A. Proof of Proposition 1

We first prove the item 1. If $\epsilon_i = 0$ for certain i, we have that

$$\sup_{\theta \in \Theta^*} \ell_i(\theta) = \ell_i^*.$$

Since ℓ_i^* is the global minimum of ℓ_i, we conclude from the above equality that $\Theta^* \subset \Theta_i^*$.

Next, we prove item 2. If $\epsilon_i = 0$ for all i, by item 1 we know that $\Theta^* \subset \Theta_i^*$ for all i, and hence $\Theta^* \subset \bigcap_{i=1}^n \Theta_i^*$. Now suppose there exists $\theta \in \bigcap_{i=1}^n \Theta_i^* \setminus \Theta^*$. Then, θ simultaneously minimizes all the sample losses and must be a minimizer of the total loss, i.e., $\theta \in \Theta^*$, contradiction.

B. Connection between Minimizer Incoherence and other Loss Conditions

The notion of minimizer incoherence is related to other loss conditions that have been studied in the existing literature. We outline their connections in this section.

► **Bounded variance** (Ghadimi & Lan, 2013): In stochastic optimization, it is standard to assume that the variance of the stochastic gradients is bounded, i.e., for all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}_\xi \|\nabla \ell_\xi(\theta) - \nabla f(\theta)\|^2 \leq \sigma^2. \tag{5}$$

In particular, when the total loss f has a unique minimizer θ^* and all sample losses $\{\ell_i\}_{i=1}^n$ are 1-gradient dominated[1], the stochastic gradient variance at θ^* satisfies

$$\mathbb{E}_\xi \|\nabla \ell_\xi(\theta^*) - \nabla f(\theta^*)\|^2 \geq \frac{1}{n} \sum_{i=1}^n (\ell_i(\theta^*) - \ell_i^*).$$

in which the right hand side corresponds to the average minimizer incoherence $\frac{1}{n}\sum_{i=1}^n \epsilon_i$. Therefore, minimizer incoherence provides an estimate of the stochastic gradient variance at the global minimum, and is weaker than the uniformly-bounded variance condition in eq. (5).

► **Second moment condition** (Bottou et al., 2018): This condition generalizes the previous bounded variance condition as: for some $C \geq 1$ and all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}_\xi \|\nabla \ell_\xi(\theta)\|^2 \leq \sigma^2 + C\|\nabla f(\theta)\|^2. \tag{6}$$

In particular, the bounded variance condition corresponds to the second moment condition with $C = 1$. In the special case that $\sigma^2 = 0$ and all the sample losses are convex, the second moment condition implies that $\nabla \ell_i(\theta^*) = 0$ for all i and all $\theta^* \in \Theta^*$, i.e., every global minimizer of the total loss also minimizes all the sample losses, which further implies full minimizer coherence.

► **Interpolation** (Ma et al., 2017): This condition assumes that the total loss f has a unique minimizer θ^* such that

$$\ell_i(\theta^*) = \ell_i^* \text{ for all } i = 1, ..., n.$$

It can be viewed a special case of the full minimizer coherence, in which the sample losses can share multiple minimizers.

► **Growth condition:** In (Tseng, 1998; Schmidt & Roux, 2013), the authors considered a strong growth condition: for some $C \geq 1$ and all $\theta \in \mathbb{R}^d$,

$$\max_i \|\nabla \ell_i(\theta)\| \leq C\|\nabla f(\theta)\|. \tag{7}$$

When all the sample losses are convex, the above condition implies full minimizer coherence. A relaxed version of this condition has been proposed in (Vaswani et al., 2018) as the weak growth condition, which relaxes the $\max_i$ in eq. (7) to $\mathbb{E}_i$.

► **Expected smoothness** (Gower et al., 2019): This condition generalizes the weak growth condition as: for some $L > 0$ all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}_\xi \big[\|\nabla \ell_\xi(\theta) - \nabla \ell_\xi(\theta^*)\|^2\big] \leq L\big(f(\theta) - f^*\big), \tag{8}$$

[1] ℓ is called 1-gradient dominated if $\ell(\theta) - \ell^* \leq \|\nabla \ell(\theta)\|^2$.

where θ^* is the unique minimizer of f. In the case of full minimizer coherence, (Gower et al., 2019) proved that expected smoothness implies the weak growth condition.

