
NONLINEAR MARKOVIAN STOCHASTIC APPROXIMATION

February 27, 2025

Mohammadhadi Hadavi
Prof. Hoi-To Wai - Chinese University of Hong Kong
Prof. Wenlong Mou - University of Toronto

1 Preliminaries

Notations The Euclidean norm is denoted by $\|\cdot\|$. The lowercase letter c and its derivatives c', c_0 , etc. denote universal numerical constants, whose value may change from line to line. As we are primarily interested in dependence of α and k , we adopt the following big- O notation: $\|f\| = \mathcal{O}(h(\alpha, k))$ if it holds that $\|f\| \leq s \cdot \|h(\alpha, k)\|$ for some constant $s > 0$.

We use of the following iteration scheme:

$$\theta_{t+1} = \theta_t + \alpha (g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t))$$

1.1 Assumptions

Assumption 1 For each $X \in \mathcal{X}$, the function $g(\theta, X)$ is three times continuously differentiable in θ with uniformly bounded first to third derivatives, i.e., $\sup_{\theta \in \mathbb{R}^d} \|g^{(i)}(\theta, X)\| < \infty$ for $i = 1, 2, 3, X \in \mathcal{X}$. Moreover, there exists a constant $L_1 > 0$ such that (1) $\|g^{(i)}(\theta, X) - g^{(i)}(\theta', X)\| \leq L_1$, for all $\theta, \theta' \in \mathbb{R}^d, i = 0, 1, 2$ and $X \in \mathcal{X}$, and (2) $\|g(0, X)\| \leq L_1$ for all $X \in \mathcal{X}$.

Assumption 1 implies that $g(\theta, X)$ is L_1 -Lipschitz w.r.t θ uniformly in X . The above assumption immediately implies that the growth of $\|g\|$ and $\|\bar{g}\|$ will be at most linear in θ , i.e., $\|g(\theta, X)\| \leq L_1(\|\theta - \theta^*\| + 1)$ and $\|\bar{g}(\theta)\| \leq L_1(\|\theta - \theta^*\| + 1)$.

Assumption 2 There exists $\mu > 0$ such that $\langle \theta - \theta', \bar{g}(\theta) - \bar{g}(\theta') \rangle \leq -\mu \|\theta - \theta'\|^2, \forall \theta, \theta' \in \mathbb{R}^d$. Consequently, the target equation $\bar{g}(\theta) = 0$ has a unique solution θ^* .

Denote by \mathcal{F}_k the filtration generated by $\{X_{t+1}, \theta_t, \xi_{t+1}\}_{t=0}^{k-1} \cup \{X_{k+1}, \theta_k\}$.

Assumption 3 Let $p \in \mathbb{Z}_+$ be given. The noise sequence $(\xi_k)_{k \geq 1}$ is a collection of i.i.d random fields satisfying the following conditions with $L_{2,p} > 0$:

$$\mathbb{E}[\xi_{k+1}(\theta) | \mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}^{1/(2p)}[\|\xi_1(\theta)\|^{2p}] \leq L_{2,p}(\|\theta - \theta^*\| + 1), \quad \forall \theta \in \mathbb{R}^d.$$

Define $C(\theta) = \mathbb{E}[\xi_1(\theta)^{\otimes 2}]$ and assume that $C(\theta)$ is at least twice differentiable. There also exists $M_\epsilon, k_\epsilon \geq 0$ such that for $\theta \in \mathbb{R}^d$, we have $\max_{i=1,2} \|C^{(i)}(\theta)\| \leq M_\epsilon \{1 + \|\theta - \theta^*\|^{k_\epsilon}\}$. In the sequel, we set $L := L_1 + L_2$, and without loss of generality, we assume $L \geq 1$.

Assumption 4 There exists a Borel measurable function $\hat{g}: \mathbb{R}^d \times \mathcal{X} \rightarrow \mathbb{R}^d$ where for each $\theta \in \mathbb{R}^d, X \in \mathcal{X}$,

$$\hat{g}(\theta, X) - P_\theta \hat{g}(\theta, X) = g(\theta, X) - \bar{g}(\theta).$$

Assumption 5 There exists $L_{PH}^{(0)} < \infty$ and $L_{PH}^{(1)} < \infty$ such that, for all $\theta \in \mathbb{R}^d$ and $X \in \mathcal{X}$, one has $\|\hat{g}(\theta, X)\| \leq L_{PH}^{(0)}, \|P_\theta \hat{g}(\theta, X)\| \leq L_{PH}^{(0)}$. Moreover, for $(\theta, \theta') \in \mathcal{H}^2$,

$$\sup_{X \in \mathcal{X}} \|P_\theta \hat{g}(\theta, X) - P_{\theta'} \hat{g}(\theta', X)\| \leq L_{PH}^{(1)} \|\theta - \theta'\|.$$

Assumption 6 For any $\theta, \theta' \in \mathbb{R}^d$, we have $\sup_{X \in \mathcal{X}} \|P_\theta(X, \cdot) - P_{\theta'}(X, \cdot)\|_{TV} \leq L_P \|\theta - \theta'\|$.

