# Reinforcement Learning - Final Project

Adi Akavia

June 9, 2000

# 1 Introduction

In this project I've studied the article: *"Finite-Sample Convergence Rates for Q-learning and Indirect Algorithms", by Michael Kearns and Satinder Singh.* The article discusses the amount of experience needed for achieving a policy with a certain level of performance guarantee by the learning algorithms: Phased-Q-Learning (which is a variant of the familiar Q-Learning) and the indirect algorithm (both algorithms are explained later on); and compares the behaviour of the direct (Phased-Q-Learning) and indirect algorithms. The article presents a theorem but does not prove it, so the heart of my project is presenting a proof to the theorem.

# 2 Notations

- M is an MDP
- S is the set of states in M
- A is the set of actions in M
- ${R^a}_M(s)$ is the return of action $a$ from state $s$. ${R^a}_M(s)$ is assumed to be constant in the article (non-stochastic returns), but this causes no loss of generality, since instead of assuming the reward is constant, we can take its expectation.
- $P^a_{sj}$ is the probability of reaching state $j$ by performing action $a$ from state $s$.

# 3 The Parallel Sampling Model

The *Parallel Sampling Model* is a model of an *ideal* exploration policy, in the sense that every state-action pair is sampled with the same frequency. We define a subroutine called $PS(M)$ that simulates the Parallel Sampling Model as follows: a *single* call to $PS(M)$ returns, for *every* state-action pair $(s, a)$, a random next state $s'$ distributed according to the transition distribution $P^a_{sj}$. Thus a single call to $PS(M)$ is simulating $|S| \times |A|$ transitions in the given $MDP$, and returns $|S| \times |A|$ next states.

Of course often this model is only an ideal and not an applicable procedure; the advantage of using such a model of perfect sampling method is that it enables us to separate the analysis of the *learning algorithms* from the quality of the policy we can sample. Using the analysis for the ideal exploration policy we can also bound the learning rate of any given exploration policy that visits every state-action pair infinitely often, by approximating the procedure $PS(M)$ with taking enough samples of the exploration function, such that with good probability every state-action pair is visited.

# 4 The Learning Algorithms

The article refers to two learning algorithms:

- A direct algorithm called *Phased-Q-Learning*, and
- An indirect algorithm

## 4.1 Direct Algorithm - Phased-Q-Learning

The Phased-Q-Learning algorithm is similar to the Q-Learning algorithm we've encountered in class, only it works in phases. In each phase the algorithm makes $m_D$ calls to $PS(M)$ (where $m_D$ is determined later by the analysis). The algorithm uses the $m_D$ samples of every state-action pair collected by the $m_D$ calls to $PS(M)$ to update the value function as follows:

$$\begin{aligned}
\forall s,a : \widehat{Q}_0(s,a) &= \widehat{v}_0(s) = 0 \\
\widehat{Q}_{l+1}(s,a) &= {R^a}_M(s) + \gamma \frac{1}{m_D} \sum_{k=1}^{m_D} \widehat{v}_l({j^l}_k) \\
&\quad \textit{where } {j^l}_1, \ldots, {j^l}_{m_D} \textit{ are the } m_D \textit{ next states observed from } (s,a) \\
&\quad \textit{on the } m_D \textit{ calls to } PS(M) \textit{ during the } l^{th} \textit{phase, and} \\
&\quad \widehat{v}_l(s) = \max_{a \in A} \left\{ \widehat{Q}(s,a) \right\}
\end{aligned}$$

Note that the Phased-Q-Learning algorithm requires $l_D \times m_D$ calls to $PS(M)$, where $l_D$ is the number of performed phases.

## 4.2 Indirect Algorithm

The indirect algorithm works as follows:

- First it makes $m_I$ calls to $PS(M)$ to obtain $m_I$ next states for each state-action pair. Here, again, $m_I$ is determined later by the analysis.
- The next step of the indirect algorithm is building an empirical model of the transition probabilities using the collected samples as follows: $\widehat{P}^a_{st} = \frac{\#(s \to_a t)}{m_I}$, note that $\widehat{P}^a_{st}$, the transition probabilities in the empirical model, is an *estimate* to the transition probability from state $s$ to state $t$ by performing action $a$, in the given MDP $M$.
- The third stage is iterating the $Value-Iteration$ algorithm (i.e. $Q_{l+1}(s,a) = R^a{}_M(s) + \gamma \sum_{t \in S} \widehat{P}^a_{st} \widehat{v}_l(t)$ ) on the model we've established in the second stage of the algorithm for $l_I$ iterations, and returns the achieved policy.

