7.1 Asymptotic convergence

For most algorithms, since the policy gradient method is gradient-based, one cannot expect to prove convergence to a globally optimal policy (I am not sure if the linear regression in training the ESN for function approximation would change this or not). The best that one could hope for is the convergence of the partial differentiation of value function approximation w.r.t the actor parameter to 0.

Many analyses of convergence are based on the TD (time difference methods) update of the critic $V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$.(see the convergence analysis in Q-learning algorithm, also see this link) The following are the most general analysis from **Policy Gradient Methods for Reinforcement Learning with Function Approximation**(this is a link), regardless of the update methods and the step length

Theorem 5. *Let π and h_w be any differentiable function approximators for the policy and advantage function respectively that satisfy $\max_{\theta,s,a,i,j} |\frac{\partial^2 \mu^\theta(s)}{\partial \theta^2}| < B < \infty$. Let$\{\alpha_k\}_{k=0}^{\infty}$ be any step-size sequence such that $\lim_{k\to\infty} \alpha_k = 0$ and $\sum_k \alpha_k = \infty$. Then, for any MDP with bounded rewards, the iteration :*

$$\begin{aligned} w_k = &\quad w \text{ such that with parameterization } w,\ h_w \text{ satisfy condition in Prop 2} \\ \theta_{k+1} = &\quad \theta_k + \alpha_k E_{\pi\sim\mu^\theta}\left[\nabla_\theta log\ \mu^\theta h_{w_k}(s,a)\right] \end{aligned}$$

converges such that $\frac{\partial h}{\partial \theta} = 0$

This convergence is obvious since from Prop 2 we can see its gradient ascend methods to maximize the value function by adjusting the policy's parameter. A more detailed proof is said to be found at some book at 1996 but I couldn't find the resources.

7.2 Non-asymptotic convergence

l

a

7.3 Non-asymptomatic Analysis of ESN Critic

Definition 2 (Expected Bellman Loss and Empirical Bellman loss)**.** *For an infinite sequence Z=$z_t(eg: Z in section 1), define the **Expected Bellman Loss** of the model $h_\theta \in \mathcal{H}$ as*

$$\begin{aligned} L(\theta) &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \|h_\theta(z_t) - \gamma h_\theta(z_{t+1}) - \mathcal{R}(z_t)\|_2^2 \\ &:= \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} l(z_t, \theta) \end{aligned}$$

*In real case, for a truncated finite sequence with length p, we have the **Empirical Bellman Loss***

$$\hat{L}(\theta) = \frac{1}{p} \sum_{t=0}^{p-1} \left\|h_\theta(T^t(z)) - \gamma h_\theta(T^{t+1}(z)) - \mathcal{R}(z)\right\|_2^2$$

To show the empirical loss's dependence on the sequence length, we could also write $\hat{L}(\theta)$ as $L_p(\theta)$

As the Bellman loss comes from the Bellman Residue $\|H - \Phi H\|_2^2$, ,the expected risk minimizer θ^* is exactly the parameter of the real optimal Q-function in the MDP/POMDP context. Set $\hat{\theta}$ as the minimizer of $\hat{L}$ obtained by regression. (Here for simplicity we omit the regularization term of ridge regression). Then

$$\begin{aligned} L(\hat{\theta}) - L(\theta^*) &\leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + \hat{L}(\hat{\theta}) - \hat{L}(\theta^*) + |\hat{L}(\theta^*) - L(\theta^*)| \\ &\leq |L(\hat{\theta}) - \hat{L}(\hat{\theta})| + 0 + |\hat{L}(\theta^*) - L(\theta^*)| \\ &\leq 2 \sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|. \end{aligned}$$

Here Θ comes from the uniform convergence such that

$$\Pr\left[|\hat{L}(\theta) - L(\theta)| \geq \varepsilon\right] \leq \delta; , \forall \theta \in \Theta$$

Assumption 1. *For the Critic model we assume:*

- ***Lipchitz Continuity:*** *$L(\theta)$ is Lipchitz continuous w.r.t θ, $|L(\theta) - L(\theta')| \leq \kappa_C \|\theta - \theta'\|$*
- ***Bounded Parameter:*** *The L2 norm of the parameter θ is bounded, i.e the hypothesis class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq M\}$ Besides, we take M such that $\kappa_C M = 1$*
- ***Bounded Loss:*** *$0 \leq l \leq 1$.*
- ***Expectation of l(z):*** *For $\mathcal{F}_m = \sigma(z_1, z_2, \ldots, z_m)$, $E[l(z_{n+1}) - L(\theta)|\mathcal{F}_n] = 0$*
- ***Unbiased estimation:*** *$\exists \theta^*$ such that $E(h_{\theta^*}(z)) = Q^*(z)$*

For Assumption 4, Notice that by Birkhoff Ergodic Theorem, $L(\theta) = \mathbb{E}[l(\theta)\,|\mathcal{I}]$ where $\mathcal{I}$ is the sub σ-algebra consists of all T-invariant set.

For finite hypothesis class we have the following result:

Theorem 6 (Convergence Rate with Finite Hypothesis Class)**.** *Suppose that the hypothesis class H is finite $\forall \delta$ s. t. $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$, we have*

$$|L(\theta) - L_n(\theta)| \leq \sqrt{\frac{2(\ln |\mathcal{H}| + \ln (2/\delta))}{n}} \quad \forall h_\theta \in \mathcal{H}.$$

Proof. Recall $l(\theta, z) = \|h_\theta(z) - \gamma h_\theta(T(z)) - \mathcal{R}(z)\|_2^2$ Notice that the input data points are not independently and identically distributed, so instead of the usual concentration inequality, we consider here the construction of a martingale in order to use the Azuma-Hoeffding inequality.

