# NONLINEAR MARKOVIAN STOCHASTIC APPROXIMATION

**October 7, 2024**

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 $\alpha$ 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:

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

## 1.1 Assumptions

**Assumption 1** *For each $X \in \mathcal{X}$, the function $g(\theta,X)$ is three times continuously differentiable in $\theta$ 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 $\theta$ uniformly in $X$. The above assumption immediately implies that the growth of $||g||$ and $||\bar{g}||$ will be at most linear in $\theta$, 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 $\theta^*$.*

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 $L \ge 1$.

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

$$\hat{g}(\theta,X) - P_\theta \hat{g}(\theta,X) = g(\theta,X) - \bar{g}(\theta). \tag{2}$$

**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 $(\theta,\theta') \in \mathcal{H}^2$,*

$$\sup_{X\in\mathcal{X}} ||P_\theta \hat{g}(\theta,X) - P_{\theta'}\hat{g}(\theta',X)|| \le L_{PH}^{(1)}||\theta-\theta'||. \tag{3}$$

**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*

$$\sup_{\theta \in \mathbb{R}^d, X \in \mathscr{X}} ||P_\theta^n(X,.) - \pi_\theta(.)||_{TV} \leq \rho^n K_P, \tag{4}$$

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

$$||\hat{g}(\theta, X)|| \leq \frac{\sigma K_P}{1-\rho}, \tag{5}$$

$$||P_\theta \hat{g}(\theta, X)|| \leq \frac{\sigma \rho K_P}{1-\rho}. \tag{6}$$

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

$$||P_\theta \hat{g}(\theta, X) - P_{\theta'} \hat{g}(\theta', X)|| \leq L_{PH}^{(1)} ||\theta - \theta'||, \tag{7}$$

*where*

$$L_{PH}^{(1)} = \frac{K_P^2 \sigma L_P}{(1-\rho)^2}(2 + K_P) + \frac{K_P}{1-\rho} L_H. \tag{8}$$

Proof of this lemma can be found in [1], Lemma 7.

# 2 Error Bound

General comments about writing math:

- Use macros as much as possible, including symbols appearing throughout the proof such as $L_{PH}^{(1)}, L_{PH}^{(0)}, \theta^*, \alpha$ (this makes life easier when we want to change notation), and basic symbols $\|\cdot\|, \mathbb{E}$, etc. Do not use $||$, use $\|$ instead.
- Try to avoid extremely long equations. This proof is not that complicated. Simplify things in the middle as much as you can as we don't care about universal constant factors. If the calculation has to be that complicated, break it down into several equations.
- Use lemmas to encapsulate your intermediate results. Structure the proofs as a tree (e.g. the proof of main theorem involves Lemmas 1,2,3. You state the lemmas in the middle of the proof, and put the proof of lemmas at the end of the theorems.)
- Use align environment instead of equation so that the numbering is for each line. Number it only when it's going to be referenced. Otherwise use align*.

## 2.1 Base Case

For the base case analysis, we can write:

$$\begin{aligned}
&\mathbb{E}\left[||\theta_{k+1} - \theta^*||^2\right] - \mathbb{E}\left[||\theta_k - \theta^*||^2\right] = \\
&2\alpha \mathbb{E}\left[\langle \theta_k - \theta^*, g(\theta_k, X_{k+1})\rangle\right] + \alpha^2 \mathbb{E}\left[||g(\theta_k, X_{k+1})||^2\right] + \alpha^2 \mathbb{E}\left[||\xi_{k+1}(\theta_k)||^2\right] = \\
&2\alpha \mathbb{E}\left[\langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k)\rangle\right] + 2\alpha \mathbb{E}\left[\langle \theta_k - \theta^*, \bar{g}(\theta_k)\rangle\right] + \alpha^2 \mathbb{E}\left[||g(\theta_k, X_{k+1})\right] + \alpha^2 \mathbb{E}\left[||\xi_{k+1}(\theta_k)||^2\right].
\end{aligned} \tag{9}$$

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

$$\langle \theta_k - \theta^*, \bar{g}(\theta_k) \rangle = \langle \theta_k - \theta^*, \bar{g}(\theta_k) + \bar{g}(\theta^*) \rangle \le -\mu ||\theta_k - \theta^*||^2. \tag{10}$$