C. Proof of Lemma 1

Consider the k-th iteration with sample $\xi(k)$. By smoothness of $\ell_{\xi(k)}$, we obtain that

$$\ell_{\xi(k)}(\theta_{k+1}) \leq \ell_{\xi(k)}(\theta_k) + \langle \theta_{k+1} - \theta_k, \nabla \ell_{\xi(k)}(\theta_k) \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2.$$

On the other hand, by restricted convexity of $\ell_{\xi(k)}$, we have: for all $\omega \in \Theta^*_{\xi(k)}$,

$$\ell_{\xi(k)}(\omega) \geq \ell_{\xi(k)}(\theta_k) + \langle \omega - \theta_k, \nabla \ell_{\xi(k)}(\theta_k) \rangle.$$

Combining the above two inequalities yields that

$$\begin{aligned}
\ell_{\xi(k)}(\theta_{k+1}) &\leq \ell_{\xi(k)}(\omega) + \langle \theta_k - \omega, \nabla \ell_{\xi(k)}(\theta_k) \rangle + \langle \theta_{k+1} - \theta_k, \nabla \ell_{\xi(k)}(\theta_k) \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \ell^*_{\xi(k)} + \langle \theta_{k+1} - \omega, \nabla \ell_{\xi(k)}(\theta_k) \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \ell^*_{\xi(k)} + \langle \theta_{k+1} - \omega, -\frac{1}{\eta}(\theta_{k+1} - \theta_k) \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \ell^*_{\xi(k)} + \frac{1}{2\eta}\left[\|\theta_k - \omega\|^2 - \|\theta_{k+1} - \omega\|^2 - \|\theta_{k+1} - \theta_k\|^2\right] + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \ell^*_{\xi(k)} + \frac{1}{2\eta}[\|\theta_k - \omega\|^2 - \|\theta_{k+1} - \omega\|^2] - \Big(\frac{1}{2\eta} - \frac{L}{2}\Big)\|\theta_{k+1} - \theta_k\|^2.
\end{aligned}$$

Rearranging the above inequality further yields that: for all $\omega \in \Theta^*_{\xi(k)}$,

$$\|\theta_{k+1} - \omega\|^2 \leq \|\theta_k - \omega\|^2 - 2\eta[\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}] - \big(1 - \eta L\big)\|\theta_{k+1} - \theta_k\|^2. \tag{9}$$

Choose $\eta \leq \frac{1}{L}$, we conclude that for all $\omega \in \Theta^*_{\xi(k)}$,

$$\|\theta_{k+1} - \omega\|^2 \leq \|\theta_k - \omega\|^2 - 2\eta\big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\big).$$

D. Proof of Lemma 2

Note that by Lemma 1, we have that for all $\omega \in \Theta^*_{\xi(k)}$,

$$\begin{aligned}
\|\theta_{k+1} - \omega\|^2 &\leq \|\theta_k - \omega\|^2 - \eta\big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\big) \\
&\leq \|\theta_k - \omega\|^2.
\end{aligned}$$

In the case of full minimizer coherence, we have $\Theta^* \subset \Theta^*_{\xi(k)}$. Therefore, the above result further implies that: for all k and any fixed $\omega \in \Theta^*$,

$$\|\theta_{k+1} - \omega\| \leq \|\theta_k - \omega\| \leq \cdots \leq \|\theta_0 - \omega\| < +\infty,$$

where we have used the fact that both Θ^* and θ_0 are bounded. Further notice that $\|\theta_{k+1}\| \leq \|\omega\| + \|\theta_{k+1} - \omega\|$, we conclude that the entire trajectory $\{\theta_k\}_k$ is bounded.