Assumption 7 For any $\theta, \theta' \in \mathbb{R}^d$, we have $\sup_{X \in \mathcal{X}} \|g(\theta, X) - g(\theta', X)\| \leq L_H \|\theta - \theta'\|$.

Assumption 8 *There exists $\rho < 1$, $K_P < \infty$ such that*

$$\sup_{\theta \in \mathbb{R}^d, X \in \mathcal{X}} \|P_\theta^n(X, \cdot) - \pi_\theta(\cdot)\|_{TV} \leq \rho^n K_P,$$

Lemma 1 *Assume that assumptions 6-8 hold. Then, for any $\theta \in \mathbb{R}^d$ and $X \in \mathcal{X}$,*

$$\|\hat{g}(\theta, X)\| \leq \frac{\sigma K_P}{1 - \rho},$$

$$\|P_\theta \hat{g}(\theta, X)\| \leq \frac{\sigma \rho K_P}{1 - \rho}.$$

Moreover, for any $\theta, \theta' \in \mathbb{R}^d$ and $X \in \mathcal{X}$,

$$\|P_\theta \hat{g}(\theta, X) - P_{\theta'} \hat{g}(\theta', X)\| \leq \|\theta - \theta'\|,$$

where

$$L_{PH}^{(1)} = \frac{K_P^2 \sigma L_P}{(1 - \rho)^2} (2 + K_P) + \frac{K_P}{1 - \rho} L_H.$$

Proof of this lemma can be found in [2], Lemma 7.

2 Error Bound

2.1 Base Case

For the base case analysis, we can write:

$$\begin{aligned} & \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] - \mathbb{E}[\|\theta_k - \theta^*\|^2] = \\ & 2\alpha \mathbb{E}[\langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) \rangle] + \alpha^2 \mathbb{E}[\|g(\theta_k, X_{k+1})\|^2] + \alpha^2 \mathbb{E}[\|\xi_{k+1}(\theta_k)\|^2] = \\ & 2\alpha \mathbb{E}[\langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) \rangle] + 2\alpha \mathbb{E}[\langle \theta_k - \theta^*, \bar{g}(\theta_k) \rangle] + \alpha^2 \mathbb{E}[\|g(\theta_k, X_{k+1})\|^2] + \alpha^2 \mathbb{E}[\|\xi_{k+1}(\theta_k)\|^2]. \end{aligned}$$

It is easy to see that under Strong Monotonicity assumption, we have

$$\langle \theta_k - \theta^*, \bar{g}(\theta_k) \rangle = \langle \theta_k - \theta^*, \bar{g}(\theta_k) - \bar{g}(\theta^*) \rangle \leq -\mu \|\theta_k - \theta^*\|^2.$$

Additionally, under Assumption 1 and 3, we have the following upper bound

$$\begin{aligned} & \alpha^2 (\mathbb{E}[\|g(\theta_k, X_{k+1})\|^2] + \mathbb{E}[\|\xi_{k+1}(\theta_k)\|^2]) \\ & \leq \alpha^2 \left(L_1^2 \mathbb{E}[(\|\theta_k - \theta^*\| + 1)^2] + L_2^2 \mathbb{E}[(\|\theta_k - \theta^*\| + 1)^2] \right) \\ & \leq 2\alpha^2 L^2 (\mathbb{E}[\|\theta_k - \theta^*\|^2] + 1). \end{aligned}$$

Therefore, we have

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] \leq (1 - 2\alpha(-\alpha L^2 + \mu)) \mathbb{E}[\|\theta_k - \theta^*\|^2] + 2\alpha^2 L^2 + 2\alpha \mathbb{E}[\langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) \rangle]$$

Solving this recursion gives us the following inequality:

$$\begin{aligned} \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] &\leq (1 - 2\alpha(-\alpha L^2 + \mu))^{k+1} \mathbb{E}[\|\theta_0 - \theta^*\|^2] \\ &\quad + \sum_{t=0}^k (1 - 2\alpha(-\alpha L^2 + \mu))^t 2\alpha^2 L^2 \\ &\quad + \sum_{t=0}^k 2\alpha (1 - 2\alpha(-\alpha L^2 + \mu))^{k-t} \mathbb{E}[\langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle]. \end{aligned}$$

For notational simplicity we define $\gamma_t := 2\alpha(1 - 2\alpha(-\alpha L^2 + \mu))^{k-t}$ for $0 \leq t \leq k$.

The second term above is just a geometric series. According to Lemma 12 of [1] this equals to $\frac{\alpha L^2 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))^{k+1}]}{-\alpha L^2 + \mu}$.

