NONLINEAR MARKOVIAN STOCHASTIC APPROXIMATION

March 15, 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 $\|.\|$. 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 $\|f\| \le s \cdot \|h(\alpha,k)\|$ for some constant $s > 0$.

We use of the following iteration scheme:

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

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) $\|g^{(i)}(\theta,X) - g^{(i)}(\theta',X)\| \le L_1$, for all $\theta,\theta' \in \mathbb{R}^d, i = 0,1,2$ and $X \in \mathcal{X}$, and (2) $\|g(0,X)\| \le L_1$ 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., $\|g(\theta,X)\| \le L_1(\|\theta - \theta^*\| + 1)$ and $\|\bar{g}(\theta)\| \le L_1(\|\theta-\theta^*\|+1)$.

Assumption 2 *There exists $\mu > 0$ such that $\langle \theta - \theta', \bar{g}(\theta) - \bar{g}(\theta')\rangle \le -\mu\|\theta-\theta'\|^2, \forall \theta,\theta' \in \mathbb{R}^d$. 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 $(\xi_k)_{k\ge1}$ is a collection of i.i.d random fields satisfying the following conditions with $L_{2,p} > 0$:*

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

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

Assumption 4 *There exists a Borel measurable function $\hat{g} : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}^d$ 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 $\|\hat{g}(\theta,X)\| \le L_{PH}^{(0)}$, $\|P_\theta \hat{g}(\theta,X)\| \le L_{PH}^{(0)}$. Moreover, for $(\theta,\theta') \in \mathcal{H}^2$,*

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

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

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

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

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

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* $X \in \mathcal{X}$,

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

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

We first prove the following lemma because we are going to use that calculation in many different parts of the proof:

Lemma2. *Using Assumptions 1, 3, 5, and for sufficiently small* α *and* $t \geq 1$, *we have*

$$\mathbb{E}[\|\theta_t - \theta_{t-1}\|] \leq \alpha L\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + 1\right).$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\|\theta_t - \theta_{t-1}\|] &\leq \alpha \mathbb{E}\left[\|g(\theta_{t-1}, X_t) + \xi_t(\theta_{t-1})\|\right] \\
&\leq \alpha L_1\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + 1\right) + \alpha L_2\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + 1\right) \\
&\leq \alpha L\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + 1\right)
\end{aligned}$$

where the first line follows from ??, second line from the Lipschitzness condition and the assumption of

$$\mathbb{E}^{1/2}\left[\|\xi_t(\theta_{t-1})\|^2 | \mathcal{F}_{t-1}\right] \leq L_2\left(\mathbb{E}[\|\theta_{t-1}\|] + 1\right),$$

and the third line from ??.

For the base case analysis, we can write:

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

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

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

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

$$\begin{aligned}
&\alpha^2\left(\mathbb{E}\left[\|g\left(\theta_k,X_{k+1}\right)\|^2\right]+\mathbb{E}\left[\|\xi_{k+1}\left(\theta_k\right)\|^2\right]\right)\\
&\le\alpha^2\left(L_1^2\mathbb{E}\left[\left(\|\theta_k-\theta^*\|+1\right)^2\right]+L_2^2\mathbb{E}\left[\left(\|\theta_k-\theta^*\|+1\right)^2\right]\right)\\
&\le2\alpha^2L^2\left(\mathbb{E}\left[\|\theta_k-\theta^*\|^2\right]+1\right).
\end{aligned}$$

Therefore, we have

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

Solving this recursion gives us the following inequality:

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^2\right]\le&\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^{k+1}\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]\\
&+\sum_{t=0}^{k}\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^t2\alpha^2L^2\\
&+\sum_{t=0}^{k}2\alpha\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^{k-t}\mathbb{E}\left[\left\langle\theta_t-\theta^*,g\left(\theta_t,X_{t+1}\right)-\bar{g}\left(\theta_t\right)\right\rangle\right].
\end{aligned}$$

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

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

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

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

with

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

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 $A_1 = 0$ by taking the total expectation.