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

$$\begin{aligned}
&\alpha^2 \left( \mathbb{E}\left[ ||g(\theta_k, X_{k+1})||^2 \right] + \mathbb{E}\left[ ||\xi_{k+1}(\theta_k)||^2 \right] \right) \\
&\le \alpha^2 \left( L_1^2 \mathbb{E}\left[ \left( ||\theta_k - \theta^*|| + 1 \right)^2 \right] + L_2^2 \mathbb{E}\left[ \left( ||\theta_k - \theta^*|| + 1 \right)^2 \right] \right) \\
&\le 2\alpha^2 L^2 \left( \mathbb{E}\left[ ||\theta_k - \theta^*||^2 \right] + 1 \right).
\end{aligned} \tag{11}$$

Therefore, we have

$$\mathbb{E}\left[ ||\theta_{k+1} - \theta^*||^2 \right] \le \left( 1 - 2\alpha \left( \alpha L^2 + \mu \right) \right) \mathbb{E}\left[ ||\theta_k - \theta^*||^2 \right] + 2\alpha^2 L^2 + 2\alpha \mathbb{E}\left[ \langle \theta_k - \theta^*, g(\theta_k, X_{k+1}) - \bar{g}(\theta_k) \rangle \right] \tag{12}$$

Solving this recursion gives us the following inequality:

$$\begin{aligned}
\mathbb{E}\left[ ||\theta_{k+1} - \theta^*||^2 \right] \le & \left( 1 - 2\alpha \left( \alpha L^2 + \mu \right) \right)^{k+1} \mathbb{E}\left[ ||\theta_0 - \theta^*||^2 \right] \\
& + \sum_{t=0}^{k} \left( 1 - 2\alpha \left( \alpha L^2 + \mu \right) \right)^t 2\alpha^2 L^2 \\
& + \sum_{t=0}^{k} 2\alpha \left( 1 - 2\alpha \left( \alpha L^2 + \mu \right) \right)^{k-t} \mathbb{E}\left[ \langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle \right].
\end{aligned} \tag{13}$$

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 which is equal to $2\alpha^2 L^2 \left( \alpha L^2 + \mu \right)^k$.
Now, we can upper bound the third summand using below decomposition:

$$\mathbb{E}\left[ \sum_{t=0}^{k} \gamma_t \langle \theta_t - \theta^*, g(\theta_t, X_{t+1}) - \bar{g}(\theta_t) \rangle \right] = \mathbb{E}\left[ A_1 + A_2 + A_3 + A_4 + A_5 \right] \tag{14}$$

with

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

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

$$\begin{aligned}A_2 &\leq \sum_{t=1}^{k} L_{PH}^{(1)}\gamma_t ||\theta_t - \theta^*||\,||\theta_t - \theta_{t-1}|| \\ &= \sum_{t=1}^{k} \alpha L_{PH}^{(1)}\gamma_t ||\theta_t - \theta^*||\,||g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)|| \\ &\leq \sum_{t=1}^{k} L_{PH}^{(1)}\gamma_t ||\theta_t - \theta^*|| \left(\alpha L_1\left(||\theta_t - \theta^*|| + 1\right) + \alpha L_2\left(||\theta_t - \theta^*|| + 1\right)\right) \\ &\leq \sum_{t=1}^{k} \frac{L_{PH}^{(1)}\gamma_t}{2}(1 + \alpha L)\left(1 + 3||\theta_t - \theta^*||^2\right)\end{aligned} \tag{15}$$

The second last step seems wrong and/or loose. You have an $\alpha$ factor there and you should keep it. Seems to me that it should be something like

$$\alpha\gamma_t\|\theta_t - \theta^*\|\Big(L_1(\|\theta_t - \theta^*\| + 1) + L_2(\|\theta_t - \theta^*\| + 1)\Big)$$

If we use your bound for $A_2$, the recursion argument cannot go through. But if we keep the $\alpha$ factor there, it will work.

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

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

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

For $A_3$, we obtain

$$\begin{aligned}A_3 &\leq \sum_{t=1}^{k} \gamma_t ||\theta_t - \theta_{t-1}||\,||P_{\theta_{t-1}}\hat{g}(\theta_{t-1}, X_t)|| \\ &\leq \sum_{t=1}^{k} \alpha L_{PH}^{(0)}\gamma_t ||g(\theta_t, X_{t+1}) + \xi_{t+1}(\theta_t)|| \\ &\leq \sum_{t=1}^{k} L_{PH}^{(0)}\gamma_t \left(\alpha L_1\left(||\theta_t - \theta^*|| + 1\right) + \alpha L_2\left(||\theta_t - \theta^*|| + 1\right)\right) \\ &\leq \sum_{t=1}^{k} \alpha L L_{PH}^{(0)}\gamma_t\left(||\theta_t - \theta^*|| + 1\right)\end{aligned} \tag{16}$$