Now, we can upper bound the third summand using below decomposition:

$$\mathbb{E} \left[\sum_{t=0}^k \gamma_t \langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle \right] = \mathbb{E}[A_1 + A_2 + A_3 + A_4 + A_5]$$

with

$$\begin{aligned} A_1 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta^*, \hat{g}(\theta_t, X_{t+1}) - P_{\theta_t} \hat{g}(\theta_t, X_t) \rangle, \\ A_2 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta^*, P_{\theta_t} \hat{g}(\theta_t, X_t) - P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t) \rangle, \\ A_3 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta_{t-1}, P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t) \rangle, \\ A_4 &:= \sum_{t=1}^k (\gamma_t - \gamma_{t-1}) \langle \theta_{t-1} - \theta^*, P_{\theta_{t-1}} \hat{g}(\theta_{t-1} - \theta^*, X_t) \rangle, \\ A_5 &:= \gamma_0 \langle \theta_0 - \theta^*, \hat{g}(\theta_0, X_0) \rangle + \gamma_k \langle \theta_k - \theta^*, P_{\theta_k} \hat{g}(\theta_k, X_{k+1}) \rangle \end{aligned}$$

For A_1 , we note that $\hat{g}(\theta_t, X_{t+1}) - P_{\theta_t} \hat{g}(\theta_t, X_t)$ is a martingale difference sequence [cf. ?] and therefore we have $\mathbb{E}[A_1] = 0$ by taking the total expectation.

For A_2 , applying Cauchy-Schwarz inequality and ??, we have

$$\begin{aligned} A_2 &\leq \sum_{t=1}^k L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\| \|\theta_t - \theta_{t-1}\| \\ &= \sum_{t=1}^k \alpha L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\| \|g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)\| \\ &\leq \sum_{t=1}^k \alpha L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\| (L_1 (\|\theta_t - \theta^*\| + 1) + L_2 (\|\theta_t - \theta^*\| + 1)) \\ &\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} (3 \|\theta_t - \theta^*\|^2 + 1) \end{aligned}$$

where the third line follows from the Lipschitzness condition and the assumption of

$$\mathbb{E}^{1/2} [\|\xi_{t+1}(\theta_t)\|^2 | \mathcal{F}_t] \leq L_2 (\|\theta_t\| + 1)$$

also, last line follows from the identity $u \leq \frac{1}{2}(1 + u^2)$.

For A_3 , we obtain

$$\begin{aligned} A_3 &\leq \sum_{t=1}^k \gamma_t \|\theta_t - \theta_{t-1}\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \\ &\leq \sum_{t=1}^k \alpha L_{PH}^{(0)} \gamma_t \|g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)\| \\ &\leq \sum_{t=1}^k \alpha L_{PH}^{(0)} \gamma_t (L_1 (\|\theta_t - \theta^*\| + 1) + L_2 (\|\theta_t - \theta^*\| + 1)) \\ &\leq \sum_{t=1}^k \alpha L L_{PH}^{(0)} \gamma_t (\|\theta_t - \theta^*\| + 1) \end{aligned}$$

where second line follows from **??** and third line is similarly done to the previous part, using Lipschitzness condition and noise assumption.

For A_4 , we have

$$\begin{aligned} A_4 &\leq \sum_{t=1}^k |\gamma_t - \gamma_{t-1}| \|\theta_{t-1} - \theta^*\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \\ &\leq \sum_{t=1}^k L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \|\theta_{t-1} - \theta^*\| \end{aligned}$$

Finally, for A_5 , we obtain

$$A_5 \leq L_{PH}^{(0)} (\gamma_0 \|\theta_0 - \theta^*\| + \gamma_k \|\theta_k - \theta^*\|)$$

which follows from Cacuchy-Schwarz inequality and **??**.

Combining the above terms and taking expectations, gives us:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=0}^k \gamma_t \langle \theta_t - \theta^*, g(\theta_t, X_{t+1} - \bar{g}(\theta_t)) \rangle \right] &\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} (1 + 3\mathbb{E}[\|\theta_t - \theta^*\|^2]) + \sum_{t=1}^k \alpha L L_{PH}^{(0)} \gamma_t (\mathbb{E}[\|\theta_t - \theta^*\|] + 1) + \\ &\quad \sum_{t=0}^{k-1} L_{PH}^{(0)} |\gamma_t - \gamma_{t+1}| \mathbb{E}[\|\theta_t - \theta^*\|] + L_{PH}^{(0)} (\gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|] + \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|]) \end{aligned}$$

now it should be noticed that as long as the α satisfies $\alpha \leq \frac{\mu}{L^2}$, we have $\gamma_t \leq \gamma_{t+1}$. Thus, we can simplify the above upper bound and write it this way:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=0}^k \gamma_t \langle \theta_t - \theta^*, g(\theta_t, X_{t+1} - \bar{g}(\theta_t)) \rangle \right] &\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} (1 + 3\mathbb{E}[\|\theta_t - \theta^*\|^2]) + \\ &\quad \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad \sum_{t=1}^k \alpha L L_{PH}^{(0)} \gamma_t + L_{PH}^{(0)} (\gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|]) \end{aligned}$$