Claim: $\mathcal{F}_m = \sigma(z_1, z_2, \ldots, z_m)$ Then given θ such that $h_\theta \in \mathcal{H}$ let $X_n := n(L_n(\theta) - L(\theta))$ then $\{X_n\}$ is a martingale with respect to $\mathcal{F}_m$.

To prove the claim, we notice that l is bounded and $\lim_{n\to\infty} X_n = 0$ for any fixed θ so $E|X_n| < \infty$. And by assumption 4, we have $E[X_{n+1}|z_0, \ldots, z_n] = X_n, \ \forall n \geq 0$

Now applying Azuma-Hoeffding inequality to $\{X_n\}$ we have:

$$\Pr\left(|X_n - X_0| \geq \epsilon\right) = \Pr\left(|X_n| \geq \epsilon\right) \leq 2 \exp\left(-\frac{\epsilon^2}{2n}\right)$$

$$\Pr\left(|L_n(\theta) - L(\theta)| \geq \epsilon\right) = \Pr\left(|X_n| \geq n\epsilon\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

$$\begin{aligned}\Pr\left(\exists\theta \text{ s.t. } |L_n(\theta) - L(\theta)| \geq \epsilon\right) &\leq \sum_{h_\theta \in \mathcal{H}} \Pr\left(|L_n(\theta) - L(\theta)| \geq \epsilon\right) \\ &\leq \sum_{h_\theta \in \mathcal{H}} 2\exp\left(-\frac{n\epsilon^2}{2}\right) \\ &= 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2}\right).\end{aligned}$$

if we take $\delta = 2|\mathcal{H}|\exp(-2n\epsilon^2)$ then

$$\epsilon = \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}$$

□

Now since in the case of ESN, the hypothesis class is infinite, we will take a detour to the concept of ϵ-net:

Definition 3 (ϵ-net)**.** *Let $\epsilon > 0$ An ϵ-net of a set S with respect to a distance metric ρ is a subset $C \subseteq S$ such that*

$$S \subseteq \bigcup_{x \in C} B_x(\epsilon), \quad \textit{where}$$

$$B_x(\epsilon) \triangleq \{x' : \rho(x, x') \leq \epsilon\}.$$

Proposition 3 (The ϵ-net of L2 Ball)**.** *Let $B,\epsilon \geq 0$ and let $S = \{x \in \mathbb{R}^p : \|x\|_2 \leq B\}$ Then there exists an ϵ-net of S (where ρ is the metric induced by L2 norm) with at most $\left(\frac{3M}{\epsilon}\right)^p$ elements.*

Proof. Set N_ϵ as the maximal ϵ-separated set, which means $d(x, y) \geq \varepsilon$ for all $x, y \in \mathcal{N}_\varepsilon$, $x \neq y$. By definition this is an ϵ-net otherwise it contradict with maximality.

While via the triangle inequality that the balls of radius $\varepsilon/2$ centered at the points in $\mathcal{N}_\varepsilon$ are disjoint, they are all contained in

$$\tilde{S} = \{x \in \mathbb{R}^p : \|x\|_2 \leq M + \frac{\epsilon}{2}\}$$

This indicates that considering the volume:

$$|N_\epsilon|(\frac{\epsilon}{2})^p < (M + \frac{\epsilon}{2})^p$$

Then

$$|N_\epsilon| < (1 + \frac{2}{\epsilon})^p < \left(\frac{3M}{\epsilon}\right)^p$$

□

Proposition 4. *For a functional H, the minimization of Bellman residue $\|H - \Phi H\|_2^2$ and the approximation of optimal Q-value functional $\|H - Q^*\|_2^2$ converge at the same rate.*

Proof. Recall that the optimal Bellman operator Φ is a strictly contracting map w.r.t superemum norm when $\gamma<1$:

$$\|H - \Phi H\|_\infty = \|(H - Q^*) + (\Phi Q^* - \Phi H)\|_\infty < (1 + \gamma)\|H - \Phi H\|_\infty$$

Similarly,

$$\|H - \Phi H\|_\infty > (1-\gamma)\|H - Q^*\|_\infty > 0$$

Thus the two converge at the same rate (The same convergence rate w.r.t infinity norm actually does not guarantee the same convergence rate w.r.t infinity norm. However, for simplicity, we assume the second claim stands) □

Now we coclude this section with the main theorem:

Theorem 7 (Convergence Rate of ESN-Critic)**.** *For an admissible input process $\mathcal{Z}$, consider the following setup:*

- *Randomly generate $A, C, \boldsymbol{\zeta}$ using a specified procedure (to be detailed in Procedure 1).*
- *Define the hypothesis class $\mathcal{H}$ as follows:*

$$\mathcal{H} = \left\{ h_W(z) := W^T \sigma(Ax + Cz + \zeta) \mid \|W\|_2 \leq M, W \in \mathcal{R}^p \right\}$$

  where W is a parameter vector, σ is an activation function, and M is a bound on the L_2 norm of W. Assume there exit the true estimation W^*

