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&\leq F(\tilde{w}_t) + \frac{3\eta_t L^2}{2}\frac{N^3}{B^3}\eta_t^2\left(\Theta\left\|\nabla F(w_0^{(t)})\right\|^2 + \sigma^2\right) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&\leq F(\tilde{w}_t) + \left(\frac{3}{2}L^2\eta^3\Theta - \frac{\eta}{2}\right)\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{3}{2}\eta^3 L^2\sigma^2 \\
&\leq F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{3}{2}\eta^3 L^2\sigma^2
\end{aligned}$$

□

Proof of Theorem 6. We have from 23

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{3}{2}\eta^3 L^2\sigma^2 \tag{87}$$

We can put expectations around F as follows:

$$\mathbb{E}[F(\tilde{w}_{t+1})] \leq \mathbb{E}[F(\tilde{w}_t)] - \frac{\eta}{4}\mathbb{E}[\Big\|\nabla F(\tilde{w}_t)\Big\|^2] + \frac{3\eta^3 L^2\sigma^2}{2}$$

Rearranging gives:

$$\mathbb{E}[\Big\|\nabla F(\tilde{w}_t)\Big\|^2] \leq \frac{4}{\eta}\mathbb{E}[F(\tilde{w}_t) - F(\tilde{w}_{t+1})] + 6\eta^2 L^2\sigma^2$$

Summing up the above for $t = 0, 1, \ldots, T$

$$\sum_{i=0}^{T}\mathbb{E}[\Big\|\nabla F(\tilde{w}_t)\Big\|^2] \leq \frac{4}{\eta}\mathbb{E}[F(\tilde{w}_0) - F(\tilde{w}_{T+1})] + 6\eta^2 L^2 T\sigma^2$$

Since $F(\tilde{w}_{T+1}) \geq F^*$, we get that:

$$\frac{1}{T}\sum_{i=0}^{T}\mathbb{E}[\Big\|\nabla F(\tilde{w}_t)\Big\|^2] \leq \frac{4}{\eta T}(F(\tilde{w}_0) - F^*) + 6\eta^2 L^2\sigma^2 \tag{88}$$

□

4.5 New Non-Convex Results (Theorem 7)

We begin by proving the following Lemma.

Lemma 20 (inspired by [12], Lemma 8)**.** *Suppose there are N vectors $X_i \in \mathbb{R}^d$, $i \in [N]$, and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, with $\frac{1}{N}\sum_{i=1}^{N}\Big\|X_i - \bar{X}\Big\|^2 \leq \Theta\Big\|\bar{X}\Big\|^2 + \sigma^2$, then for a randomly sampled permutation $\sigma \in \mathcal{S}_N$ and for $k < N$, we have that with probability at least $1 - \delta$ that:*

$$\Big\|\frac{1}{k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\Big\| \leq \frac{2\sqrt{\ln(\frac{2}{\delta})N\left(\Theta\bar{X}^2 + \sigma^2\right)}}{k} \tag{89}$$

We will use the following result from [9]:

Lemma 21. *Suppose that a sequence of random variables $\{x_j\}_{j\geq 1}$ is a martingale taking values in $\mathbb{R}^d$ and $\sum_{j=2}^{\infty} ess\ sup\|x_j - x_{j-1}\| \leq c^2$, for some $c > 0$. Then, for some $\lambda > 0$ we have that:*

$$\mathbb{P}\Big(sup\{\|x_j\| : j \geq 1\} \geq \lambda\Big) \leq 2exp\Big(-\frac{\lambda^2}{2c^2}\Big) \tag{90}$$

Proof of Lemma 20. Consider the following sequence of random variables for $j = 1, 2, \ldots$

$$x_j = \frac{1}{N - \min(k, j)}\sum_{t=1}^{\min(k,j)}(X_{\sigma(t)} - \bar{X})$$

We want to prove that the sequence of the random variables is a martingale, i.e.