Hence, using the derived upper bounds from the above terms, we have:

$$\begin{aligned} \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] &\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} (1 + 3\mathbb{E}[\|\theta_t - \theta^*\|^2]) + \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu)) \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^2] + L_{PH}^{(0)} \gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) L_{PH}^{(0)} \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|] + \\ &\quad \left(\frac{L}{L_{PH}^{(0)}} + 1 \right) \frac{\gamma_1 (1 - (1 - 2\alpha(-\alpha L^2 + \mu))^k)}{(1 - (1 - 2\alpha(-\alpha L^2 + \mu)))} + \frac{\alpha L^2 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))^{k+1}]}{-\alpha L^2 + \mu} \end{aligned}$$

for further notation simplicity we define $c_{1,t} := \left(\frac{L}{L_{PH}^{(0)}} + 1 \right) \frac{\gamma_1 (1 - (1 - 2\alpha(-\alpha L^2 + \mu))^t)}{(1 - (1 - 2\alpha(-\alpha L^2 + \mu)))} + \frac{\alpha L^2 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))^{t+1}]}{-\alpha L^2 + \mu}$ for $0 \leq t \leq k$. Now to write down this upper bound in a way in which it only depends on $\|\theta_0 - \theta^*\|$ related terms and constants, we can write:

$$\begin{aligned} \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] &\leq \sum_{t=1}^k \left[\frac{3\alpha L L_{PH}^{(1)} \gamma_t}{2} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} \right] + \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu)) \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^2] + L_{PH}^{(0)} \gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) L_{PH}^{(0)} \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|] + c_{1,k} \\ &= \sum_{t=1}^k \frac{3\alpha L L_{PH}^{(1)} \gamma_t}{2} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \frac{\alpha L L_{PH}^{(1)}}{2} \sum_{t=1}^k \gamma_t + \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu)) \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^2] + L_{PH}^{(0)} \gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) L_{PH}^{(0)} \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|] + c_{1,k} \\ &= \sum_{t=1}^k \frac{3\alpha L L_{PH}^{(1)} \gamma_t}{2} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu)) \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^2] + L_{PH}^{(0)} \gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) L_{PH}^{(0)} \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|] + c_{1,k} + \\ &\quad \frac{L L_{PH}^{(1)} \gamma_1 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))^k]}{4 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))]} \end{aligned}$$

where the last equality follows from the definition of γ_t s. Similarly we define $c_{2,t} := \frac{L L_{PH}^{(1)} \gamma_1 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))^t]}{4 [1 - (1 - 2\alpha(-\alpha L^2 + \mu))]}$ for $0 \leq t \leq k$. So we can write it as

$$\begin{aligned} \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] &\leq \sum_{t=1}^k \frac{3\alpha L L_{PH}^{(1)} \gamma_t}{2} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1) \gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] + \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu)) \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^2] + L_{PH}^{(0)} \gamma_1 \mathbb{E}[\|\theta_0 - \theta^*\|] + (\alpha L + 1) L_{PH}^{(0)} \gamma_k \mathbb{E}[\|\theta_k - \theta^*\|] + c_{1,k} + c_{2,k} \end{aligned}$$

Now for the second term on RHS, we note that

$$(\alpha L - 1) \gamma_t + \gamma_{t+1} \leq \alpha L \gamma_{t+1}, \quad \mathbb{E}[\|\theta_t - \theta^*\|] \leq \sqrt{\mathbb{E}[\|\theta_t - \theta^*\|^2]},$$

and consequently

$$\begin{aligned}
& \frac{1}{(1-2\alpha(-\alpha L^2 + \mu))^k} \sum_{t=1}^{k-1} L_{PH}^{(0)} ((\alpha L - 1)\gamma_t + \gamma_{t+1}) \mathbb{E}[\|\theta_t - \theta^*\|] \\
& \leq 2L_{PH}^{(0)} L \alpha^2 \sum_{t=1}^{k-1} \frac{1}{(1-2\alpha(-\alpha L^2 + \mu))^{t+1}} \sqrt{\mathbb{E}[\|\theta_t - \theta^*\|^2]} \\
& \leq 2L_{PH}^{(0)} L \alpha^2 \left(\sum_{t=1}^{k-1} \frac{1}{(1-2\alpha(-\alpha L^2 + \mu))^{t+1}} \right)^{1/2} \left(\sum_{t=1}^{k-1} \frac{1}{(1-2\alpha(-\alpha L^2 + \mu))^{t+1}} \mathbb{E}[\|\theta_t - \theta^*\|^2] \right)^{1/2} \\
& \leq 2L_{PH}^{(0)} L \alpha^2 \cdot \sum_{t=1}^{k-1} \frac{1}{(1-2\alpha(-\alpha L^2 + \mu))^{t+1}} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \frac{1}{-\alpha L^2 + \mu} \cdot \frac{2L_{PH}^{(0)} L \alpha}{(1-2\alpha(-\alpha L^2 + \mu))^k}.
\end{aligned}$$