- *For each roll out, the length of the trajectory (the sample collected) is bounded. The parameter W is updated once per roll out. That is when updated k times, we have W_k as the minimizer of $L_n(W)$ or $\hat{L}(W)$ where $n/k \sim \mathcal{O}(1)$ For simplicity, set n=k.*

Suppose the Bellman loss generated by models in $\mathcal{H}$ satisfies conditions in Assumptions 1 . By minimizing the empirical Bellman loss with linear regression k times and obtaining the minimizer of L_k(i.e. $\hat{L}$) W_k, we have the following convergence result: With probability at least $1 - \mathcal{O}(e^{-p})$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \left\| H_{W_k}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(z_t) - H_{W^*}^{\boldsymbol{A},\boldsymbol{C},\zeta}(z_t) \right\|_2^2 \leq \mathcal{O}(\sqrt{\frac{\log k}{k}})$$

or equivalently

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \left\| H_{W_k}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(z_t) - H_{W^*}^{\boldsymbol{A},\boldsymbol{C},\zeta}(z_t) \right\|_2^2 \leq \tilde{\mathcal{O}}(k^{\frac{1}{2}})$$

where: $H_W^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(z) := W^T \sigma(Ax + Cz + \zeta)$

Proof. By Proposition 4, we only need to prove

$$L(W_k) - L(W^*) \leq 2 \sup_{W \in S} |L(W) - \hat{L}(W)| \leq \mathcal{O}(\sqrt{\frac{\log k}{k}})$$

To prove this, fix parameter$\delta, \epsilon > 0$, by proposition 3, let C be the ϵ-net of parameter space $S = \{w : \|W\|_2 \leq M\}$ w.r.t l2 norm. Define $E = \left\{ \forall W \in C, |\hat{L}(W) - L(W)| \leq \delta \right\}$ As in Theorem 6, we have $\Pr(E) \geq 1 - 2|C| \exp(-2k\delta^2)$ For any $W \in S$, pick $W_0 \in C$ such that $\|W - W_0\|_2 \leq \epsilon$ By assumption, L and $\hat{L}$ are Lipchitz continuous. Set $\epsilon = \delta/(2\kappa)$ then:

$$|\hat{L}(W) - L(W)| \leq |\hat{L}(W) - \hat{L}(W_0)| + |\hat{L}(W_0) - L(W_0)| + |L(W_0) - L(W)| \leq 2\kappa\epsilon + \delta = 2\delta$$

By Proposition 3, $\log|C| \leq p\log(3M/(\delta/2))$. Recall that we take M s.t. $\kappa M = 1$, so take $\delta = \sqrt{\frac{cp\log k}{k}}$

$$\begin{aligned}
\log|C| - 2k\delta^2 &\leq p\log\left(\frac{6M}{\delta}\right) - 2k\delta^2 \\
&= p\log\left(\frac{6M\sqrt{k}}{\sqrt{cp\log k}}\right) - 2k\frac{cp}{k}\log k \\
&< p\log\left(\frac{6M\sqrt{k}}{\sqrt{cpk}}\right) - 2cp\log k \\
&< (\left(\frac{1}{2} - \frac{5c}{2}\right)\log k + \log 6M - \frac{1}{2}logcp)\cdot p \\
&\leq -p,
\end{aligned}$$

The last inequality comes from the choosing of constant $c > \frac{1}{5}$. Then we have:
With probability of at least $1 - \mathcal{O}(e^{-p}), L(W_k) - L(W^*) \leq \mathcal{O}(\sqrt{\frac{logk}{k}})$ □

Assume that all reservoir states x of the reservoir system is bounded such that $0 < a < \|x\|_2 < b < \infty$ (this property has to do with the initialization of ESN), then from the above theorem(when we take p big enough (i.e reservoir system generated with high-dim matrices) such that $P \to 1$), we have the following corollary:

Corollary 1 (Convergence of the parameter)**.** *W_k is obtained from k-th iteration of regression in Algorithm (), W^* is the true estimate of the parameter of optimal Q-value functional, then :*

$$\mathbb{E}[\|W_k - W^*\|_2] \leq \mathcal{O}((\frac{\log k}{k})^{\frac{1}{4}})$$

or equivalently

$$\mathbb{E}[\|W_k - W^*\|_2] \leq \tilde{\mathcal{O}}(k^{-\frac{1}{4}})$$

7.4 Asymptomatic Analysis of ESN Critic

(ps:the asymptotic analysis should come before the non-asymptotic ones, so that we could see there are certain limitaions on the asymptotic analysis so the theoretical part should be rearranged. I put it here for simplicity (all the definitions are given in the previous part) for now)

The asymptotic approach gives a bound on $L(\hat{\theta}) - L(\theta^*)$ when the number of training sample $n \to \infty$ given the convergence of parameter θ.

Most of the asymptotic analysis are based on Central Limit Theorem which require the i.i.d samples. But there are also weaker results with the "almost independent" setting:

The MDP setting is such an ideal case:

Theorem 8 (Markov chain central limit theorem)**.** *Suppose that:*

- *The sequence $X_1, X_2, X_3, \ldots$ of Markovian random process that has a stationary probability distribution*
- *The initial distribution of the process, i.e. the distribution of X_1, is the stationary distribution, so that $X_1, X_2, X_3, \ldots$ are identically distributed.*

- *g is some (measurable) real-valued function for which var$(g(X_1)) < +\infty$.*

Set

$\mu = \mathrm{E}(g(X_1))$,

$\widehat{\mu}_n = \dfrac{1}{n}\sum_{k=1}^{n} g(X_k)$

$\sigma^2 := \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\widehat{\mu}_n) = \lim_{n\to\infty} n\,\mathrm{var}(\widehat{\mu}_n) = \mathrm{var}(g(X_1)) + 2\sum_{k=1}^{\infty} \mathrm{cov}(g(X_1), g(X_{1+k}))$.