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

$$
\begin{aligned}
A_2 &\le \sum_{t=1}^{k} L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta^*\| \, \|\theta_t - \theta_{t-1}\|\right] \\
&= \sum_{t=1}^{k} \alpha L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta^*\| \, \|g(\theta_{t-1}, X_t) + \xi_t(\theta_{t-1})\|\right] \\
&\le \sum_{t=1}^{k} \alpha L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\left(\|\theta_t - \theta_{t-1}\| + \|\theta_{t-1} - \theta^*\|\right)\left(\|g(\theta_{t-1}, X_t)\| + \|\xi_t(\theta_{t-1})\|\right)\right] \\
&\le \sum_{t=1}^{k} \alpha L_{PH}^{(1)} \gamma_t \Big(L_1 \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^2\right] + \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\|\right] + \mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + \mathbb{E}\left[\|\theta_t - \theta_{t-1}\|\right]\right) \\
&\quad + \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\xi_t(\theta_{t-1})\|\right] + \mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \, \|\xi_t(\theta_{t-1})\|\right]\Big)
\end{aligned}
$$

where the second line follows from ?? and the third line follows from the triangle inequality. Now we upper the compound terms in the last line's parentheses:

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\|\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\| | \mathscr{F}_{t-1}\right]\right] \\
&\le \mathbb{E}\left[\alpha L\left(\|\theta_{t-1} - \theta^*\| + 1\right) \|\theta_{t-1} - \theta^*\|\right] \\
&\le \frac{\alpha L\left(3\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^2\right] + 1\right)}{2}
\end{aligned}
$$

where in the second line we used Lemma 2 and in the last line we used $u \le \frac{u^2+1}{2}$.

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\xi_t(\theta_{t-1})\|\right] &\le \mathbb{E}\left[\alpha\left(\|g(\theta_{t-1}, X_t)\| + \|\xi_t(\theta_{t-1})\|\right) \|\xi_t(\theta_{t-1})\|\right] \\
&\le \mathbb{E}\left[\alpha \|\xi_t(\theta_{t-1})\|^2 + \alpha L_1\left(\|\theta_{t-1} - \theta^*\| + 1\right) \|\xi_t(\theta_{t-1})\|\right] \\
&\le L\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|\right] + 1\right)
\end{aligned}
$$

where the first inequality follows from ??, second line from Lemma 2 and in the last line we used boundedness property of the noise and sufficiently small α.

$$\begin{aligned}\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\,\|\xi_t(\theta_{t-1})\|\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\,\|\xi_t(\theta_{t-1})\||\mathscr{F}_{t-1}\right]\right] \\ &\le \mathbb{E}\left[L_2\|\theta_{t-1}-\theta^*\|\left(\|\theta_{t-1}-\theta^*\|+1\right)\right] \\ &\le \frac{L_2\left(3\|\theta_{t-1}-\theta^*\|^2+1\right)}{2}\end{aligned}$$

where the second line follows from ?? and in the last line we used $u \le \frac{u^2+1}{2}$.

Summing up all these bounds, we can write for A_2:

$$\begin{aligned}A_2 &\le \sum_{t=1}^{k} \alpha L_{PH}^{(1)}\gamma_t\left(2L\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\right]+1\right)+2L\left(3\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^2\right]+1\right)+\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^2\right]\right) \\ &\le \sum_{t=1}^{k} \alpha L L_{PH}^{(1)}\gamma_t\left(8\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^2\right]+5\right)\end{aligned}$$

which in the last line we again used $u \le \frac{u^2+1}{2}$ property.

For A_3, we obtain

$$\begin{aligned}A_3 &\le \sum_{t=1}^{k} \gamma_t\mathbb{E}\left[\|\theta_t-\theta_{t-1}\|\,\|P_{\theta_{t-1}}\hat{g}(\theta_{t-1},X_t)\|\right] \\ &\le \sum_{t=1}^{k} L_{PH}^{(0)}\gamma_t\mathbb{E}\left[\|g(\theta_{t-1},X_t)+\xi_t(\theta_{t-1})\|\right] \\ &\le \sum_{t=1}^{k} L_{PH}^{(0)}\gamma_t\left(L_1\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\right]+1\right)+L_2\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\right]+1\right)\right) \\ &\le \sum_{t=1}^{k} \alpha L L_{PH}^{(0)}\gamma_t\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\right]+1\right)\end{aligned}$$

where second line follows from **??** and third line follows from ?? .

For A_4, we have

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

Finally, for A_5, we obtain

$$A_5 \le L_{PH}^{(0)}\left(\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]\right)$$

which follows from Cacuhy-Scwarz inequality and **??**.