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

$$\begin{aligned}A_4 &\leq \sum_{t=1}^{k} |\gamma_t - \gamma_{t-1}|\,||\theta_{t-1} - \theta^*||\,||P_{\theta_{t-1}\hat{g}(\theta_{t-1}, X_t)}|| \\ &\leq \sum_{t=1}^{k} L_{PH}^{(0)}|\gamma_t - \gamma_{t-1}|\,||\theta_{t-1} - \theta^*||\end{aligned} \tag{17}$$

Finally, for $A_5$, we obtain

$$A_5 \leq L_{PH}^{(0)}\left(\gamma_0||\theta_0 - \theta^*|| + \gamma_k||\theta_k - \theta^*||\right) \tag{18}$$

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

Combining the above terms and taking expectations, gives us:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{k}\gamma_t\left\langle\theta_t-\theta^*, g\left(\theta_t, X_{t+1}-\bar{g}\left(\theta_t\right)\right)\right\rangle\right] \leq{}& \sum_{t=1}^{k}\frac{L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\left(1+3\mathbb{E}\left[||\theta_t-\theta^*||^2\right]\right)+\sum_{t=1}^{k}\alpha L L_{PH}^{(0)}\gamma_t\left(\mathbb{E}\left[||\theta_t-\theta^*||\right]+1\right)+\\ &\sum_{t=0}^{k-1}L_{PH}^{(0)}|\gamma_t-\gamma_{t+1}|\,\mathbb{E}\left[||\theta_t-\theta^*||\right]+L_{PH}^{(0)}\left(\gamma_0\mathbb{E}\left[||\theta_0-\theta^*||\right]+\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]\right)\end{aligned} \tag{19}$$

now it should be noticed that as long as we have $\alpha \leq \frac{\sqrt{2\mu^2+4L^2}-\mu}{2L^2}$, we have $\gamma_{t+1} \leq \gamma_t$ Seems wrong direction. Thus, we can simplify the above upper bound and write it this way:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{k}\gamma_t\left\langle\theta_t-\theta^*, g\left(\theta_t, X_{t+1}-\bar{g}\left(\theta_t\right)\right)\right\rangle\right] \leq{}& \sum_{t=1}^{k}\frac{L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\left(1+3\mathbb{E}\left[||\theta_t-\theta^*||^2\right]\right)+\\ &\sum_{t=1}^{k-1}L_{PH}^{(0)}\left((\alpha L+1)\gamma_t-\gamma_{t+1}\right)\mathbb{E}\left[||\theta_t-\theta^*||\right]+\\ &\sum_{t=1}^{k}\alpha L L_{PH}^{(0)}\gamma_t+L_{PH}^{(0)}\left((2\gamma_0-\gamma_1)\mathbb{E}\left[||\theta_0-\theta^*||\right]+(\alpha L+1)\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]\right)\end{aligned} \tag{20}$$

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

$$\begin{aligned}\mathbb{E}\left[||\theta_{k+1}-\theta^*||^2\right] \leq{}& \sum_{t=1}^{k}\frac{L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\left(1+3\mathbb{E}\left[||\theta_t-\theta^*||^2\right]\right)+\sum_{t=1}^{k-1}L_{PH}^{(0)}\left((\alpha L+1)\gamma_t-\gamma_{t+1}\right)\mathbb{E}\left[||\theta_t-\theta^*||\right]+\\ &\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)\gamma_0\mathbb{E}\left[||\theta_0-\theta^*||^2\right]+L_{PH}^{(0)}\left(2\gamma_0-\gamma_1\right)\mathbb{E}\left[||\theta_0-\theta^*||\right]+(\alpha L+1)\,L_{PH}^{(0)}\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]+\\ &\left(\frac{L_{PH}^{(0)}}{L}+1\right)\frac{\alpha L^2\left(1-\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k\right)}{\left(\alpha L^2+\mu\right)}+\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k 2\alpha^2L^2\end{aligned} \tag{21}$$