Then when $n \to \infty$

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$$

Then we have the following results:

Theorem 9 (Asymptotic Convergence of parameter)**.** *Suppose that the process $\nabla l(z_t)$ satisfies the conditions in Theorem 8. Moreover, for the k-th iteration with total sample number $n \sim k$. Then if W^* is the real estimate, suppose:*

- *$\hat{W}_k \xrightarrow{p} W^*$ as $k \to \infty$*
- *For any k, L_k is twice differentiable w.r.t W (which is true in ESN case). $\nabla^2 L(w^*)$ is full rank.*
- *Similar to the term "expectation of l" in assumption one, here we assume $\nabla L = E(\nabla l)$*

Then

$$\sqrt{k}(W_k - W^*) = O_P(1)$$

i.e.for every $\epsilon > 0$.there is an M such that

$$\sup_k \mathbb{P}(\|\sqrt{k}(W_k - W^*)\|_2 > M) < \epsilon.$$

or when $k \to \infty$ $|W_k - W^*| \le \frac{c}{k} + o\left(\frac{1}{k}\right)$

Proof. By Taylor expansion of L_k at W^*:

$$0 = \nabla L_k(W_k) = \nabla L_k(W^*) + \nabla^2 L_k(W^*)(W_k - W^*) + O(\|W_k - W^*\|_2^2)$$

Rearrange it and multiple n, we get:

$$\sqrt{k}(W_k - W^*) = -(\nabla^2 L_k(W^*))^{-1}\sqrt{k}(\nabla L_k(W^*)) + O(\sqrt{k}||W_k - W^*||_2^2)$$

Then by Markov CLT (Theorem 8) with $X_i = \nabla\ell(z_i, w^*)$and $\widehat{X} = \nabla\widehat{L}(W^*)$ we have

$$\sqrt{k}(\nabla\widehat{L}(W^*) - \nabla L(W^*)) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\nabla\ell(\mathcal{Z}, \theta^*)))$$

Since $\nabla L(W^*) = 0$, $\sqrt{k}(\nabla\widehat{L}(W^*) \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\nabla\ell(\mathcal{Z}, \theta^*)))$

By the law of large number, $\nabla^2\widehat{L}(W^*) \xrightarrow{p} \nabla^2 L(W^*)$ Then

$\sqrt{k}(W_k - W^*) \xrightarrow{d} \nabla^2 L(\theta^*)^{-1}\mathcal{N}(0, \mathrm{Cov}(\nabla\ell(\mathcal{Z}, \theta^*)))$

$$\stackrel{d}{=} \mathcal{N}\left(0, \nabla^2 L(k^*)^{-1}\,\mathrm{Cov}(\nabla\ell(\mathcal{Z}, k^*))\nabla^2 L(k^*)^{-1}\right)$$

This indicates $\sqrt{k}(W_k - W^*) = O_P(1)$ □

7.5 Overall Convergence

To give a finite time analysis of the Actor-Critic algorithm with ESN, we have to make some assumptions on the actor model first. (Without further explanation, the following $\|\cdot\|$ are all L2 norm)

Assumption 2. *For the Actor model we have:*

- *The policy π_θ is differentiable with respect to θ, the log probability of choosing certain action a in certain state s can be written as $\nabla \log \pi_\theta(a \mid s)$*
- ***Bounded w.r.t** θ: Without loss of generality $\|\nabla \log \pi_\theta(a \mid s)\| \leq 1$*
- ***Lipchitz continuous w.r.t** θ: $\|\nabla \log \pi_{\theta_1}(a \mid s) - \nabla \log \pi_{\theta_2}(a \mid s)\| \leq \kappa_A \cdot \|\theta_1 - \theta_2\|$,for any θ_1, θ_2*

Besides, the following is discussed under the MDP/POMDP setting with reward bounded by κ_r and finite horizon.

Denote the trajectory length at k-th roll out as h(k), denote

$$F_t := Q^*(s_t, a_t)\nabla_\theta \log \pi_\theta(s_t, a_t),$$

$$F_{AC,t} := H_{W_k}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(s_t, a_t)\nabla_\theta \log \pi_\theta(s_t, a_t),$$

$$g_h = \sum_{t=1}^{h} \gamma^{t-1} F_t$$

then for the parameter of the actor, it's updated by

$$\theta_{k+1} = \theta_k + \eta_k g_h^{AC} := \theta_k + \frac{1}{1-\gamma}\eta_k \sum_{t=1}^{h(k)} \gamma^{t-1} H_{W_k}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(s_t, a_t)\nabla \log \pi_{\theta_k}(s_t, a_t|\theta_k)$$

With $\tau := \mathcal{F}_h = \sigma(z_1, z_2, \ldots, z_h)$

We further define:

Definition 4 (Ergodic distribution)**.** *For initial state s_0, the ergodic distribution associated with the fixed policy π_θ is defined as :*

$$\rho_{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(s_t = s \mid s_0, \pi_\theta)$$