Combining the above terms gives us:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{k}\gamma_t\left\langle\theta_t-\theta^*, g\left(\theta_t, X_{t+1}-\bar{g}(\theta_t)\right)\right\rangle\right] \le& \sum_{t=0}^{k-1}\alpha L L_{PH}^{(1)}\gamma_{t+1}\left(5+8\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]\right)+\sum_{t=0}^{k-1}\alpha L L_{PH}^{(0)}\gamma_{t+1}\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|\right]+1\right)+ \\ &\sum_{t=0}^{k-1} L_{PH}^{(0)}|\gamma_t-\gamma_{t+1}|\,\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+L_{PH}^{(0)}\left(\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]\right)\end{aligned}$$

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

$$\mathbb{E}\left[\sum_{t=0}^{k}\gamma_t\left\langle\theta_t-\theta^*,g\left(\theta_t,X_{t+1}-\bar{g}\left(\theta_t\right)\right)\right\rangle\right]\leq\sum_{t=0}^{k-1}\alpha LL_{PH}^{(1)}\gamma_{t+1}\left(5+8\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]\right)+$$
$$\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L+1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\sum_{t=0}^{k-1}\alpha LL_{PH}^{(0)}\gamma_{t+1}+L_{PH}^{(0)}\left(\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]\right)$$

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

$$\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^2\right]\leq\sum_{t=0}^{k-1}\alpha LL_{PH}^{(1)}\gamma_{t+1}\left(5+8\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]\right)+\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L+1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+L_{PH}^{(0)}\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+L_{PH}^{(0)}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]+$$
$$2\alpha^2LL_{PH}^{(0)}\left[\frac{1-\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^k}{1-\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)}\right]+\frac{\alpha L^2\left[1-\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^{k+1}\right]}{-\alpha L^2+\mu}$$

for further notation simplicity we define $c_{1,t}:=2\alpha^2LL_{PH}^{(0)}\left[\frac{1-(1-2\alpha(-\alpha L^2+\mu))^t}{1-(1-2\alpha(-\alpha L^2+\mu))}\right]+\frac{\alpha L^2\left[1-(1-2\alpha(-\alpha L^2+\mu))^{t+1}\right]}{-\alpha L^2+\mu}$ 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:

$$\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^2\right]\leq\sum_{t=0}^{k-1}\left[8\alpha LL_{PH}^{(1)}\gamma_{t+1}\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]+5\alpha LL_{PH}^{(1)}\gamma_{t+1}\right]+\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L+1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+L_{PH}^{(0)}\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+L_{PH}^{(0)}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]+c_{1,k}$$
$$=\sum_{t=0}^{k-1}8\alpha LL_{PH}^{(1)}\gamma_{t+1}\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]+5\alpha LL_{PH}^{(1)}\sum_{t=0}^{k-1}\gamma_{t+1}+\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L_1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+L_{PH}^{(0)}\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+L_{PH}^{(0)}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]+c_{1,k}$$
$$=\sum_{t=0}^{k-1}8\alpha LL_{PH}^{(1)}\gamma_{t+1}\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]+\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L+1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+L_{PH}^{(0)}\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+L_{PH}^{(0)}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]+c_{1,k}+$$
$$\frac{10\alpha^2LL_{PH}^{(1)}\left[1-\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^k\right]}{\left[1-\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\right]}$$

where the last equality follows from the definition of γ_ts. Similarly we define $c_{2,t}:=\frac{10\alpha^2LL_{PH}^{(1)}\left[1-(1-2\alpha(-\alpha L^2+\mu))^t\right]}{[1-(1-2\alpha(-\alpha L^2+\mu))]}$ for $0\leq t\leq k$. So we can write it as

$$\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^2\right]\leq\sum_{t=0}^{k-1}8\alpha LL_{PH}^{(1)}\gamma_{t+1}\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]+\sum_{t=0}^{k-1}L_{PH}^{(0)}\left(\left(\alpha L+1\right)\gamma_{t+1}-\gamma_t\right)\mathbb{E}\left[\|\theta_t-\theta^*\|\right]+$$
$$\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+L_{PH}^{(0)}\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|\right]+L_{PH}^{(0)}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|\right]+c_{1,k}+c_{2,k}$$