E. Proof of Proposition 2

We first prove item 1. Note that by Proposition 1 we have $\Theta^* = \bigcap_{i=1}^n \Theta^*_i$. In the proof of Lemma 1 we have shown in eq. (9) that for any $\omega \in \Theta^*_{\xi(k)}$

$$\|\theta_{k+1} - \omega\|^2 \leq \|\theta_k - \omega\|^2 - 2\eta[\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}] - \big(1 - \eta L\big)\|\theta_{k+1} - \theta_k\|^2.$$

We can choose any $\omega \in \Theta^*$ and sum the above bound over the B-th epoch to obtain that

$$\|\theta_{n(B+1)} - \omega\|^2 \leq \|\theta_{nB} - \omega\|^2 - 2\eta \sum_{k=nB}^{n(B+1)-1} \big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\big) - \sum_{k=nB}^{n(B+1)-1} \big(1 - \eta L\big)\|\theta_{k+1} - \theta_k\|^2.$$

Rearranging the above inequality yields that

$$\sum_{k=nB}^{n(B+1)-1} \Big[\big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\big) + \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|\theta_{k+1} - \theta_k\|^2\Big] \leq \frac{1}{2\eta}\big(\|\theta_{nB} - \omega\|^2 - \|\theta_{n(B+1)} - \omega\|^2\big).$$

Further summing the above bound over the epochs $K = 0, ..., B-1$ yields that

$$\sum_{K=0}^{B-1} \sum_{k=nK}^{n(K+1)-1} \Big[\big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\big) + \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|\theta_{k+1} - \theta_k\|^2\Big] \leq \frac{1}{2\eta}\|\theta_0 - \omega\|^2. \tag{10}$$

Note that $\ell_{\xi_k}(\theta_{k+1}) - \ell^*_{\xi_k}$ is non-negative, and $\big(\frac{1}{2\eta} - \frac{L}{2}\big)\|\theta_{k+1} - \theta_k\|^2$ is also non-negative if we choose $\eta \leq \frac{1}{L}$. Also, the left hand side of the above inequality is bounded above for all B. Therefore, it implies that $\ell_{\xi_k}(\theta_{k+1}) - \ell^*_{\xi_k} \overset{k}{\to} 0$, $\|\theta_{k+1} - \theta_k\| \overset{k}{\to} 0$. In particular, for all subsequences $\{i(T)\}_T, i = 1, ..., n$, we have $\ell_i(\theta_{i(T)+1}) - \ell^*_i \overset{T}{\to} 0$. Therefore, by continuity of the sample losses, we conclude that all the limit points of $\{\theta_{i(T)+1}\}_T$ belong to the set Θ^*_i for all i. Since $\|\theta_{k+1} - \theta_k\| \overset{k}{\to} 0$, we conclude that all the limit points $\mathfrak{X}_i$ of $\{\theta_{i(T)}\}_T$ belong to the set Θ^*_i for all i, and item 1 is proved.

Next, we prove item 2. It suffices to show that $\mathfrak{X}_i = \mathfrak{X}_j$ for all $i \neq j$. Consider any $\omega \in \mathfrak{X}_i$ with a corresponding subsequence $\theta_{i(T_k)} \overset{k}{\to} \omega$. By the random reshuffle sampling, we have $|i(T_k) - j(T_k)| \leq n$ for all i, j, k. Also, note that $\|\theta_{k+1} - \theta_k\| \overset{k}{\to} 0$. We obtain that

$$\|\theta_{j(T_k)} - \omega\| \leq \|\theta_{j(T_k)} - \theta_{i(T_k)}\| + \|\theta_{i(T_k)} - \omega\| \overset{k}{\to} 0. \tag{11}$$

Therefore, we showed that every $\omega \in \mathfrak{X}_i$ is also in any other $\mathfrak{X}_j$. In summary, $\mathfrak{X}_i = \mathfrak{X}_j = \mathfrak{X}$. Moreover, since item 1 shows that $\mathfrak{X}_i \subset \Theta^*_i$, we further obtain that $\mathfrak{X} \subset \bigcap_{i=1}^n \Theta^*_i$.

F. Proof of Theorem 1

We prove it by contradiction. Assume there exists $\omega_1, \omega_2 \in \mathfrak{X}$ such that $\omega_1 \neq \omega_2$. Let $\theta_{q(k)} \to \omega_1$ and $\theta_{p(k)} \to \omega_2$ be two converged subsequences. Without loss of generality, we can always assume that $p(k) > q(k)$ (if not, simply take a subsequence of $\{p(k)\}_k$ such that this property is satisfied).