, also for reference we define $\rho_\theta(s, a) = \rho_{\pi_\theta}(s) \cdot \pi_\theta(a \mid s)$

Recall Policy Gradient Theorem:

Theorem 10 (Policy Gradient)**.** *For value function $J(\theta)$*

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \mid s) \cdot Q_{\pi_\theta}(s, a)\right]$$

Lemma 3 (Approximation error of $\nabla J(\theta)$)**.** *Let Assumption 2 and Proposition 4 about the critic parameter in effect. For the k-th roll out, With $\tau := \mathcal{F}_{h(k)} = \sigma(z_1, z_2, \ldots, z_{h(k)})$,*

$$\left\|\mathbb{E}_\tau\left[g_{h(k)}^{AC}\right] - \nabla_\theta J(\theta)\right\| \leq C_1 \gamma^{h(k)} + C_2 (\frac{\log k}{k})^{\frac{1}{4}}$$

Proof. Since

$$\left\|\mathbb{E}_\tau[\hat{g}_h^{AC}] - \nabla_\theta J(\theta)\right\| \leq \left\|\mathbb{E}_\tau[g_\infty^{AC}] - \nabla_\theta J(\theta)\right\| + \left\|\mathbb{E}_\tau[g_h^{AC}] - \mathbb{E}_\tau[g_\infty^{AC}]\right\|$$

Consider

$$\begin{aligned}
\mathbb{E}_\tau[g_\infty^{AC}] &= \mathbb{E}_\tau\left[\sum_{t=1}^{\infty}\gamma^{t-1}F_{AC,t}\right] \\
&= \mathbb{E}_\tau\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(F_t + F_{AC,t} - F_t\right)\right] \\
&= \mathbb{E}_\tau\left[\sum_{t=1}^{\infty}\gamma^{t-1}F_t\right] + \mathbb{E}_\tau\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(F_{AC,t} - F_t\right)\right]
\end{aligned}$$

We shall first prove that the fist term equals $\nabla_\theta J(\theta)$. This comes from:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}F_t\right] &= \sum_{t=1}^{\infty}\gamma^{t-1}\int_S \mathbb{E}\big[F_t|s_t = s\big]\Pr\big(s_t = s|s_1\big)\mathrm{d}s \\
&= \sum_{t=1}^{\infty}\gamma^{t-1}\int_S\int_{\mathcal{A}} Q(s,a)\nabla_\theta \log\pi_\theta(s,a)\mathrm{d}a\Pr(s_t = s|s_1)\mathrm{d}s \\
&= \int_{\mathcal{S}}\int_{\mathcal{A}} Q(s,a)\nabla_\theta \log\pi_\theta(s,a)\mathrm{d}a\sum_{t=1}^{\infty}\gamma^{t-1}\Pr(s_t = s|s_1)\mathrm{d}s \\
&= \int_{\mathcal{S}}\int_{\mathcal{A}} Q(s,a)\nabla_\theta \log\pi_\theta(s,a)\mathrm{d}a\rho^{\pi_\theta}(s)\mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^{\pi_\theta}(s)}\left[\int_{\mathcal{A}} Q(s,a)\nabla_\theta \log\pi_\theta(s,a)\mathrm{d}a\right] \\
&= \mathbb{E}_{s\sim\rho^{\pi_\theta}(s),a\sim\pi_\theta(s,\cdot)}[Q(s,a)\nabla_\theta \log\pi_\theta(s,a)] \\
&= \nabla_\theta J(\theta)
\end{aligned}$$

(regularity assumptions are used here in the Fubini's theorem for exchanging sum and intergral)

With bounded reservoir states x in proposition 4 assume $\sigma(\boldsymbol{A}x + \boldsymbol{C}z + \zeta) \leq \kappa_s$ We then consider using the conclusion in proposition 4 to bound the second term:

$$\begin{aligned}
F_{AC,t} - F_t &= \left(H_{W_k}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(s_t,a_t) - H_{W^*}^{\boldsymbol{A},\boldsymbol{C},\boldsymbol{\zeta}}(s_t,a_t)\right)\nabla\log\pi_\theta(s_t,a_t) \\
&\leq \kappa_s(\frac{\log k}{k})^{\frac{1}{4}}
\end{aligned}$$

Then

$$\left\|g_\infty^{AC} - \nabla_\theta J(\theta)\right\| \leq \frac{\kappa_s}{1-\gamma}(\frac{\log k}{k})^{\frac{1}{4}} := C_2(\frac{\log k}{k})^{\frac{1}{4}}$$

Since

$$g_\infty - g_h = \gamma^{h-1}\sum_{t=0}^{\infty}\gamma^t F_{t+h+1} \leq \gamma^{h-1}\sum_{t=0}^{\infty}\gamma^t\kappa_r/(1-\gamma) \leq (\frac{\kappa_r}{(1-\gamma)^2})\gamma^{h-1}$$

Similarly with bounded out put of the critic

$$g_\infty^{AC} - g_h^{AC} \leq C_1\gamma^{H-1}$$

We now have

$$\left\|\mathbb{E}_\tau\left[g_h^{AC}\right] - \nabla_\theta J(\theta)\right\| \leq C_1\gamma^h + C_2(\frac{\log k}{k})^{\frac{1}{4}}$$

□

Proposition 5 (Lipchitz Continuity of Policy Gradient)**.**

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq \kappa_J \cdot \|\theta_1 - \theta_2\|$$

For some κ_J

Proof.