We also note that

$$\frac{\gamma_k}{(1-2\alpha(-\alpha L^2 + \mu))^k} \mathbb{E}[\|\theta_k - \theta^*\|] \leq \alpha \frac{\mathbb{E}[\|\theta_k - \theta^*\|^2]}{(1-2\alpha(-\alpha L^2 + \mu))^k} + \frac{\alpha}{(1-2\alpha(-\alpha L^2 + \mu))^k}.$$

similarly

$$\frac{\gamma_1}{(1-2\alpha(-\alpha L^2 + \mu))^k} \mathbb{E}[\|\theta_0 - \theta^*\|] \leq \alpha \frac{\mathbb{E}[\|\theta_0 - \theta^*\|^2]}{(1-2\alpha(-\alpha L^2 + \mu))^1} + \frac{\alpha}{(1-2\alpha(-\alpha L^2 + \mu))^1}.$$

and we also define for $0 \leq t \leq k$

$$c_{3,t} := \frac{1}{-\alpha L^2 + \mu} \frac{2\alpha L_{PH}^{(0)} L}{(1-2\alpha(-\alpha L^2 + \mu))^t} + \frac{\alpha(\alpha L + 1) L_{PH}^{(0)}}{(1-2\alpha(-\alpha L^2 + \mu))^t} + \frac{\alpha L_{PH}^{(0)}}{(1-2\alpha(-\alpha L^2 + \mu))}$$

to wrap up all the remainder terms.

Substituting back and rearranging with also defining $c'_{2,t} := \frac{c_{2,t}}{(1-2\alpha(-\alpha L^2 + \mu))^t}$ and $c'_{1,t} := \frac{c_{1,t}}{(1-2\alpha(-\alpha L^2 + \mu))^t}$, yields

$$\begin{aligned}
\frac{\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2]}{(1-2\alpha(-\alpha L^2 + \mu))^k} & \leq \frac{\alpha \left((\alpha L + 1) L_{PH}^{(0)} + \frac{3}{2} L L_{PH}^{(1)} \right)}{(1-2\alpha(-\alpha L^2 + \mu))^k} \mathbb{E}[\|\theta_k - \theta^*\|^2] + \sum_{t=1}^{k-1} \frac{\alpha \left(\frac{3}{2} L L_{PH}^{(1)} + 2\alpha(1-2\alpha(-\alpha L^2 + \mu))^{-1} L L_{PH}^{(0)} \right)}{(1-2\alpha(-\alpha L^2 + \mu))^t} \mathbb{E}[\|\theta_t - \theta^*\|^2] + \\
& \quad \left((1-2\alpha(-\alpha L^2 + \mu)) + \alpha L L_{PH}^{(1)} (1-2\alpha(-\alpha L^2 + \mu))^{-1} \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + c'_{1,k} + c'_{2,k} + c_{3,k}.
\end{aligned}$$

for sufficiently small α s, we have

$$\alpha \left(\frac{3}{2} L L_{PH}^{(1)} + 2\alpha(1-2\alpha(-\alpha L^2 + \mu))^{-1} L L_{PH}^{(0)} \right) \leq 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + L L_{PH}^{(1)} \right)$$

and again in a similar fashion we have

$$\left((1-2\alpha(-\alpha L^2 + \mu)) + \alpha L L_{PH}^{(1)} (1-2\alpha(-\alpha L^2 + \mu))^{-1} \right) \leq 2\alpha L L_{PH}^{(1)} + 1$$

using above simplifications we can rewrite our upper bound as

$$\frac{\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2]}{(1-2\alpha(-\alpha L^2 + \mu))^k} \leq 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + L L_{PH}^{(1)} \right) \sum_{t=1}^k \frac{\mathbb{E}[\|\theta_t - \theta^*\|^2]}{(1-2\alpha(-\alpha L^2 + \mu))^t} + \left(2\alpha L L_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + c'_{1,k} + c'_{2,k} + c_{3,k}.$$

For solving the above recursion, we first define $S_t := 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \sum_{l=1}^t \frac{\mathbb{E}[\|\theta_l - \theta^*\|^2]}{(1 - 2\alpha(-\alpha L^2 + \mu))^l}$ for $1 \leq t \leq k$. Also we use $C_t := c'_{1,t} + c'_{2,t} + c_{3,t}$ and for $0 \leq t \leq k$, defining constant terms. Now we can write

$$\frac{\mathbb{E}[\|\theta_{t+1} - \theta^*\|^2]}{(1 - 2\alpha(-\alpha L^2 + \mu))^t} \leq S_t + \left(2\alpha LL_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + C_t.$$

using this expansion, we should first notice that

$$\begin{aligned} \frac{S_t}{S_{t-1}} &= \frac{S_{t-1} + 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \frac{\mathbb{E}[\|\theta_t - \theta^*\|^2]}{(1 - 2\alpha(-\alpha L^2 + \mu))^t}}{S_{t-1}} \\ &= 1 + 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \frac{S_{t-1} + \left(2\alpha LL_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + C_{t-1}}{S_{t-1}} \\ &\leq 1 + 4\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right). \end{aligned}$$