$$\mathbb{E}[x_j | x_{j-1}, x_{j-2}, ..., x_1] = x_{j-1}$$

As we can see from the formula, we have that $x_j = x_{j-1}$ for $j \geq k + 1$, so we want to prove it for $j \leq k$. Let us rewrite x_j:

$$\begin{aligned} x_j &= \frac{1}{N - \min(k, j)}\sum_{t=1}^{j-1}(X_{\sigma(t)} - \bar{X}) + \frac{1}{N - \min(k, j)}(X_{\sigma(j)} - \bar{X}) \\ &= \frac{1}{N - \min(k, j-1)}\sum_{t=1}^{j-1}(X_{\sigma(t)} - \bar{X}) + \Big(\frac{1}{N - \min(k, j)}\sum_{t=1}^{j-1}(X_{\sigma(t)} - \bar{X}) - \\ &- \frac{1}{N - \min(k, j-1)}\sum_{t=1}^{j-1}(X_{\sigma(t)} - \bar{X})\Big) + \frac{1}{N - \min(k, j)}(X_{\sigma(j)} - \bar{X}) \end{aligned}$$

If $j \leq k$, then

$$x_j = x_{j-1} + \frac{1}{(N-j)(N-j+1)} \sum_{t=1}^{j-1} (X_{\sigma(t)} - \bar{X}) + \frac{1}{N-j}(X_{\sigma(j)} - \bar{X})$$

Note that:

$$\mathbb{E}[X_{\sigma(j)} - \bar{X} | \sigma(1), ..., \sigma(j-1)] = \frac{1}{N-j+1} \sum_{t=j}^{N} (X_{\sigma(t)} - \bar{X}) = -\frac{1}{N-j+1} \sum_{t=1}^{j-1} (X_{\sigma(t)} - \bar{X})$$

Thus when $j \leq k$, we have (using the above):

$$\mathbb{E}[x_j | x_1, ..., x_{j-1}] = x_{j-1} + \frac{1}{(N-j)(N-j+1)} \sum_{t=1}^{j-1} (X_{\sigma(t)} - \bar{X}) + \frac{1}{N-j} \mathbb{E}[X_{\sigma(j)} - \bar{X}] = x_{j-1}$$

Now, note that for $j \geq k+1$, $x_j - x_{j-1} = 0$. Now we see that for $2 \leq j \leq k$:

$$\begin{aligned}
\|x_j - x_{j-1}\|^2 &= \Big\| \frac{1}{(N-j)(N-j+1)} \sum_{t=1}^{j-1} (X_{\sigma(t)} - \bar{X}) + \frac{1}{N-j}(X_{\sigma(j)} - \bar{X}) \Big\|^2 \\
&\leq 2 \Big\| \frac{1}{(N-j)(N-j+1)} \sum_{t=1}^{j-1} (X_{\sigma(t)} - \bar{X}) \Big\|^2 + 2 \Big\| \frac{1}{N-j}(X_{\sigma(j)} - \bar{X}) \Big\|^2 \\
&= \frac{2}{(N-j)^2(N-j+1)^2} \Big\| \sum_{t=j}^{N} (X_{\sigma(t)} - \bar{X}) \Big\|^2 + \frac{2}{(N-j)^2} \Big\| (X_{\sigma(j)} - \bar{X}) \Big\|^2 \\
&\leq \frac{2}{(N-j)^2(N-j+1)^2} \left((N-j+1) \sum_{t=j}^{N} \Big\| X_{\sigma(t)} - \bar{X} \Big\|^2 + (N-j)^2 \Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \right) \\
&\leq \frac{2}{(N-j)^3} \left(\sum_{t=j}^{N} \Big\| X_{\sigma(t)} - \bar{X} \Big\|^2 + (N-j) \Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \right)
\end{aligned}$$

Since we know that:

$$\sum_{t=1}^{j-1} \Big(X_{\sigma(t)} - \bar{X} \Big) = -\sum_{t=j}^{N} \Big(X_{\sigma(t)} - \bar{X} \Big) \tag{91}$$