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\|$$

$$\begin{aligned}
\|\nabla J(\theta_1) - \nabla J(\theta_2)\| &= \sum_{t=1}^{\infty} \gamma^{t-1} \int_S \int_{\mathcal{A}} Q(s,a)(\nabla_{\theta_1} \log \pi_{\theta_1}(s,a) - \nabla_{\theta_2} \log \pi_{\theta_2}(s,a)) \mathrm{Pr}_{\theta_1}(s_t = s|s_1) \,\mathrm{d}a \,\mathrm{d}s \\
&\quad + \sum_{t=1}^{\infty} \gamma^{t-1} \int_S \int_{\mathcal{A}} Q(s,a)(\nabla_{\theta_2} \log \pi_{\theta_2}(s,a))(\mathrm{Pr}_{\theta_1}(s_t = s|s_1) - \mathrm{Pr}_{\theta_2}(s_t = s|s_1)) \,\mathrm{d}a \,\mathrm{d}s \\
&= \sum_{t=1}^{\infty} \gamma^{t-1} \int_S \int_{\mathcal{A}} |Q(s,a)| \|\nabla_{\theta_1} \log \pi_{\theta_1}(s,a) - \nabla_{\theta_2} \log \pi_{\theta_2}(s,a)\| |\mathrm{Pr}_{\theta_1}(s_t = s|s_1)| \,\mathrm{d}a \,\mathrm{d}s \\
&\quad + \sum_{t=1}^{\infty} \gamma^{t-1} \int_S \int_{\mathcal{A}} |Q(s,a)| \|\nabla_{\theta_2} \log \pi_{\theta_2}(s,a)\| |\mathrm{Pr}_{\theta_1}(s_t = s|s_1) - \mathrm{Pr}_{\theta_2}(s_t = s|s_1)| \,\mathrm{d}a \,\mathrm{d}s \\
&:= \sum_{t=1}^{\infty} \gamma^{t-1} (I_1 + I_2)
\end{aligned}$$

For the first term:

$$I_1 \leq \kappa_A \kappa_r \|\theta_1 - \theta_2\| \int_S \int_{\mathcal{A}} |\mathrm{Pr}_{\theta_1}(s_t = s|s_1)| \,\mathrm{d}a \,\mathrm{d}s = \kappa_A \kappa_r \|\theta_1 - \theta_2\|$$

For the second term:
Set $\mathcal{U}_t = \{u : u = 0, \cdots, t\}$, then

$$|\mathrm{Pr}_{\theta_1}(s_t = s|s_1) - \mathrm{Pr}_{\theta_2}(s_t = s|s_1)| = \left[\prod_{u=0}^{t-1} \mathrm{Pr}(s_{u+1}|s_u, a_u)\right] \cdot \left[\prod_{u \in \mathcal{U}_t} \pi_{\theta_1}(a_u|s_u) - \prod_{u \in \mathcal{U}_t} \pi_{\theta_2}(a_u|s_u)\right]$$

According to the Taylor expansion of $\prod_{u \in \mathcal{U}_t} \pi_\theta(a_u|s_u)$, $\exists \tilde{\theta}$

$$\left[\prod_{u \in \mathcal{U}_t} \pi_{\theta_1}(a_u|s_u) - \prod_{u \in \mathcal{U}_t} \pi_{\theta_2}(a_u|s_u)\right] \leq \|\theta_1 - \theta_2\| \left[\sum_{m \in \mathcal{U}_t} \nabla \pi_{\tilde{\theta}}(a_m|s_m) \prod_{u \in \mathcal{U}_t} \pi_{\tilde{\theta}}(a_u|s_u)\right] \leq \|\theta_1 - \theta_2\| \prod_{u \in \mathcal{U}_t} \pi_{\tilde{\theta}}(a_u|s_u)$$

$$I_2 \leq \|\theta_1 - \theta_2\| \cdot \kappa_r \cdot \int \left[\prod_{u=0}^{t-1} p(s_{u+1}|s_u, a_u)\right] \cdot (t+1) \cdot \prod_{u \in \mathcal{U}_t} \pi_{\tilde{\theta}}(a_u|s_u) \mathrm{d}a \,\mathrm{d}s = \|\theta_1 - \theta_2\|(t+1)\kappa_r$$

Combining I_1 and I_2, then $\exists \kappa_J$

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq \kappa_J \cdot \|\theta_1 - \theta_2\|$$

□

For the convergence, we still have to make an assumption about the length of the roll-out:

Assumption 3. *h(k) is generally a non-decreasing sequence, for any $0 < a < 1$,we have*

$$\int_0^k a^{h(s)} ds < \mathcal{O}(k^{1-\delta(a)})$$

for some $\delta(a)$

For example, for h(k)=k,

$$\int_0^k a^{h(s)} ds < \mathcal{O}(k^0) = Constant$$

We can now draw the conclusion of the main theorem

Theorem 11 (Convergence of Actor-Critic with ESN)**.** *For the k-th roll out, we have the critic parameter W_k and the actor is also updated k times with step-size $\eta_k = k^{-a}$Suppose we have $\delta(\gamma)$ as define in assumption 3. Then with all the above assumptions in effect, for any $\hat{\epsilon}$ by Algorithm () we have:*