Apply the inequality in Lemma 1 with any $\omega \in \mathfrak{X} \subset \Theta^*$ and note that $p(k) > q(k)$, we obtain that

$$\|\theta_{p(k)} - \omega\| \leq \|\theta_{q(k)} - \omega\|. \tag{12}$$

In particular, set $\omega = \omega_1$, the right hand side of the above inequality converges to 0 because ω_1 is the unique limit point of $\theta_{q(k)}$ by our choice. Therefore, we conclude that ω_1 is also a limit point of $\{\theta_{p(k)}\}_k$, and hence $\omega_1 = \omega_2$, contradiction.

G. Proof of Theorem 2

Consider the k-th iteration with sample $\xi(k)$. By smoothness of $\ell_{\xi(k)}$, we obtain that

$$\ell_{\xi(k)}(\theta_{k+1}) \leq \ell_{\xi(k)}(\theta_k) + \langle \theta_{k+1} - \theta_k, \nabla \ell_{\xi(k)}(\theta_k)\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2.$$

On the other hand, by restricted strong convexity of $\ell_{\xi(k)}$, we have: for all $\omega \in \Theta^*_{\xi(k)}$,

$$\ell_{\xi(k)}(\omega) \geq \ell_{\xi(k)}(\theta_k) + \langle \omega - \theta_k, \nabla \ell_{\xi(k)}(\theta_k)\rangle + \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2. \tag{13}$$

Combining both inequalities above, we obtain that: for all $\omega \in \Theta^*$,

$$\begin{aligned}\ell_{\xi(k)}(\theta_{k+1}) &\le \ell_{\xi(k)}(\omega) + \langle \theta_{k+1} - \omega, \nabla \ell_{\xi(k)}(\theta_k)\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2 \\ &= \ell_{\xi(k)}(\omega) + \langle \theta_{k+1} - \omega, \frac{1}{\eta}(\theta_k - \theta_{k+1})\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2 \\ &= \ell_{\xi(k)}(\omega) + \frac{1}{2\eta}(\|\theta_k - \omega\|^2 - \|\theta_{k+1} - \omega\|^2) - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2.\end{aligned}$$

Now let $\eta = \frac{1}{L}$. We further obtain that: for all $\omega \in \Theta^*$,

$$\begin{aligned}\|\theta_{k+1} - \omega\|^2 &\le \big(1 - \mu_{\xi(k)}\eta\big)\|\theta_k - \omega\|^2 - 2\eta\Big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)}\Big) \\ &\le \big(1 - \frac{\mu_{\xi(k)}}{L}\big)\|\theta_k - \omega\|^2. \end{aligned} \tag{14}$$

Telescoping the above inequality over the B-th epoch and by sampling with random reshuffle, we conclude that: for all $\omega \in \Theta^*$,

$$\begin{aligned}\|\theta_{n(B+1)} - \omega\|^2 &\le \prod_{i=1}^{n}\big(1 - \frac{\mu_i}{L}\big)\|\theta_{nB} - \omega\|^2 \\ &= \alpha\|\theta_{nB} - \omega\|^2.\end{aligned}$$

In particular, choose $\omega = \arg\min_{u\in\Theta^*}\|\theta_{nB} - u\|$, the above inequality further implies that

$$\mathrm{dist}^2_{\Theta^*}(\theta_{n(B+1)}) \le \|\theta_{n(B+1)} - \omega\|^2 \le \alpha\|\theta_{nB} - \omega\|^2 = \alpha\mathrm{dist}^2_{\Theta^*}(\theta_{nB}).$$

The desired result follows by telescoping the above inequality over the epoch index B.

H. Proof of Proposition 3

One can check that eq. (14) still holds for SGD with random sampling, i.e.,

$$\|\theta_{k+1} - \omega\|^2 \le \big(1 - \frac{\mu_{\xi(k)}}{L}\big)\|\theta_k - \omega\|^2.$$

Taking expectation on both sides of the above inequality yields that

$$\mathbb{E}\|\theta_{k+1} - \omega\|^2 \le \big(1 - \frac{\bar{\mu}}{L}\big)\mathbb{E}\|\theta_k - \omega\|^2,$$

where $\bar{\mu} := \frac{1}{n}\sum_{i=1}^n \mu_i$. Telescoping the above inequality over the B epochs yields that

$$\mathbb{E}\|\theta_{nB} - \omega\|^2 \le \big(1 - \frac{\bar{\mu}}{L}\big)^{nB}\mathbb{E}\|\theta_0 - \omega\|^2.$$