So that gives:

$$\begin{aligned}
\sum_{j=2}^{\infty} \Big\| x_j - x_{j-1} \Big\|^2 &= \sum_{j=2}^{k} \Big\| x_j - x_{j-1} \Big\|^2 \\
&\leq \sum_{j=2}^{k} \frac{2}{(N-j)^3} \left(\sum_{t=j}^{N} \Big\| X_{\sigma(t)} - \bar{X} \Big\|^2 + (N-j) \Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \right) \\
&\leq \sum_{j=2}^{k} \Big(\Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \Big) \Big(\frac{2}{(N-j)^2} + \sum_{t=1}^{j} \frac{2}{(N-t)^3} \Big) \\
&\leq \sum_{j=2}^{k} \Big(\Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \Big) \Big(\frac{2}{(N-j)^2} + \sum_{t=1}^{j} \frac{2}{(N-j)^3} \Big) \\
&\leq \sum_{j=2}^{k} \Big(\Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \Big) \Big(\frac{2}{(N-j)^2} + \frac{2j}{(N-j)^3} \Big) \\
&\leq \sum_{j=2}^{k} \Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \frac{2}{(N-j)^2} \Big(1 + \frac{j}{(N-j)} \Big) \\
&\leq \sum_{j=2}^{k} \Big\| X_{\sigma(j)} - \bar{X} \Big\|^2 \frac{2N}{(N-j)^3} \leq \frac{2N^2}{(N-k)^3} \left(\Theta \Big\| \bar{X} \Big\|^2 + \sigma^2 \right)
\end{aligned}$$

Now, setting $c^2 = \frac{2N^2}{(N-k)^3}\left(\Theta\left\|\bar{X}\right\|^2 + \sigma^2\right)$ and applying lemma 21, we have:

$$\mathbb{P}(\|x_k\| \geq \lambda) \leq \mathbb{P}(\sup\{\|x_j\| : j \geq 0\} \geq \lambda) \leq 2\exp\left(-\frac{\lambda^2(N-k)^3}{2N^2(\Theta\bar{X}^2 + \sigma^2)}\right) \tag{92}$$

$$\mathbb{P}(\left\|\frac{1}{N-k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \lambda) \leq 2\exp\left(-\frac{\lambda^2(N-k)^3}{2N^2(\Theta\bar{X}^2 + \sigma^2)}\right) \tag{93}$$

$$\tag{94}$$

Setting first $\lambda = \frac{k\delta'}{N-k}$ we have:

$$\mathbb{P}(\left\|\frac{1}{k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \delta') \leq 2\exp\left(-\frac{\delta'^2 k^2(N-k)}{2N^2(\Theta\bar{X}^2 + \sigma^2)}\right) \tag{95}$$

$$\tag{96}$$

Substituting $\delta' = \frac{\sqrt{2\delta'' N^2\left(\Theta\bar{X}^2 + \sigma^2\right)}}{k\sqrt{N-k}}$ we end up having:

$$\mathbb{P}\left(\left\|\frac{1}{k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \frac{\sqrt{2\delta'' N^2\left(\Theta\bar{X}^2 + \sigma^2\right)}}{k\sqrt{N-k}}\right) \leq 2\exp(-\delta'') \tag{97}$$

$$\tag{98}$$

and thus setting up $\delta'' = \ln(\frac{2}{\delta})$, with probability $1-\delta$ we have:

$$\left\|\frac{1}{k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \frac{\sqrt{2\ln(2/\delta)N^2\left(\Theta\bar{X}^2 + \sigma^2\right)}}{k\sqrt{N-k}} \tag{99}$$

Note that if we have $k \leq \frac{N}{2}$, with probability $1-\delta$ we have:

$$\left\|\frac{1}{k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \frac{\sqrt{4\ln(2/\delta)N\left(\Theta\bar{X}^2 + \sigma^2\right)}}{k} \tag{100}$$

In order to prove the bound for a batch of size greater than $\frac{N}{2}$, take again $k \leq \frac{N}{2}$, and note that:

$$\mathbb{P}(\left\|\frac{1}{N-k}\sum_{t=1}^{j}(X_{\sigma(t)} - \bar{X})\right\| \geq \lambda) = \mathbb{P}(\left\|\frac{1}{N-k}\sum_{t=j+1}^{N}(X_{\sigma(t)} - \bar{X})\right\| \geq \lambda) \leq 2\exp\left(-\frac{\lambda^2(N-k)^3}{2N^2(\Theta\bar{X}^2 + \sigma^2)}\right) \tag{101}$$

$$\tag{102}$$

which becomes

$$\left\|\frac{1}{N-k}\sum_{t=j+1}^{N}(X_{\sigma(t)} - \bar{X})\right\| \geq \frac{\sqrt{2\ln(2/\delta)N^2\left(\Theta\bar{X}^2 + \sigma^2\right)}}{\sqrt{(N-k)^3}} \tag{103}$$