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}\left[||\theta_{k+1}-\theta^*||^2\right] \le{}& \sum_{t=1}^{k}\left[\frac{3L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\mathbb{E}\left[||\theta_t-\theta^*||^2\right]+\frac{L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\right]+\sum_{t=1}^{k-1}L_{PH}^{(0)}\left((\alpha L+1)\gamma_t-\gamma_{t+1}\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)}\left(2\gamma_0-\gamma_1\right)\mathbb{E}\left[||\theta_0-\theta^*||\right]+(\alpha L+1)L_{PH}^{(0)}\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]+\\
&\left(\frac{L_{PH}^{(0)}}{L}+1\right)\frac{\alpha L^2\left(1-\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k\right)}{\left(\alpha L^2+\mu\right)}+\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k 2\alpha^2L^2\\
={}&\sum_{t=1}^{k}\frac{3L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\mathbb{E}\left[||\theta_t-\theta^*||^2\right]+\frac{L_{PH}^{(1)}}{2}(1+\alpha L)\sum_{t=1}^{k}\gamma_t+\sum_{t=1}^{k-1}L_{PH}^{(0)}\left((\alpha L+1)\gamma_t-\gamma_{t+1}\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)}\left(2\gamma_0-\gamma_1\right)\mathbb{E}\left[||\theta_0-\theta^*||\right]+(\alpha L+1)L_{PH}^{(0)}\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]+\\
&\left(\frac{L_{PH}^{(0)}}{L}+1\right)\frac{\alpha L^2\left(1-\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k\right)}{\left(\alpha L^2+\mu\right)}+\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k 2\alpha^2L^2\\
={}&\sum_{t=1}^{k}\frac{3L_{PH}^{(1)}\gamma_t}{2}(1+\alpha L)\mathbb{E}\left[||\theta_t-\theta^*||^2\right]+\sum_{t=1}^{k-1}L_{PH}^{(0)}\left((\alpha L+1)\gamma_t-\gamma_{t+1}\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)}\left(2\gamma_0-\gamma_1\right)\mathbb{E}\left[||\theta_0-\theta^*||\right]+(\alpha L+1)L_{PH}^{(0)}\gamma_k\mathbb{E}\left[||\theta_k-\theta^*||\right]+\\
&\left(\frac{L_{PH}^{(0)}}{L}+1\right)\frac{\alpha L^2\left(1-\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k\right)}{\left(\alpha L^2+\mu\right)}+\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k 2\alpha^2L^2+\\
&\frac{L_{PH}^{(1)}(1+\alpha L)\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)\left[1-\left(1-2\alpha\left(\alpha L^2+\mu\right)\right)^k\right]}{4\alpha\left(\alpha L^2+\mu\right)}
\end{aligned}\tag{22}$$

Where does the last term in the last line come from? This term seems very large and will ruin your bound...

where the last equality follows from the definition of $\gamma_t$'s.

BEGIN – Wenlong edits

For the second term on RHS, we note that

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

and consequently

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

We also note that

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

Substituting back and rearranging yields

$$\begin{aligned}\frac{\mathbb{E}[\|\theta_{k+1}-\theta^*\|^2]}{(1-2\alpha(\alpha L^2+\mu))^k} &\leq \left\{6\alpha L_{PH}^{(1)}(1+\alpha L)+6\alpha^2 L_{PH}^{(0)} L\right\}\sum_{t=1}^{k}\frac{\mathbb{E}[\|\theta_t-\theta^*\|^2]}{(1-2\alpha(\alpha L^2+\mu))^{t-1}}+2\alpha L_{PH}^{(0)}\frac{\mathbb{E}[\|\theta_k-\theta^*\|^2]}{(1-2\alpha(\alpha L^2+\mu))^{k-1}}\\ &+\frac{2L_{PH}^{(0)}L(\alpha L^2+\mu)^{-1}+L_{PH}^{(0)}(1+\alpha L)}{(1-2\alpha(\alpha L^2+\mu))^k}\alpha+\left(\frac{L_{PH}^{(0)}}{L}+1\right)\frac{L^2}{(\alpha L^2+\mu)}\alpha+2\alpha^2L^2+3\alpha\mathbb{E}[\|\theta_0-\theta^*\|^2]+3\alpha\left(L_{PH}^{(0)}\right)^2.\end{aligned}$$

The factor in the first term of the first parenthesis should also be some constant times $\alpha^2$. See my comments on the bound for $A_2$. Using the corrected bound you'll be able to solve this recursion. Please fix this and make necessary downstream edits.

END – Wenlong edits

## 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 $||\theta_{k+1}-\theta^*||^{2n}$

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

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 $\alpha$, i.e., $\alpha 2n\left\langle\theta_k-\theta^*, g\left(\theta_k, X_{k+1}\right)+\xi_{k+1}\left(\theta_k\right)\right\rangle^j ||\theta_k-\theta^*||^{2(n-1)}$. We can further decompose it as

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

   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

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

3. For the remaining terms, we see that they are of higher orders of $\alpha$. Therefore, when $\alpha$ 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

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

Subsequently, we focus on analyzing $T_1'$.

# References

[1] 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.