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{E}[\|\nabla J(\theta_j)\|^2] < C_4 k^{a-1} + \kappa_J \sigma^2 k^{1-a} + \kappa_\nabla C_1 k^{-\delta(\gamma)} + \kappa_\nabla C_2 k^{-1/4+\hat{\epsilon}}$$

Proof. For the k-th roll out, we have the critic parameter W_k. Then $\exists \tilde{W}_k$ such that

$$J(W_{k+1}) = J(W_k) + (W_{k+1} - W_k)^T \nabla J(\tilde{W}_k)$$

$$(W_{k+1} - W_k)^\top \left(\nabla J(\tilde{W}_k) - J(W_k)\right) \geq -\kappa_J \|\tilde{W}_k - W_k\| \cdot \|W_{k+1} - W_k\| \geq -\kappa_J \|W_{k+1} - W_k\|^2$$

$$J(W_{k+1}) \geq J(W_k) + (W_{k+1} - W_k)^\top \nabla J(W_k) - \kappa_J \|W_{k+1} - W_k\|^2$$

Take the expectation with respect to $\mathcal{F}_k$, and since the reservoir states and the parameter W is bounded, $\exists \sigma$ such that $\mathbb{E}(\|g_h^{AC}\|) \leq \mathbb{E}(\|g_\infty^{AC}\|) \leq \sigma$. And by theorem 10 since both term is bounded in the assumption $\exists \kappa_\nabla$ such that $|\nabla_\theta J(\theta)\| \leq \kappa_\nabla$

Denote the σ-algebra generated by the set $\{s_u, a_u, w_u\}_{u \leq k}$ as $\mathcal{F}_k$

$$\begin{aligned}
\mathbb{E}[J(W_{k+1})|\mathcal{F}_k] &\geq J(W_k) + \mathbb{E}[W_{k+1} - W_k|\mathcal{F}_k]^\top \nabla J(W_k) - \kappa_J \mathbb{E}[\|\eta_k g_{h(k)}^{AC}\|^2|\mathcal{F}_k] \\
&\geq J(W_k) + \mathbb{E}[W_{k+1} - W_k|\mathcal{F}_k]^\top \nabla J(W_k) - \kappa_J \sigma^2 \eta_k^2 \\
&\geq J(W_k) + \eta_k \|\nabla_W J(W_k)\|^2 + \eta_k \mathbb{E}\left[g_{h(k)}^{AC} - \nabla_W J(W_k)\big|\mathcal{F}_k\right]^\top \nabla_W J(W_k) - \kappa_J \sigma^2 \eta_k^2 \\
&\geq J(W_k) + \eta_k \|\nabla_W J(W_k)\|^2 - \eta_k \left|\mathbb{E}\left[g_{h(k)}^{AC} - \nabla_W J(W_k)|\mathcal{F}_k\right]^\top \nabla_W J(W_k)\right| - \kappa_J \sigma^2 \eta_k^2 \\
&\geq J(W_k) + \eta_k \|\nabla_W J(W_k)\|^2 - \eta_k \|\mathbb{E}\left[g_{H(k)}^{AC}\right] - \nabla_W J(W_k)\| \cdot \|\nabla_W J(W_k)\| - \kappa_J \sigma^2 \eta_k^2 \\
&\geq J(W_k) + \eta_k \|\nabla_W J(W_k)\|^2 - \eta_k \kappa_\nabla \left(C_1 \gamma^{h(k)} + C_2 (\frac{\log k}{k})^{\frac{1}{4}}\right) - \kappa_J \sigma^2 \eta_k^2
\end{aligned}$$

Define $U_k := J(W^*) - J(W_k)$, then rearrange the above inequality and take the total expectation, we have:

$$\eta_k \mathbb{E}[\|\nabla J(W_k)\|^2] \leq \mathbb{E}[U_k] - \mathbb{E}[U_{k+1}] + \kappa_J \sigma^2 \eta_k^2 + \eta_k \kappa_\nabla C_1 \gamma^{h(k)-1} + \eta_k \kappa_\nabla C_2 (\frac{\log k}{k})^{\frac{1}{4}}$$

Then taking the sum over $\{k-N, \ldots, k\}$. Notice that $0 < \mathbb{E}(U_k) < 2J(\theta^*) < \frac{2\kappa_r}{1-\gamma} := C_3$