Now for the second term on RHS, we note that since $L \geq 2\mu$,

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

and consequently

$$\begin{aligned}
&\frac{1}{(1-2\alpha(-\alpha L^2+\mu))^k}\sum_{t=0}^{k-1} L_{PH}^{(0)}\big((\alpha L+1)\gamma_{t+1}+\gamma_t\big)\mathbb{E}\big[\|\theta_t-\theta^*\|\big] \\
&\leq 4L_{PH}^{(0)}L\alpha^2\sum_{t=0}^{k-1}\frac{1}{(1-2\alpha(-\alpha L^2+\mu))^{t+1}}\sqrt{\mathbb{E}\big[\|\theta_t-\theta^*\|^2\big]} \\
&\leq 4L_{PH}^{(0)}L\alpha^2\Big(\sum_{t=0}^{k-1}\frac{1}{(1-2\alpha(-\alpha L^2+\mu))^{t+1}}\Big)^{1/2}\Big(\sum_{t=1}^{k-1}\frac{1}{(1-2\alpha(-\alpha L^2+\mu))^{t+1}}\mathbb{E}\big[\|\theta_t-\theta^*\|^2\big]\Big)^{1/2} \\
&\leq 4L_{PH}^{(0)}L\alpha^2\cdot\sum_{t=0}^{k-1}\frac{1}{(1-2\alpha(-\alpha L^2+\mu))^{t+1}}\mathbb{E}\big[\|\theta_t-\theta^*\|^2\big]+\frac{1}{-\alpha L^2+\mu}\cdot\frac{2\alpha LL_{PH}^{(0)}}{(1-2\alpha(-\alpha L^2+\mu))^k}.
\end{aligned}$$

We also note that

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

similarly

$$\frac{\gamma_0}{(1-2\alpha(-\alpha L^2+\mu))^k}\mathbb{E}\big[\|\theta_0-\theta^*\|\big] \leq \alpha\mathbb{E}\big[\|\theta_0-\theta^*\|^2\big]+\alpha.$$

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

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

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

$$\begin{aligned}
\frac{\mathbb{E}\big[\|\theta_{k+1}-\theta^*\|^2\big]}{(1-2\alpha(-\alpha L^2+\mu))^k} \leq{}& \frac{\alpha}{\big(1-2\alpha\left(-\alpha L^2+\mu\right)\big)^k}\mathbb{E}\left[\|\theta_k-\theta^*\|^2\right]+\sum_{t=0}^{k-1}\frac{\alpha\left(8LL_{PH}^{(1)}+4\alpha\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^{-1}LL_{PH}^{(0)}\right)}{\big(1-2\alpha\left(-\alpha L^2+\mu\right)\big)^t}\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]+ \\
&\alpha\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+c'_{1,k}+c'_{2,k}+c_{3,k}.
\end{aligned}$$

for sufficiently small αs, we have

$$\alpha\left(\frac{3}{2}LL_{PH}^{(1)}+2\alpha\left(1-2\alpha\left(-\alpha L^2+\mu\right)\right)^{-1}LL_{PH}^{(0)}\right) \leq 4\alpha\left((\alpha L+1)\,L_{PH}^{(0)}+2LL_{PH}^{(1)}\right)$$

using above simplification we can rewrite our upper bound as

$$\frac{\mathbb{E}\big[\|\theta_{k+1}-\theta^*\|^2\big]}{(1-2\alpha(-\alpha L^2+\mu))^k} \leq 4\alpha\left((\alpha L+1)\,L_{PH}^{(0)}+2LL_{PH}^{(1)}\right)\sum_{t=0}^{k}\frac{\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]}{\big(1-2\alpha\left(-\alpha L^2+\mu\right)\big)^t}+\alpha\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]+c'_{1,k}+c'_{2,k}+c_{3,k}.$$

For solving the above recursion, we first define $S_t := 4\alpha\left((\alpha L+1)\,L_{PH}^{(0)}+2LL_{PH}^{(1)}\right)\sum_{l=0}^{t}\frac{\mathbb{E}[\|\theta_l-\theta^*\|^2]}{(1-2\alpha(-\alpha L^2+\mu))^l}$ for

$0 \le t \le k$. Also we use $C_t := c'_{1,t} + c'_{2,t} + c_{3,t}$ and for $0 \le t \le k$, defining constant terms. Now we can write

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

using this expansion, we should first notice that

$$\begin{aligned}
\frac{S_t}{S_{t-1}} &= \frac{S_{t-1} + 4\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right) \frac{\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right]}{\left(1 - 2\alpha\left(-\alpha L^2 + \mu\right)\right)^t}}{S_{t-1}} \\
&= 1 + 4\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right) \frac{S_{t-1} + \alpha \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right] + C_{t-1}}{S_{t-1}} \\
&\le 1 + 8\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right).
\end{aligned}$$

Now, since we have $S_0 = 4\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right) \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right]$, thus

$$S_t \le 4\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right) \left[1 + 8\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right)\right]^t \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right].$$

Substituting this upper bound into previous equations we get

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{k+1} - \theta^*\|^2\right] \le \\
&\left(1 - 2\alpha\left(-\alpha L^2 + \mu\right)\right)^k \left[4\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right)\left(1 + 8\alpha\left((\alpha L + 1) L_{PH}^{(0)} + 2LL_{PH}^{(1)}\right)\right)^k \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right] + \alpha \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right] + C_k\right]
\end{aligned}$$