Finally, observe that $N-k > \frac{N}{2}$ so that becomes:

$$\left\|\frac{1}{N-k}\sum_{t=j+1}^{N}(X_{\sigma(t)} - \bar{X})\right\| \geq \frac{\sqrt{4N\ln(2/\delta)\left(\Theta\bar{X}^2 + \sigma^2\right)}}{N-k} \tag{104}$$

$$\tag{105}$$

which is what we wanted. Substituting back to our problem we have with probability at least $1-\delta$:

$$\left\|\frac{1}{iB}\sum_{k=0}^{i}\sum_{j=1}^{B}\nabla f(w) - \nabla F(w)\right\| \leq \frac{\sqrt{4\ln(2/\delta)N\left(\Theta\left\|\nabla F(w)\right\|^2 + \sigma^2\right)}}{iB} \tag{106}$$

$$\tag{107}$$

□

Lemma 22. *We have that with probability* $1-\delta$

$$I=\sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)}-w_0^{(t)}\right\|^2\leq\frac{N^2\eta_t^2}{B^2}\Big(\Big(3\Theta+2\Big)\left\|\nabla F(w_0^{(t)})\right\|^2+\sigma^2\Big) \tag{108}$$

Proof. We have with probability $1-\delta$ that:

$$\begin{aligned}
\left\|w_i^{(t)}-w_0^{(t)}\right\|^2&=\eta_t^2\Big\|\frac{1}{B}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)};\sigma^{(t)}(kB+j))\Big\|^2\\
&=i^2\cdot\eta_t^2\Big\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\nabla f(w_k^{(t)};\sigma^{(t)}(kB+j))\Big\|^2\\
&\overset{(a)}{\leq}3i^2\eta_t^2\Big\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\Big(\nabla f(w_k^{(t)};\sigma^{(t)}(kB+j))-\nabla f(w_0^{(t)};\sigma^{(t)}(kB+j))\Big)\Big\|^2+\\
&+3i^2\eta_t^2\Big(\Big\|\frac{1}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\Big(\nabla f(w_0^{(t)};\sigma^{(t)}(kB+j))-\nabla F(w_0^{(t)})\Big)\Big\|^2+\left\|\nabla F(w_0^{(t)})\right\|^2\Big)\\
&\overset{(b)}{\leq}\frac{3i^2\eta_t^2}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}\left\|\nabla f(w_k^{(t)};\sigma^{(t)}(kB+j))-\nabla f(w_0^{(t)};\sigma^{(t)}(kB+j))\right\|^2+\\
&+3i^2\eta_t^2\left(\frac{4\ln(2/\delta)N\Big(\Theta\left\|\nabla F(w_0^{(t)})\right\|^2+\sigma^2\Big)}{(iB)^2}+\left\|\nabla F(w_0^{(t)})\right\|^2\right)\\
&\overset{(c)}{\leq}\frac{3i^2\eta_t^2}{iB}\sum_{k=0}^{i-1}\sum_{j=1}^{B}L^2\left\|w_k^{(t)}-w_0^{(t)}\right\|^2+3\eta_t^2i^2\Big(\frac{4\ln(2/\delta)N\sigma^2}{(iB)^2}+\left\|\nabla F(w_0^{(t)})\right\|^2\Big(i^2+\frac{4\ln(2/\delta)N\Theta}{(iB)^2}\Big)\Big)\\
&=\frac{3i^2\eta_t^2L^2}{i}\sum_{k=0}^{i-1}\left\|w_k^{(t)}-w_0^{(t)}\right\|^2+3\eta_t^2i^2\Big(\frac{4\ln(2/\delta)N\sigma^2}{((iB)^2}+\left\|\nabla F(w_0^{(t)})\right\|^2\Big(i^2+\frac{4\ln(2/\delta)N\Theta}{(iB)^2}\Big)\Big)\\
&\overset{(d)}{\leq}3i\eta_t^2L^2I+3\eta_t^2i^2\Big(\frac{4\ln(2/\delta)N\sigma^2}{(iB)^2}+\left\|\nabla F(w_0^{(t)})\right\|^2\Big(i^2+\frac{4\ln(2/\delta)N\Theta}{(iB)^2}\Big)\Big)
\end{aligned}$$