Now, since we have $S_0 = 0$ and

$$S_1 = 2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \frac{\mathbb{E}[\|\theta_1 - \theta^*\|^2]}{(1 - 2\alpha(-\alpha L^2 + \mu))} = \frac{2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \left(\left(2\alpha LL_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + C_0 \right)}{(1 - 2\alpha(-\alpha L^2 + \mu))}$$

it follows that $S_t \leq \left(1 + 4\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \right)^{t-1} \left[\frac{2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \left(\left(2\alpha LL_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + C_0 \right)}{(1 - 2\alpha(-\alpha L^2 + \mu))} \right]$. Substituting this upper bound into previous equations we get

$$\begin{aligned} \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] &\leq \\ &\quad (1 - 2\alpha(-\alpha L^2 + \mu))^{k-1} \left(1 + 4\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \right)^k \left[2\alpha \left((\alpha L + 1) L_{PH}^{(0)} + LL_{PH}^{(1)} \right) \left(\left(2\alpha LL_{PH}^{(1)} + 1 \right) \mathbb{E}[\|\theta_0 - \theta^*\|^2] + C_0 \right) \right] \end{aligned}$$

Choosing α sufficiently small, we can simplify the above inequality and write it as follows

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] \leq \tilde{c}_1 \cdot (1 - 2\alpha\mu)^k \mathbb{E}[\|\theta_0 - \theta^*\|^2] + \tilde{c}_2 \cdot 2\alpha L \left(L_{PH}^{(0)} + L_{PH}^{(1)} \right)$$

where \tilde{c}_1 and \tilde{c}_2 are $\mathcal{O}(1)$ constants.

2.2 General Case

In this case, we assume that the moment bound in [??] has been proven for $k \leq n-1$, we now proceed to show that the desired moment convergence holds for n with $2 \leq n \leq p$.

We start with the following decomposition of $\|\theta_{k+1} - \theta^*\|^{2n}$

$$\begin{aligned} \|\theta_{k+1} - \theta^*\|^{2n} &= \left(\|\theta_k - \theta^*\|^2 + 2\alpha \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k) \rangle + \alpha^2 \|g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k)\|^2 \right)^n \\ &= \sum_{\substack{i,j,l \\ i+j+l=n}} \binom{n}{i,j,l} \|\theta_k - \theta^*\|^{2i} \left(2\alpha \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k) \rangle \right)^j \left(\alpha \|g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k)\|^2 \right)^l \end{aligned}$$

We note the following cases.

1. $i = n, j = l = 0$. In this case, the summand is simply $\|\theta_k - \theta^*\|^{2i}$.

2. When $i = n - 1, j = 1$ and $l = 0$. In this case, the summand is of order α , i.e.,

$$\alpha \cdot 2n \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k) \rangle^j \|\theta_k - \theta^*\|^{2(n-1)}.$$

We can further decompose it as

$$\begin{aligned} & 2n\alpha \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) + \xi_{k+1}(\theta_k) \rangle \|\theta_k - \theta^*\|^{2(n-1)} \\ &= \underbrace{2n\alpha \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) + \xi_{k+1}(\theta_k) \rangle \|\theta_k - \theta^*\|^{2(n-1)}}_{T_1} + \underbrace{2n\alpha \langle \theta_k - \theta^*, \bar{g}(\theta_k) \rangle \|\theta_k - \theta^*\|^{2(n-1)}}_{T_2}. \end{aligned}$$

Note that, when (X_k) is i.i.d or from a martingale noise, we have

$$\mathbb{E}[T_1 | \theta_k] = 0$$

However, when (X_k) is Markovian, the above inequality does not hold and T_1 requires careful analysis.

Nonetheless, under the strong monotonicity assumption, we have

$$T_2 \leq -2n\alpha\mu \|\theta_k - \theta^*\|^{2n}.$$

3. For the remaining terms, we see that they are of higher orders of α . Therefore, when α is selected sufficiently small, these terms do not raise concern.