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

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

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

2.2 General Case

Similar to the previous case, we first prove a useful lemma:

Lemma3. *Using Assumptions 1, 3, 5, and $m, n \ge 1$, we have*

$$\mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^m \|\theta_{t-1} - \theta^*\|^n\right] \le 2^{2m+n-2} \alpha^m L^m \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{m+n}\right] + 1\right).$$

Also when $m = 0$, this upper bound can be written as

$$\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^n\right] \le 2^{n-1}\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^n\right] + 1\right). \tag{1}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t-\theta_{t-1}\|^m\|\theta_{t-1}-\theta^*\|^n\right] &\le \alpha^m\mathbb{E}\left[\|g(\theta_{t-1},X_t)+\xi_t(\theta_{t-1})\|^m\|\theta_{t-1}-\theta^*\|^n\right]\\
&\le 2^{m-1}\alpha^m\mathbb{E}\left[\left(\|g(\theta_{t-1},X_t)\|^m+\|\xi_t(\theta_{t-1})\|^m\right)\|\theta_{t-1}-\theta^*\|^n\right]\\
&= 2^{m-1}\alpha^m\mathbb{E}\left[\mathbb{E}\left[\left(\|g(\theta_{t-1},X_t)\|^m+\|\xi_t(\theta_{t-1})\|^m\right)|\mathscr{F}_{t-1}\right]\|\theta_{t-1}-\theta^*\|^n\right]\\
&\le 2^{m-1}\alpha^m\mathbb{E}\left[\left(L_1^m\mathbb{E}\left[\left(\|\theta_{t-1}-\theta^*\|+1\right)^m\right]+L_2^m\mathbb{E}\left[\left(\|\theta_{t-1}-\theta^*\|+1\right)^m\right]\right)\|\theta_{t-1}-\theta^*\|^n\right]\\
&\le 2^{m-1}\alpha^m L^m\mathbb{E}\left[\left(\|\theta_{t-1}-\theta^*\|+1\right)^{m+n}\right]\\
&\le 2^{2m+n-2}\alpha^m L^m\left(\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^{m+n}\right]+1\right)
\end{aligned}$$

where the second and the last line follows from the mclaurin's inequality. Also the fourth line follows from ??. Proof for the other case is a trivial consequence of these calculations.

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

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

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

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.,

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

   We can further decompose it as

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

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

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

   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

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

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

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

Subsequently, we focus on analyzing T_1'; but before that, we write the general recursion of the error bound. First, we define $T_{1,t}' := \left\langle\theta_t-\theta^*, g(\theta_t, X_{t+1})-\bar{g}(\theta_t)\|\theta_t-\theta^*\|^{2(n-1)}\right\rangle$ to make T_1' dependent on the iteration index. Now, following the above decomposition and taking the expectations, we have:

$$\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^{2n}\right] \le \mathbb{E}\left[\|\theta_k-\theta^*\|^{2n}\right]+2n\alpha\mathbb{E}\left[T_{1,k}'\right]-2n\alpha\mu\mathbb{E}\left[\|\theta_k-\theta^*\|^{2n}\right]+o(\alpha) = \left(1-2n\alpha\mu\right)\mathbb{E}\left[\|\theta_k-\theta^*\|^{2n}\right]+2n\alpha\mathbb{E}\left[T_{1,k}'\right]+o(\alpha)$$

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

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

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:

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

with

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

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 $A_1 = 0$ by taking the total expectation.

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

$$
\begin{aligned}
A_2 &\le \sum_{t=1}^{k} L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|\theta_t - \theta^*\|^{2n-1}\right] \\
&\le \sum_{t=1}^{k} L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \left(\|\theta_t - \theta_{t-1}\| + \|\theta_{t-1} - \theta^*\|\right)^{2n-1}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-2} L_{PH}^{(1)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^{2n} + \|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\|^{2n-1}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-2} L_{PH}^{(1)} \gamma_t \left(2^{4n} \alpha^{2n} L^{2n} \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n}\right] + 1\right) + 2^{2n-1} \alpha L \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n}\right] + 1\right)\right) \\
&\le \sum_{t=1}^{k} 2^{4n-2} \alpha L L_{PH}^{(1)} \gamma_t \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n}\right] + 1\right)
\end{aligned}
$$

where the second line follows from triangle inequality, third line from mclaurin's inequality, fourth line from Lemma 3, and the last line for $\alpha < \frac{1}{4L}$.