I. Proof of Lemma 3

Consider the k-th iteration with sample $\xi(k)$. By smoothness of $\ell_{\xi(k)}$, we obtain that

$$\ell_{\xi(k)}(\theta_{k+1}) \le \ell_{\xi(k)}(\theta_k) + \langle \theta_{k+1} - \theta_k, \nabla\ell_{\xi(k)}(\theta_k)\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2.$$

On the other hand, by restricted strong convexity of $\ell_{\xi(k)}$, we have for $\omega = \mathrm{proj}_{\Theta^*}(\theta_k)$,

$$\ell_{\xi(k)}(\omega) \ge \ell_{\xi(k)}(\theta_k) + \langle \omega - \theta_k, \nabla\ell_{\xi(k)}(\theta_k)\rangle + \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2. \tag{15}$$

Combining both of the above inequalities, we obtain that

$$\begin{aligned}\ell_{\xi(k)}(\theta_{k+1}) &\leq \ell_{\xi(k)}(\omega) + \langle \theta_{k+1} - \omega, \nabla \ell_{\xi(k)}(\theta_k)\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2 \\ &= \ell_{\xi(k)}(\omega) + \langle \theta_{k+1} - \omega, \frac{1}{\eta}(\theta_k - \theta_{k+1})\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2 \\ &= \ell_{\xi(k)}(\omega) + \frac{1}{2\eta}(\|\theta_k - \omega\|^2 - \|\theta_{k+1} - \omega\|^2) - (\frac{1}{2\eta} - \frac{L}{2})\|\theta_{k+1} - \theta_k\|^2 - \frac{\mu_{\xi(k)}}{2}\|\theta_k - \omega\|^2.\end{aligned}$$

Choose $\eta = \frac{1}{L}$ and rearrange the above inequality, we obtain that

$$\begin{aligned}\|\theta_{k+1} - \omega\|^2 &\leq \big(1 - \mu_{\xi(k)}\eta\big)\|\theta_k - \omega\|^2 - 2\eta\Big(\ell_{\xi(k)}(\theta_{k+1}) - \ell_{\xi(k)}(\omega)\Big) \\ &\leq \big(1 - \frac{\mu_{\xi(k)}}{L}\big)\|\theta_k - \omega\|^2 - 2\eta\Big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)} + \ell^*_{\xi(k)} - \ell_{\xi(k)}(\omega)\Big) \\ &\leq \big(1 - \frac{\mu_{\xi(k)}}{L}\big)\|\theta_k - \omega\|^2 - 2\eta\Big(\ell_{\xi(k)}(\theta_{k+1}) - \ell^*_{\xi(k)} + \ell^*_{\xi(k)} - \sup_{\omega\in\Theta^*}\ell_{\xi(k)}(\omega)\Big) \\ &\leq \big(1 - \frac{\mu_{\xi(k)}}{L}\big)\|\theta_k - \omega\|^2 + 2\eta\epsilon, \end{aligned} \tag{16}$$

where the last inequality uses the definition of minimizer incoherence, which is bounded by ϵ. Telescoping the above inequality over the iterations of the B-th epoch, we obtain that

$$\|\theta_{n(B+1)} - \omega\|^2 \leq \prod_{i=1}^{n}\Big(1 - \frac{\mu_i}{L}\Big)\|\theta_{nB} - \omega\|^2 + 2\eta\epsilon\sum_{k=nB}^{n(B+1)-1}\prod_{s=k+1}^{n(B+1)-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big), \tag{17}$$

where we define $\prod_{s=n(B+1)}^{n(B+1)-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big) = 1$ by default. Note that the above inequality is an epochwise contraction with a bounded error term $\eta\epsilon\sum_{k=nB}^{n(B+1)-1}\prod_{s=k+1}^{n(B+1)-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big)$, we conclude that $\|\theta_{n(B+1)} - \omega\|^2$ is bounded for all B and hence $\{\theta_k\}_k$ is bounded.