The above holds because of the following:

1. (a) $(a+b+c)^2\leq 3\|a\|^2+3\|b\|^2+3\|c\|^2$
2. (b) This is due to 106 and 20.
3. (c)
$$\left\|\nabla f(w_k^{(t)};\sigma^{(t)}(kB+j))-\nabla f(w_0^{(t)};\sigma^{(t)}(kB+j))\right\|\leq L\left\|w_k^{(t)}-w_0^{(t)}\right\|$$
4. (d)
$$I=\sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)}-w_0^{(t)}\right\|^2\geq\sum_{i=1}^{i-1}\left\|w_i^{(t)}-w_0^{(t)}\right\|^2,\forall i\leq\frac{N}{B}$$

We can now finish of in the following way:

$$\begin{aligned}
I=\sum_{i=1}^{\frac{N}{B}-1}\left\|w_i^{(t)}-w_0^{(t)}\right\|^2&\leq\sum_{i=1}^{\frac{N}{B}-1}\left(3i\eta_t^2L^2I+3\eta_t^2\Big(\frac{4i^2\ln(2/\delta)N\sigma^2}{(iB)^2}+\left\|\nabla F(w_0^{(t)})\right\|^2\Big(i^2+\frac{4i^2\ln(2/\delta)N\Theta}{(iB)^2}\Big)\Big)\right)\\
&\leq\frac{3\eta_t^2L^2IN^2}{2B^2}+\frac{12\eta_t^2\ln(2/\delta)N^2\sigma^2}{B^3}+3\eta_t^2\left\|\nabla F(w_0^{(t)})\right\|^2\left(\frac{N^3}{3B^3}+\frac{4\ln(2/\delta)N^2\Theta}{2B^3}\right)
\end{aligned}$$

Now choosing η_t such that:

$$\frac{3\eta_t^2L^2IN^2}{2B^2}\leq\frac{I}{2}\iff\eta_t\leq\frac{B}{NL}\sqrt{\frac{1}{3}} \tag{109}$$

The above becomes

$$I \leq \frac{N^2\eta_t^2}{B^3}\left(12\ln(2/\delta)\sigma^2 + \left\|\nabla F(w_0^{(t)})\right\|^2 (N + 6\ln(2/\delta)\Theta)\right)$$
$$= \eta^2\left(\frac{1}{B}12\ln(2/\delta)\sigma^2 + \left\|\nabla F(w_0^{(t)})\right\|^2\left(\frac{N + 6\ln(2/\delta)\Theta}{B}\right)\right)$$

□

Lemma 23. *We have that if $\eta \leq \frac{1}{2L}$, $\frac{C}{6} \leq N$, then with probability $1 - 2e^{-C/6}$:*

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{C}{N}\eta^3 L^2\sigma^2 \tag{110}$$

Proof. Here we assume that $\tilde{w}_t = w_0^{(t)}$. We have:

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \langle \nabla F(\tilde{w}_t), \tilde{w}_{t+1} - \tilde{w}_t\rangle + \frac{L}{2}\left\|\tilde{w}_{t+1} - \tilde{w}_t\right\|^2$$
$$\leq F(\tilde{w}_t) - \left\langle \nabla F(\tilde{w}_t), \frac{\eta_t}{B}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\rangle + \frac{L\eta_t^2}{2B^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$= F(\tilde{w}_t) + \frac{L\eta_t^2}{2B^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2$$
$$- \frac{\eta_t N}{2B}\left\|\frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 + \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$= F(\tilde{w}_t) + \left(\frac{L\eta_t^2}{2B^2} - \frac{\eta_t}{2BN}\right)\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$+ \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2 - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2$$