For A_3, we obtain

$$
\begin{aligned}
A_3 &\le \sum_{t=1}^{k} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \, \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \, \|\theta_t - \theta^*\|^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} L_{PH}^{(0)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\| \left(\|\theta_t - \theta_{t-1}\| + \|\theta_{t-1} - \theta^*\|\right)^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-3} L_{PH}^{(0)} \gamma_t \mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^{2n-1} + \|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\|^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-3} L_{PH}^{(0)} \gamma_t \left(2^{4n-1} \alpha^{2n-1} L^{2n-1} \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right) + 2^{2n-2} \alpha L \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)\right) \\
&\le \sum_{t=1}^{k} 2^{4n-4} \alpha L L_{PH}^{(0)} \gamma_t \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)
\end{aligned}
$$

where second line follows from triangle inequality, third line from mclaurin's inequality, fourth line from Lemma 3, and the last line for $\alpha < \frac{1}{4L}$.

For A_4, we have

$$
\begin{aligned}
A_4 &\le \sum_{t=1}^{k} |\gamma_t - \gamma_{t-1}| \, \mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \|P_{\theta_{t-1}} \hat{g}(\theta_{t-1}, X_t)\| \, \|\theta_t - \theta^*\|^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \, \mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \, \|\theta_t - \theta^*\|^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \left(\|\theta_t - \theta_{t-1}\| + \|\theta_{t-1} - \theta^*\|\right)^{2(n-1)}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-3} L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^{2(n-1)} \, \|\theta_{t-1} - \theta^*\| \; + \|\theta_{t-1} - \theta^*\|^{2n-1}\right] \\
&\le \sum_{t=1}^{k} 2^{2n-3} L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \left(2^{4n-5} \alpha^{2n-2} L^{2n-2} \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right) + 2^{2n-2} \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)\right) \\
&\le \sum_{t=1}^{k} 2^{4n-4} L_{PH}^{(0)} |\gamma_t - \gamma_{t-1}| \left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)
\end{aligned}
$$

where the third line follows from triangle inequality, fourth line from mclaurin's inequality, fifth line from Lemma 3, and the last line for $\alpha < \frac{1}{4L}$.

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

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

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

$$
\begin{aligned}
A_5 &\leq \sum_{t=1}^{k} 2^{2n-3}(n-1)\gamma_{t-1}\mathbb{E}\left[\langle \theta_{t-1} - \theta^*, P_{\theta_{t-1}}\hat{g}\left(\theta_{t-1} - \theta^*, X_t\right)\rangle \|\theta_t - \theta_{t-1}\|\left(\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}\right)\right] \\
&\leq \sum_{t=1}^{k} 2^{2n-3}(n-1)\gamma_{t-1}\mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \, \|P_{\theta_{t-1}}\hat{g}(\theta_{t-1}, X_t)\| \, \|\theta_t - \theta_{t-1}\|\left(\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}\right)\right] \\
&\leq \sum_{t=1}^{k} 2^{2n-3}(n-1)L_{PH}^{(0)}\gamma_{t-1}\mathbb{E}\left[\|\theta_{t-1} - \theta^*\| \, \|\theta_t - \theta_{t-1}\|\left(\|\theta_t - \theta_{t-1}\|^{2n-3} + \|\theta_{t-1} - \theta^*\|^{2n-3}\right)\right] \\
&\leq \sum_{t=1}^{k} 2^{2n-3}(n-1)L_{PH}^{(0)}\gamma_{t-1}\mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^{2n-2} \, \|\theta_{t-1} - \theta^*\| + \|\theta_t - \theta_{t-1}\| \, \|\theta_{t-1} - \theta^*\|^{2n-2}\right] \\
&\leq \sum_{t=1}^{k} 2^{2n-3}(n-1)L_{PH}^{(0)}\gamma_{t-1}\left(2^{4n-5}\alpha^{2n-2}L^{2n-2}\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right) + 2^{2n-2}\alpha L\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)\right) \\
&\leq \sum_{t=1}^{k} 2^{4n-4}(n-1)\alpha L L_{PH}^{(0)}\gamma_{t-1}\left(\mathbb{E}\left[\|\theta_{t-1} - \theta^*\|^{2n-1}\right] + 1\right)
\end{aligned}
$$

where the fifth line follows from Lemma 3 and the last line follows for $\alpha < \frac{1}{4L}$.

Finally, for A_6, we obtain

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

which follows from Cauchy-Schwarz inequality and ??.