Note that the indirect algorithm requires $m_I$ calls to $PS(M)$.

# 5 The Main Theorem - Bound on the Number of Samples

The main theorem of the article bounds the number of calls to the subroutine $PS(M)$, required by the learning algorithms to ensure with probability of at least $1-\delta$ that the achieved policy is an $\varepsilon$-optimal policy. The bounds stated by the article are:

**Theorem 5.1** *Main Theorem*

- *For an appropriate choice of the parameters $m_D$ and $l_D$, the total number of calls to $PS(M)$ required by the Phased-Q-Learning algorithm in order to ensure that, with probability at least $1-\delta$, the expected return of the resulting policy will be within $\varepsilon$ of the optimal policy, is:*

$$O\left(\left(\frac{1}{\varepsilon^2} \cdot \ln \frac{1}{\varepsilon}\right) \cdot \left(\ln \frac{|S|}{\delta} + \ln \ln \frac{1}{\varepsilon}\right)\right) \qquad (1)$$

- *For an appropriate choice of the parameters $m_I$ and $l_I$, the total number of calls to $PS(M)$ required by the indirect algorithm in order to ensure that, with probability at least $1-\delta$, the expected return of the resulting policy will be within $\varepsilon$ of the optimal policy, is*

$$O((\frac{1}{\varepsilon^2}) \cdot (\ln \frac{|S|}{\delta} \; + \; \ln \ln \frac{1}{\varepsilon})) \tag{2}$$

# 6 My Results

The article presents no proof to its theorem, so the core of my project is providing a proof to the theorem. The results I've accomplished depend on $|A|$ (the size of the actions-set), but under the assumption of a constant $|A|$, the result are the same:

- For an appropriate choice of the parameters $m_D$ and $l_D$, the total number of calls to $PS(M)$ required by the Phased-Q-Learning algorithm in order to ensure that, with probability at least $1-\delta$, the expected return of the resulting policy will be within $\varepsilon$ of the optimal policy, is:

$$O((\frac{1}{\varepsilon^2} \cdot \ln \frac{1}{\varepsilon}) \cdot (\ln (\frac{|S| \cdot |A|}{\delta}) + \ln \ln \frac{1}{\varepsilon})) \tag{3}$$

- For an appropriate choice of the parameters $m_I$ and $l_I$, the total number of calls to $PS(M)$ required by the indirect algorithm in order to ensure that, with probability at least $1-\delta$, the expected return of the resulting policy will be within $\varepsilon$ of the optimal policy, is

$$O(\frac{1}{\varepsilon^2} \cdot (\ln (\frac{|S| \cdot |A|}{\delta}) + \ln \ln \frac{1}{\varepsilon})) \tag{4}$$

# 7 Proof for Phased Q-Learning

## 7.1 Notations

- $Q_l$: A $(state, action)$ value function defined by $Q_{l+1}(s,a) = {R_M}^a(s) + \gamma \sum_{t \in S} {P^a}_{st} v_l(t)$, where $v_l(t) = \max_{b \in A} \{Q_l(t,b)\}$. Note, this is the operation of the Value-Iteration algorithm

- $\widehat{Q}_l$: A $(state, action)$ value function defined by $\widehat{Q}_{l+1}(s,a) = R_M{}^a(s) + \gamma \frac{1}{m_D} \sum_{k=1}^{m_D} \widehat{v}_l(t_k)$, where $\widehat{v}_l(t_k) = \max_{b \in A} \left\{ \widehat{Q_l}(t_k, b) \right\}$, and $t_k$ are the $m_D$ next states observed from $(s,a)$ on the $m_D$ calls to $PS(M)$. Note, this is the operation of the phased-Q-learning algorithm

- $Q^*$ denotes, as usually, the optimal value function.