J. Proof of Theorem 3

Note that eq. (17) further implies that

$$\begin{aligned}\mathrm{dist}^2_{\Theta^*}(\theta_{n(B+1)}) &\leq \|\theta_{n(B+1)} - \omega\|^2 \\ &\leq \prod_{i=1}^{n}\Big(1 - \frac{\mu_i}{L}\Big)\|\theta_{nB} - \omega\|^2 + 2\eta\epsilon\sum_{k=nB}^{n(B+1)-1}\prod_{s=k+1}^{n(B+1)-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big) \\ &= \prod_{i=1}^{n}\Big(1 - \frac{\mu_i}{L}\Big)\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) + 2\eta\epsilon\sum_{k=nB}^{n(B+1)-1}\prod_{s=k+1}^{n(B+1)-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big). \end{aligned} \tag{18}$$

Next, denote σ_B as the random shuffle permutation performed in epoch B and define the quantity

$$M(\sigma_B) := \sum_{k=n(B-1)}^{nB-1}\prod_{s=k+1}^{nB-1}\Big(1 - \frac{\mu_{\xi(s)}}{L}\Big).$$

It is clear that $M(\sigma_B)$ is a random variable that depends on the permutation σ_B. We define its expectation as $\mathbb{E}_\sigma M(\sigma_B) := \overline{M}$, which is a fixed constant for every epoch B. Then, taking expectation on both sides of eq. (17) yields that

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{n(B+1)}) \leq \alpha\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) + 2\eta\epsilon\overline{M}.$$

Rearranging the above inequality further yields that

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{n(B+1)}) - \frac{2\eta\epsilon\overline{M}}{1-\alpha} \leq \alpha\Big(\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) - \frac{2\eta\epsilon\overline{M}}{1-\alpha}\Big),$$

which, after telescoping over B, further gives that: for all B,

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) \leq \alpha^B\Big(\mathrm{dist}^2_{\Theta^*}(\theta_0) - \frac{2\eta\epsilon\overline{M}}{1-\alpha}\Big) + \frac{2\eta\epsilon\overline{M}}{1-\alpha}.$$

Lastly, note that we choose $\eta = \frac{1}{L}$.

K. Proof of Corollary 2

The proof is similar to that of Theorem 3. The only difference is that the sampling order of the index $\{\sigma(k)\}_k$ is now deterministic.

One can check that eq. (18) is valid for SGD with incremental sampling by replacing $\xi(s)$ with $\sigma(s)$, and we have

$$\begin{aligned}\mathrm{dist}^2_{\Theta^*}(\theta_{n(B+1)}) &\leq \prod_{i=1}^{n}\Big(1-\frac{\mu_i}{L}\Big)\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) + 2\eta\epsilon\sum_{k=nB}^{n(B+1)-1}\prod_{s=k+1}^{n(B+1)-1}\Big(1-\frac{\mu_{\sigma(s)}}{L}\Big)\\ &\overset{\text{def}}{:=} \prod_{i=1}^{n}\Big(1-\frac{\mu_i}{L}\Big)\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) + 2\eta\epsilon\widetilde{M}.\end{aligned}$$

Then, the desired result follows from a standard telescoping over B and $\eta = \frac{1}{L}$.

L. Proof of Proposition 4

One can check that eq. (16) still holds for SGD with random sampling and step size $\eta = \frac{1}{L}$. Taking expectations on both sides of the inequality and simplifying yields that

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{k+1}) \leq \big(1-\frac{\overline{\mu}}{L}\big)\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_k) + 2\eta\epsilon, \tag{19}$$

Rearranging and simplifying the above inequality yields that

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{k+1}) - \frac{2\eta\epsilon}{1-(1-\overline{\mu}/L)} \leq \big(1-\frac{\overline{\mu}}{L}\big)\Big(\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_k) - \frac{2\eta\epsilon}{1-(1-\overline{\mu}/L)}\Big),$$

which, after telescoping over k, further gives that: for all $k = nB$,

$$\mathbb{E}\mathrm{dist}^2_{\Theta^*}(\theta_{nB}) \leq \big(1-\frac{\overline{\mu}}{L}\big)^{nB}\Big(\mathrm{dist}^2_{\Theta^*}(\theta_0) - \frac{2\eta\epsilon L}{\overline{\mu}}\Big) + \frac{2\eta\epsilon L}{\overline{\mu}}.$$