Now we take $\eta_t = \frac{\eta}{\frac{N}{B}}$, and $\eta \leq \frac{1}{L}$. We also have:

$$S = \left\|\nabla F(\tilde{w}_t) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$= \left\|\frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_0^{(t)}; \sigma(iB+j)) - \frac{1}{N}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$= \frac{1}{N^2}\left\|\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left(\nabla f(w_0^{(t)}; \sigma(iB+j)) - \nabla f(w_i^{(t)}; \sigma(iB+j))\right)\right\|^2$$
$$\leq \frac{N}{N^2}\sum_{i=1}^{\frac{N}{B}-1}\sum_{j=1}^{B}\left\|\nabla f(w_0^{(t)}; \sigma(iB+j)) - \nabla f(w_i^{(t)}; \sigma(iB+j))\right\|^2$$
$$\leq \frac{BL^2}{N}I$$

This gives with probability $1-2e^{-C/6}$, where we have $\eta_t = \frac{\eta}{\frac{N}{B}}$:

$$\begin{aligned}
F(\tilde{w}_{t+1}) &\leq F(\tilde{w}_t) + \frac{\eta_t N}{2B}\frac{L^2 I}{\frac{N}{B}} - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&\leq F(\tilde{w}_t) + \frac{\eta_t L^2}{2}\eta_t^2\frac{N^2}{B^2}\left(\frac{1}{B}12\ln(2/\delta)\sigma^2 + \left\|\nabla F(w_0^{(t)})\right\|^2\left(\frac{N+6\ln(2/\delta)\Theta}{B}\right)\right) - \frac{\eta_t N}{2B}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&= F(\tilde{w}_t) + \frac{\eta^3 L^2}{2N}\left(12\ln(2/\delta)\sigma^2 + \left\|\nabla F(w_0^{(t)})\right\|^2(N+6\ln(2/\delta)\Theta)\right) - \frac{\eta}{2}\left\|\nabla F(\tilde{w}_t)\right\|^2 \\
&\leq F(\tilde{w}_t) + \left(\frac{\eta^3 L^2}{2}\left(1+\frac{6\ln(2/\delta)}{N}\right) - \frac{\eta}{2}\right)\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{6}{N}\ln(2/\delta)\eta^3 L^2\sigma^2 \\
&\overset{(a)}{\leq} F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(w_0^{(t)})\right\|^2 + \frac{C}{N}\eta^3 L^2\sigma^2
\end{aligned}$$

Where for (a) we used $\delta = 2e^{-C/6}$, for which means that for $\eta \leq \frac{1}{2L}, C \leq \frac{N}{6}$

$$\frac{\eta^3 L^2}{2}\left(1+\frac{6\ln(2/\delta)}{N}\right) - \frac{\eta}{2} \leq \eta^3 L^2 - \frac{\eta}{2} \leq -\frac{\eta}{4}$$

□

Proof of Theorem 7. We have from 23 with probability $1-2e^{-C/6}$

$$F(\tilde{w}_{t+1}) \leq F(\tilde{w}_t) - \frac{\eta}{4}\left\|\nabla F(\tilde{w}_t)\right\|^2 + \frac{C}{N}\eta^3 L^2\sigma^2 \tag{111}$$

Rearranging gives:

$$\left\|\nabla F(\tilde{w}_t)\right\|^2 \leq \frac{4}{\eta}\left(F(\tilde{w}_t) - F(\tilde{w}_{t+1})\right) + \frac{4}{N}C\eta^2 L^2\sigma^2$$

Summing up the above for $t = 0, 1, \ldots, T$

$$\sum_{i=0}^{T}\left\|\nabla F(\tilde{w}_t)\right\|^2 \leq \frac{4}{\eta}(F(\tilde{w}_0) - F(\tilde{w}_{T+1})) + \frac{4}{N}CT\eta^2 L^2\sigma^2$$

Since $F(\tilde{w}_{T+1}) \geq F^*$, we get that with probability $1-\frac{T}{2e^{C/6}}$:

$$\frac{1}{T}\sum_{i=0}^{T}\left\|\nabla F(\tilde{w}_t)\right\|^2 \leq \frac{4}{\eta T}(F(\tilde{w}_0) - F^*) + \frac{4}{N}C\eta^2 L^2\sigma^2 \tag{112}$$

□