## 7.2 Bounding $\left\| \widehat{Q}_{l_D} - Q^* \right\|$

After $l_D$ iterations, the algorithm Phased-Q-Learning returns a policy within $\varepsilon$ of the optimal policy, if

$$\left\| \widehat{Q}_{l_D} - Q^* \right\| \leq \varepsilon \tag{5}$$

(due to the definition of $Q^*$).
By the triangle inequality:

$$\left\| \widehat{Q}_l - Q^* \right\| \leq \left\| \widehat{Q}_l - Q_l \right\| + \left\| Q_l - Q^* \right\| \tag{6}$$

Let's bound each of this values separately:

**Claim 7.1**

$$\forall (s,a) \in S \times A,\ 0 \leq i \leq l_D \Pr(\left\| \frac{1}{m_D} \sum_{k=1}^{m_D} \widehat{v}_i(x_k) - \sum_{j \in S} P_{sj}^a \widehat{v}_i(j) \right\| \geq w) \leq 2e^{-2m_D w^2} \tag{7}$$

**Proof:** Note that

- $\sum_{j \in S} P_{sj}^a \widehat{v}_i(j)$ is the expectation for the $\widehat{v}$ value of the next samples received from $PS(M)$ for the pair $(s,a)$ (since the samples are taken from the distribution $P_{sj}^a$); i.e. $E[\frac{1}{m_D} \sum_{k=1}^{m_D} \widehat{v}_i(x_k)] = \sum_{j \in S} P_{sj}^a \widehat{v}_i(j)$

- $\widehat{v}_k$ is bounded (since the immediate return is bounded and $\gamma < 1$)

- The samples are independent and identically distributed.

Therefore, from *Chernoff Inequality* we receive that $\forall w > 0$

$$\forall s \in S,\ a \in A,\ 0 \leq i \leq l_D \Pr(\left\| \frac{1}{m_D} \sum_{k=1}^{m_D} \widehat{v}_i(x_k) - \sum_{j \in S} P_{sj}^a \widehat{v}_i(j) \right\| \geq w) \leq 2e^{-2m_D w^2} \tag{8}$$

□

**Claim 7.2** *Bounding* $\left\|\widehat{Q}_l - Q_l\right\|$*:*

- $\left\|\widehat{Q}_l - Q_l\right\| \leq \max_{(s,a)\in S\times A,\ 0\leq i\leq l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k) - \sum_{j\in S}P_{sj}^a\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=1}^{l}\gamma^i\right\}$
- $\Pr(\left\|\widehat{Q}_l - Q_l\right\| \geq w\cdot\sum_{i=1}^{l}\gamma^i) \leq l\cdot|S|\cdot|A|\cdot 2e^{-2m_D w^2}$

**Proof:** First by induction let's see that:

$$\left\|\widehat{Q}_l - Q_l\right\| \leq \max_{(s,a)\in S\times A,\ 0\leq i\leq l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k) - \sum_{j\in S}P_{sj}^a\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=1}^{l}\gamma^i\right\} \tag{9}$$

Induction's Base: For $l = 0$ both sides of the equation are zero, and the inequality holds.
Induction's Step: Assume the claim for $l-1$ and prove it for $l$:

$$\begin{aligned}
\left\|\widehat{Q}_l - Q_l\right\| &= \max_{(s,a)\in S\times A}\left\|\widehat{Q}_l(s,a) - Q_l(s,a)\right\| \\
&= \max_{(s,a)\in S\times A}\left\|{R_M}^a(s) + \gamma\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k) - {R_M}^a(s) - \gamma\sum_{j\in S}P_{sj}^a v_{l-1}(j)\right\| \\
&\quad \textit{where } x_1,\ \ldots,\ x_{m_D} \textit{ are the } m_D \textit{ next states observed from } (s,a) \\
&\quad \textit{on the } m_D \textit{ calls to } PS(M). \\
&= \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k) - \sum_{j\in S}P_{sj}^a v_{l-1}(j)\right\|\right\} \\
&\quad \textit{By triangle inequality}: \\
&\leq \max_{(s,a)\in S\times A}\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k) - \sum_{j\in S}P_{sj}^a\widehat{v}_{l-1}(j)\right\| + \\
&\quad \gamma\left\|\sum_{j\in S}P_{sj}^a\widehat{v}_{l-1}(j) - \sum_{j\in S}P_{sj}^a v_{l-1}(j)\right\|\} \\
&\leq \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k) - \sum_{j\in S}P_{sj}^a\widehat{v}_{l-1}(j)\right\|\right\} + \\
&\quad \max_{(s,a)\in S\times A}\left\{\gamma\left\|\sum_{j\in S}P_{sj}^a\widehat{v}_{l-1}(j) - \sum_{j\in S}P_{sj}^a v_{l-1}(j)\right\|\right\} \\
&\quad \textit{By triangle inequality}:
\end{aligned}$$