Therefore, to prove the desired moment bound, we spend the remaining section analyzing T_1 . Immediately, we note that

$$\begin{aligned} \mathbb{E}[T_1] &= \mathbb{E}[2n\alpha \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) + \mathbb{E}[\xi_{k+1}(\theta_k) | \theta_k] \rangle \|\theta_k - \theta^*\|^{2(n-1)}] \\ &= 2n\alpha \mathbb{E} \left[\underbrace{\langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) \rangle \|\theta_k - \theta^*\|^{2(n-1)}}_{T'_1} \right]. \end{aligned}$$

Subsequently, we focus on analyzing T'_1 ; but before that, we write the general recursion of the error bound. First, we define $T'_{1,t} := \langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)}$ to make T'_1 dependent on the iteration index. Now, following the above decomposition and taking the expectations, we have:

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^{2n}] \leq \mathbb{E}[\|\theta_k - \theta^*\|^{2n}] + \mathbb{E}[T'_{1,k}] - 2n\alpha\mu \mathbb{E}[\|\theta_k - \theta^*\|^{2n}] + o(\alpha) = (1 - 2n\alpha\mu) \mathbb{E}[\|\theta_k - \theta^*\|^{2n}] + \mathbb{E}[T'_{1,k}] + o(\alpha)$$

similarly to the previous case we define $\gamma_t := 2n\alpha(1 - 2n\alpha\mu)^{k-t}$ for $0 \leq t \leq k$. Solving the above recursion will give us

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^{2n}] \leq \sum_{t=0}^k \gamma_t \mathbb{E}[T'_{1,t}] + \gamma_0 \mathbb{E}[\|\theta_0 - \theta^*\|^{2n}] + o(\alpha)$$

We have to upper bound the first term in the RHS above. For this purpose, we use a similar decomposition to our base case analysis:

$$\sum_{t=0}^k \gamma_t \mathbb{E}[T'_{1,t}] = \left[\sum_{t=0}^k \gamma_t \langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)} \right] = \mathbb{E}[A_1 + A_2 + A_3 + A_4 + A_5]$$

with

$$\begin{aligned}
A_1 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta^*, \hat{g}(\theta_t, X_{t+1}) - P_{\theta_t} \hat{g}(\theta_t, X_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)}, \\
A_2 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta^*, P_{\theta_t} \hat{g}(\theta_t, X_t) - P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)}, \\
A_3 &:= \sum_{t=1}^k \gamma_t \langle \theta_t - \theta_{t-1}, P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)} \\
A_4 &:= \sum_{t=1}^k (\gamma_t - \gamma_{t-1}) \langle \theta_{t-1} - \theta^*, P_{\theta_{t-1}} \hat{g}(\theta_{t-1} - \theta^*, X_t) \rangle \|\theta_t - \theta^*\|^{2(n-1)}, \\
A_5 &:= \sum_{t=1}^k \gamma_{t-1} \langle \theta_{t-1} - \theta^*, P_{\theta_{t-1}} \hat{g}(\theta_{t-1} - \theta^*, X_t) \rangle (\|\theta_t - \theta^*\|^{2(n-1)} - \|\theta_{t-1} - \theta^*\|^{2(n-1)}) \\
A_6 &:= \gamma_0 \langle \theta_0 - \theta^*, \hat{g}(\theta_0, X_0) \rangle \|\theta_0 - \theta^*\|^{2(n-1)} + \gamma_k \langle \theta_k - \theta^*, P_{\theta_k} \hat{g}(\theta_k, X_{k+1}) \rangle \|\theta_k - \theta^*\|^{2(n-1)}
\end{aligned}$$

For A_1 , we note that $\hat{g}(\theta_t, X_{t+1}) - P_{\theta_t} \hat{g}(\theta_t, X_t)$ is a martingale difference sequence [cf. ?] and therefore we have $\mathbb{E}[A_1] = 0$ by taking the total expectation.

For A_2 , applying Cauchy-Schwarz inequality and ??, we have

$$\begin{aligned}
A_2 &\leq \sum_{t=1}^k L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\|^{2n-1} \|\theta_t - \theta_{t-1}\| \\
&= \sum_{t=1}^k \alpha L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\|^{2n-1} \|g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)\| \\
&\leq \sum_{t=1}^k \alpha L_{PH}^{(1)} \gamma_t \|\theta_t - \theta^*\|^{2n-1} (L_1 (\|\theta_t - \theta^*\| + 1) + L_2 (\|\theta_t - \theta^*\| + 1)) \\
&\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(1)} \gamma_t}{2} (3 \|\theta_t - \theta^*\|^{2n} + 1)
\end{aligned}$$

where the third line follows from the Lipschitzness condition and the assumption of

$$\mathbb{E}^{1/2} [\|\xi_{t+1}(\theta_t)\|^2 | \mathcal{F}_t] \leq L_2 (\|\theta_t\| + 1).$$

Also, last line follows from the identity $u \leq \frac{1}{2}(1 + u^2)$.

For A_3 , we obtain

$$\begin{aligned}
A_3 &\leq \sum_{t=1}^k \gamma_t \|\theta_t - \theta_{t-1}\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k \alpha L_{PH}^{(0)} \gamma_t \|g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)\| \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k \alpha L_{PH}^{(0)} \gamma_t (L_1 (\|\theta_t - \theta^*\| + 1) + L_2 (\|\theta_t - \theta^*\| + 1)) \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k \frac{\alpha L L_{PH}^{(0)} \gamma_t}{2} (3 \|\theta_t - \theta^*\|^{2n-1} + 1)
\end{aligned}$$

where second line follows from ?? and third line is similarly done to the previous part, using Lipschitzness condition and noise assumption.