Combining the above terms gives us:

$$
\begin{aligned}
\sum_{t=0}^{k}\gamma_t\mathbb{E}\left[T'_{1,t}\right] &= \mathbb{E}\left[\sum_{t=0}^{k}\gamma_t\langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t)\rangle \|\theta_t - \theta^*\|^{2(n-1)}\right] \\
&\leq \sum_{t=0}^{k-1} 2^{4n-2}\alpha L L_{PH}^{(1)}\gamma_{t+1}\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n}\right] + 1\right) + \sum_{t=0}^{k-1} 2^{4n-4}\alpha L L_{PH}^{(0)}\gamma_{t+1}\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n-1}\right] + 1\right) \\
&\quad + \sum_{t=0}^{k-1} 2^{4n-4}L_{PH}^{(0)}|\gamma_{t+1} - \gamma_t|\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n-1}\right] + 1\right) + \sum_{t=0}^{k-1} 2^{4n-4}(n-1)\,\alpha L L_{PH}^{(0)}\gamma_{t+1}\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n-1}\right] + 1\right) \\
&\quad + L_{PH}^{(0)}\left(\gamma_0\|\theta_0 - \theta^*\|^{2n-1} + \gamma_k\|\theta_k - \theta^*\|^{2n-1}\right) \\
&= \sum_{t=0}^{k-1} 2^{4n-2}\alpha L L_{PH}^{(1)}\gamma_{t+1}\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n}\right] + 1\right) + \sum_{t=0}^{k-1} 2^{4n-4}L_{PH}^{(0)}\left((n\alpha L + 1)\gamma_{t+1} - \gamma_t\right)\left(\mathbb{E}\left[\|\theta_t - \theta^*\|^{2n-1}\right] + 1\right) \\
&\quad + L_{PH}^{(0)}\left(\gamma_0\|\theta_0 - \theta^*\|^{2n-1} + \gamma_k\|\theta_k - \theta^*\|^{2n-1}\right)
\end{aligned}
$$

where in the last equality we used the fact that for $\alpha \leq \frac{\mu}{L^2}$, $\gamma_t \leq \gamma_{t+1}$.

Now, if we make use of the inequality $2|x|^3 \leq x^2 + x^4$ and consolidating the terms, for sufficiently small

α, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^{2n}\right] &\leq \sum_{t=0}^{k-1} 2^{4n-5}\left[8\alpha L L_{PH}^{(1)}\gamma_{t+1} + L_{PH}^{(0)}\left((n\alpha L+1)\gamma_{t+1}-\gamma_t\right)\right]\left(\mathbb{E}\left[\|\theta_t-\theta^*\|^{2n}\right]+1\right)\\
&+\sum_{t=0}^{k-1} 2^{4n-5}L_{PH}^{(0)}\left((n\alpha L+1)\gamma_{t+1}-\gamma_t\right)\left(\mathbb{E}\left[\|\theta_t-\theta^*\|^{2(n-1)}\right]+1\right)+\frac{L_{PH}^{(0)}}{2}\left(\gamma_0\|\theta_0-\theta^*\|^{2(n-1)}+\gamma_k\|\theta_k-\theta^*\|^{2(n-1)}\right)\\
&+\left(\frac{L_{PH}^{(0)}}{2}+1\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\frac{L_{PH}^{(0)}}{2}\gamma_k\|\theta_k-\theta^*\|^{2n}+o(\alpha)\\
&\leq \sum_{t=0}^{k-1} 2^{4n-4}\alpha L\gamma_{t+1}\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]\mathbb{E}\left[\|\theta_t-\theta^*\|^{2n}\right]+\sum_{t=0}^{k-1}2^{4n-4}n\alpha L L_{PH}^{(0)}\mathbb{E}\left[\|\theta_t-\theta^*\|^{2(n-1)}\right]\\
&+\frac{L_{PH}^{(0)}}{2}\left(\gamma_0\|\theta_0-\theta^*\|^{2(n-1)}+\gamma_k\|\theta_k-\theta^*\|^{2(n-1)}\right)+\left(\frac{L_{PH}^{(0)}}{2}+1\right)\gamma_0\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\frac{L_{PH}^{(0)}}{2}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|^{2n}\right]+\tilde{c}_{1,k}+o(\alpha)
\end{aligned}
$$

where the last inequality follows for $(n\alpha L+1)\gamma_{t+1}-\gamma_t \leq 2n\alpha L\gamma_{t+1}$ and also with

$$
\tilde{c}_{1,t} := \sum_{l=0}^{t-1} 2^{4n-4}n\alpha L\gamma_{t+1}\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]+\sum_{l=0}^{t-1}2^{4n-4}n\alpha L L_{PH}^{(0)}
$$