$$\begin{aligned}
\sum_{j=k-N}^{k} \mathbb{E}[\|\nabla J(\theta_j)\|^2] &\leq \sum_{j=k-N}^{k} \frac{1}{\eta_j} (\mathbb{E}[U_j] - \mathbb{E}[U_{j+1}]) + \kappa_J \sigma^2 \sum_{j=k-N}^{k} \eta_j + \sum_{j=k-N}^{k} \left(\kappa_\nabla C_1 \gamma^{h(j)-1} + \kappa_\nabla C_2 (\frac{\log k}{k})^{\frac{1}{4}} \right) \\
&\leq \sum_{j=k-N}^{k} \left(\frac{1}{\eta_j} - \frac{1}{\eta_{j-1}} \right) \mathbb{E}[U_j] - \frac{1}{\eta_k} \mathbb{E}[U_{k+1}] + \frac{1}{\eta_{k-N-1}} \mathbb{E}[U_{k-N}] + \kappa_J \sigma^2 \sum_{j=k-N}^{k} \eta_j \\
&+ \sum_{j=k-N}^{k} \left(\kappa_\nabla C_1 \gamma^{h(j)-1} + \kappa_\nabla C_2 (\frac{\log k}{k})^{\frac{1}{4}} \right) \\
&< \sum_{j=k-N}^{k} \left(\frac{1}{\eta_j} - \frac{1}{\eta_{j-1}} \right) C_3 + \frac{1}{\eta_{k-N-1}} C_3 + \kappa_J \sigma^2 \sum_{j=k-N}^{k} \eta_j \\
&+ \sum_{j=k-N}^{k} \left(\kappa_\nabla C_1 \gamma^{h(j)-1} + \kappa_\nabla C_2 (\frac{\log k}{k})^{\frac{1}{4}} \right) \\
&= \frac{C_3}{\eta_k} + \kappa_J \sigma^2 \sum_{j=k-N}^{k} \eta_j + \sum_{j=k-N}^{k} \left(\kappa_\nabla C_1 \gamma^{h(j)-1} + \kappa_\nabla C_2 (\frac{\log k}{k})^{\frac{1}{4}} \right) \\
&= C_4 k^a + \kappa_J \sigma^2 \sum_{j=k-N}^{k} j^{-a} + \sum_{j=k-N}^{k} \left(\kappa_\nabla C_1 \gamma^{h(j)-1} + \kappa_\nabla C_2 (\frac{\log j}{j})^{\frac{1}{4}} \right)
\end{aligned}$$

Analyzing each term, we have

$$\kappa_J \sigma^2 \sum_{j=k-N}^{k} j^{-a} \leq \kappa_J \sigma^2 \int_{k-N}^{k} j^{-a} dj = \kappa_J \sigma^2 (k^{1-a} - (k-N-1)^{1-a}) < \kappa_J \sigma^2 k^{1-a}$$

$$\sum_{j=k-N}^{k} \kappa_\nabla C_1 \gamma^{h(j)-1} \leq \kappa_\nabla C_1 k^{1-\delta(\gamma)}$$

And for any $\hat{\epsilon}$

$$\sum_{j=k-N}^{k} \kappa_\nabla C_2 (\frac{\log j}{j})^{\frac{1}{4}} \leq \int_{k-N}^{k} \kappa_\nabla C_2 k^{(-1/4+\hat{\epsilon})} = \kappa_\nabla C_2 (k^{1-1/4+\hat{\epsilon}} - (k-N-1)^{3/4+\hat{\epsilon}}) < \kappa_\nabla C_2 k^{3/4+\hat{\epsilon}}$$

Now divide both sides by k and set $N = k-1$, we have for any $\hat{\epsilon}$

$$\frac{1}{k} \sum_{j=1}^{k} \mathbb{E}[\|\nabla J(\theta_j)\|^2] < C_4 k^{a-1} + \kappa_J \sigma^2 k^{-a} + \kappa_\nabla C_1 k^{-\delta(\gamma)} + \kappa_\nabla C_2 k^{-1/4+\hat{\epsilon}}$$

□

7.6 Discussion on sample complexity

Definition 5 (Sample Complexity)**.** *Since the actor and ESN critic update at the same time scale, Define*

$$K_\epsilon = \min\{k : \inf_{0\le m\le k} \|\nabla J(\theta_m)\|^2 < \epsilon\}$$

as the sample complexity.

Proposition 6 (Sample Complexity of Actor- Critic with ESN)**.** *Recall that we set the step-size of the actor as $\eta_k = k^{-a}$ where $a \in (0,1)$. Then we have*

$$K_\epsilon \le \begin{cases} \mathcal{O}(\epsilon^{-1/l}) & \text{where} \quad \mathrm{l} := \min\{\mathrm{a}, 1-\mathrm{a}, \delta(\gamma)\} \text{ if } l < 1/4 \\ \mathcal{O}(\epsilon^{-\frac{4}{1+4\hat{\epsilon}}}). & \text{Otherwise} \quad \text{for} \quad \forall \hat{\epsilon} > 0 \end{cases}$$

Here $\mathcal{W}$ is the Lambert W function.

Proof. With finite horizon, we have $K_\epsilon \sim k$. Then

$$\epsilon \le \frac{1}{K_\epsilon}\sum_{j=1}^{K_\epsilon} \mathbb{E}[\|\nabla J(\theta_j)\|^2] \le \mathcal{O}\big(K_\epsilon^{a-1} + K_\epsilon^{-a} + K_\epsilon^{-\delta\gamma} + K_\epsilon^{-1/4+\hat{\epsilon}}\big)$$

Then if $l := \min\{a, 1-a, \delta(\gamma)\} < 1/4$

$$\epsilon \le \frac{1}{K_\epsilon}\sum_{j=1}^{K_\epsilon} \mathbb{E}[\|\nabla J(\theta_j)\|^2] \le \mathcal{O}\big(K_\epsilon^{-l}\big)$$

$$K_\epsilon \le \mathcal{O}(\epsilon^{-1/l})$$

Otherwise

$$\epsilon \le \frac{1}{K_\epsilon}\sum_{j=1}^{K_\epsilon} \mathbb{E}[\|\nabla J(\theta_j)\|^2] \le \mathcal{O}\big(K_\epsilon^{-1/4+\hat{\epsilon}}\big)$$

$$K_\epsilon \le \mathcal{O}(\epsilon^{-\frac{4}{1+4\hat{\epsilon}}})$$

□

References