For A_4 , we have

$$\begin{aligned}
A_4 &\leq \sum_{t=1}^k |\gamma_t - \gamma_{t-1}| \|\theta_{t-1} - \theta^*\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \|\theta_{t-1} - \theta^*\| \|\theta_t - \theta^*\|^{2(n-1)} \\
&= \sum_{t=1}^k L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \|\theta_{t-1} - \theta_t + \theta_t - \theta^*\| \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| (\|\theta_{t-1} - \theta_t\| + \|\theta_t - \theta^*\|) \|\theta_t - \theta^*\|^{2(n-1)} \\
&\leq \sum_{t=1}^k L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \left(\frac{\alpha L}{2} (3\|\theta_t - \theta^*\|^{2n-1} + 1) + \|\theta_t - \theta^*\|^{2n-1} \right) \\
&\leq \sum_{t=1}^k 4L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \|\theta_t - \theta^*\|^{2n-1}
\end{aligned}$$

where the fourth line follows from the triangle inequality, fifth line uses a similar argument to the previous part and the last line uses the fact that $\alpha < \frac{1}{L}$.

Now for A_5 , we have to first note that, using mean-value theorem and with $a \in [0, 1]$, we'll get

$$\begin{aligned}
\|\theta_t - \theta^*\|^{2(n-1)} - \|\theta_{t-1} - \theta^*\|^{2(n-1)} &= \|\theta_t - \theta_{t-1}\| \cdot 2(n-1) \|a(\theta_t - \theta^*) + (1-a)(\theta_{t-1} - \theta^*)\|^{2n-3} \\
&\leq \|\theta_t - \theta_{t-1}\| \cdot 2(n-1) \|a(\theta_t - \theta_{t-1}) + \theta_{t-1} - \theta^*\|^{2n-3} \\
&\leq 2^{2n-3} (n-1) \|\theta_t - \theta_{t-1}\| (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3})
\end{aligned}$$

where the last line follows using the mclaurin's inequality. Plugging in the above upper bound to A_5 gives us

$$\begin{aligned}
A_5 &\leq \sum_{t=1}^k 2^{2n-3} (n-1) \gamma_{t-1} \|\theta_t - \theta_{t-1}\| \langle \theta_{t-1} - \theta^*, P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t) \rangle (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \gamma_{t-1} \|\theta_{t-1} - \theta^*\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \|\theta_t - \theta_{t-1}\| (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) L_{PH}^{(0)} \gamma_{t-1} \|\theta_{t-1} - \theta^*\| \|\theta_t - \theta_{t-1}\| (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} \|\theta_{t-1} - \theta^*\| (\|\theta_t - \theta^*\| + 1) (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} (\|\theta_t - \theta_{t-1}\| + \|\theta_t - \theta^*\|) (\|\theta_t - \theta^*\| + 1) (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} (\alpha L + 1) (\|\theta_t - \theta^*\| + 1) (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} (\alpha L + 1) (\|\theta_t - \theta^*\| + 1) \left(\alpha^{2n-3} L^{2n-3} (\|\theta_t - \theta^*\| + 1)^{2n-3} + \|\theta_{t-1} - \theta_t\| + \|\theta_t - \theta^*\|^{2n-3} \right) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} (\alpha L + 1) (\|\theta_t - \theta^*\| + 1) \left(\alpha^{2n-3} L^{2n-3} (\|\theta_t - \theta^*\| + 1)^{2n-3} + 2^{2n-4} (\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_t - \theta^*\|^{2n-3}) \right) \\
&\leq \sum_{t=1}^k 2^{2n-3} (n-1) \alpha L L_{PH}^{(0)} \gamma_{t-1} (\alpha L + 1) (\|\theta_t - \theta^*\| + 1) \left((2^{2n-4} + 1) \alpha^{2n-3} L^{2n-3} (\|\theta_t - \theta^*\| + 1)^{2n-3} + 2^{2n-4} \|\theta_t - \theta^*\|^{2n-3} \right)
\end{aligned}$$

where in the fourth, sixth and seventh line we used $\|\theta_t - \theta_{t-1}\| \leq \alpha L (\|\theta_t - \theta^*\| + 1)$ which we have showed

in our previous calculations. Also in the fifth and eight line, we used triangle inequality. McLaurin's inequality is another time used in eight line too.

Finally, for A_6 , we obtain

$$A_6 \leq L_{PH}^{(0)} (\gamma_0 \|\theta_0 - \theta^*\|^{2n-1} + \gamma_k \|\theta_k - \theta^*\|^{2n-1})$$

which follows from Cauchy-Schwarz inequality and ??.

References

- [1] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.
- [2] B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.