being defined for $1 \leq t \leq k-1$. It can be seen that for sufficiently small α, $\tilde{c}_{1,t}$ would be of order $4^n n\alpha L\left(nL_{PH}^{(0)}+4\alpha L_{PH}^{(1)}\right)$. Now from induction hypothesis we know that

$$
\mathbb{E}\left[\|\theta_t-\theta^*\|^{2(n-1)}\right] \leq \tilde{C}_{n-1,1}\cdot\left(1-2n\alpha\mu\right)^t\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\tilde{C}_{n-1,2}\cdot 4^{n-1}(n-1)\alpha L\left((n-1)L_{PH}^{(0)}+2\alpha L_{PH}^{(1)}\right).
$$

Thus, if we plug in this upper bound into the previous inequality, for $\alpha < \frac{1}{4^n}$ we will have

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^{2n}\right] &\leq \sum_{t=0}^{k-1}2^{4n-4}\alpha L\gamma_{t+1}\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]\mathbb{E}\left[\|\theta_t-\theta^*\|^{2n}\right]\\
&+\frac{L_{PH}^{(0)}}{2}\gamma_k\mathbb{E}\left[\|\theta_k-\theta^*\|^{2n}\right]+\tilde{c}_{2,k}\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\tilde{c}_{1,k}+\tilde{c}_{3,k}+o(\alpha)
\end{aligned}
$$

in which $\tilde{c}_{3,k}$ would be of order $4^n n\alpha L\left(nL_{PH}^{(0)}+4\alpha L_{PH}^{(1)}\right)$ and $\tilde{c}_{2,k}$ is an $\mathcal{O}(1)$ constant dependent to k.

To solve the above recursion, we define $\tilde{S}_t := \frac{L_{PH}^{(0)}}{2}\gamma_t\mathbb{E}\left[\|\theta_t-\theta^*\|^{2n}\right]+\sum_{l=0}^{t-1}2^{4n-4}\alpha L\gamma_{l+1}\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]\mathbb{E}\left[\|\theta_l-\theta^*\|^{2n}\right]$ for $1 \leq t \leq k$. Using this we can write

$$
\mathbb{E}\left[\|\theta_{t+1}-\theta^*\|^{2n}\right] \leq \tilde{S}_t+\tilde{c}_{2,t}\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\tilde{c}_{1,t}+\tilde{c}_{3,t}+o(\alpha).
$$

Now notice that we have(recall that $\gamma_t = 2n\alpha\left(1-2n\alpha\mu\right)^{k-t}$.)

$$
\begin{aligned}
\frac{\tilde{S}_t}{\tilde{S}_{t-1}} &\leq \frac{\tilde{S}_{t-1}+\left(2^{4n-4}\alpha L\gamma_t\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]-\frac{L_{PH}^{(0)}}{2}\gamma_{t-1}\right)\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^{2n}\right]+\frac{L_{PH}^{(0)}}{2}\gamma_t\mathbb{E}\left[\|\theta_t-\theta^*\|^{2n}\right]}{\tilde{S}_{t-1}}\\
&= \frac{\tilde{S}_{t-1}+\left(2^{4n-4}\alpha L\gamma_t\left[4L_{PH}^{(1)}+nL_{PH}^{(0)}\right]-\frac{L_{PH}^{(0)}}{2}\gamma_{t-1}\right)\mathbb{E}\left[\|\theta_{t-1}-\theta^*\|^{2n}\right]+\frac{L_{PH}^{(0)}}{2}\gamma_t\left[\tilde{S}_{t-1}+\tilde{c}_{2,t-1}\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\tilde{c}_{1,t}+\tilde{c}_{3,t}+o(\alpha)\right]}{\tilde{S}_{t-1}}
\end{aligned}
$$

it is not very hard to see that the if we take α small enough above fraction could get as close as possible to 1. This means that we can write our final error bound as the following

$$\mathbb{E}\left[\|\theta_{k+1}-\theta^*\|^{2n}\right] \leq \tilde{C}_{n,1}\cdot\left(1-2n\alpha\mu\right)^{k+1}\mathbb{E}\left[\|\theta_0-\theta^*\|^{2n}\right]+\tilde{C}_{n,2}\cdot 4^n n\alpha L\left(nL_{PH}^{(0)}+4\alpha L_{PH}^{(1)}\right)$$

in which $\tilde{C}_{n,1}$ and $\tilde{C}_{n,2}$ are $\mathcal{O}(1)$ constants and also α has been taken small enough.

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.