$$
\begin{aligned}
\leq\ & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-1}(j)\right\|\right\}+\\
& \max_{(s,a)\in S\times A}\left\{\gamma\sum_{j\in S}P^a_{sj}\left\|\widehat{v}_{l-1}(j)-v_{l-1}(j)\right\|\right\}\\
\leq\ & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-1}(j)\right\|\right\}+\\
& \max_{(s,a)\in S\times A}\left\{\gamma\left\|\widehat{v}_{l-1}-v_{l-1}\right\|\cdot\sum_{j\in S}P^a_{sj}\right\}\\
=\ & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-1}(j)\right\|\right\}+\\
& \max_{(s,a)\in S\times A}\left\{\gamma\left\|\widehat{v}_{l-1}-v_{l-1}\right\|\cdot 1\right\}\\
& Note\ that:\\
& \left\|\widehat{v}_l-v_l\right\|=\max_{s\in S}\left\|\widehat{v}_l(s)-v_l(s)\right\|=\max_{s\in S}\left\|\max_{a\in A}\widehat{Q}_l(s,a)-\max_{a\in A}Q_l(s,a)\right\|\\
& \leq\max_{s\in S}\max_{a\in A}\left\|\widehat{Q}_l(s,a)-Q_l(s,a)\right\|\leq\max_{(s,a)\in S\times A}\left\|\widehat{Q}_l(s,a)-Q_l(s,a)\right\|=\left\|\widehat{Q}_l-Q_l\right\|\\
& Therefore:\\
\leq\ & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-1}(j)\right\|\right\}+\\
& \max_{(s,a)\in S\times A}\left\{\gamma\left\|\widehat{Q}_{l-1}-Q_{l-1}\right\|\right\}\\
& By\ induction's\ assumption:\\
\leq\ & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-1}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-1}(j)\right\|\right\}+\\
& \max_{(s,a)\in S\times A}\left\{\gamma(\max_{(t,b)\in S\times A,\ 0\leq i\leq l-1}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(y_k)-\sum_{j\in S}P^b{}_{tj}\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=1}^{l-1}\gamma^i\right\})\right\}\\
& where\ y_1,\ \ldots,\ y_{m_D}\ are\ the\ m_D\ next\ states\ observed\ from\ (t,b)\\
& on\ the\ m_D\ calls\ to\ PS(M).\\
\leq\ & \max_{(s,a)\in S\times A,\ o\leq i\leq l}\left\{\gamma\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\right\}+
\end{aligned}
$$

$$\max_{(s,a)\in S\times A,\ 0\le i\le l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=2}^{l}\gamma^i\right\}$$
$$=\max_{(s,a)\in S\times A,\ 0\le i\le l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=1}^{l}\gamma^i\right\}$$

Now, after we've accomplished the bound for $\left\|\widehat{Q}_l-Q_l\right\|$, let's show that:

$$\Pr(\left\|\widehat{Q}_l-Q_l\right\|\ge w\cdot\sum_{i=1}^{l}\gamma^i)\le l\cdot|S|\cdot|A|\cdot 2e^{-2m_Dw^2} \tag{10}$$

$$\begin{aligned}
&\Pr(\left\|\widehat{Q}_l-Q_l\right\|\ge w\cdot\sum_{i=1}^{l}\gamma^i)\\
&\quad \textit{By the first bound}:\\
&\le \Pr(\max_{(s,a)\in S\times A,\ 0\le i\le l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\cdot\sum_{i=1}^{l}\gamma^i\right\}\ge w\cdot\sum_{i=1}^{l}\gamma^i)\\
&= \Pr(\max_{(s,a)\in S\times A,\ 0\le i\le l}\left\{\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\right\}\ge w)\\
&= \Pr(\exists(s,a)\in S\times A,\ 0\le i\le l\ s.t.\ \left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\ge w)\\
&\le \sum_{(s,a)\in S\times A,\ 0\le i\le l}\Pr(\left\|\frac{1}{m_D}\sum_{k=1}^{m_D}\widehat{v}_{l-i}(x_k)-\sum_{j\in S}P^a_{sj}\widehat{v}_{l-i}(j)\right\|\ge w)\\
&\quad \textit{By claim } 7.1:\\
&\le \sum_{(s,a)\in S\times A,\ 0\le i\le l} 2e^{-2m_Dw^2}\\
&= l\cdot|S|\cdot|A|\cdot 2e^{-2m_Dw^2}
\end{aligned}$$

□

**Claim 7.3** *Bounding* $\|Q_l-Q^*\|$*:*

$$\|Q_l-Q^*\|\le\frac{\gamma^{l-1}}{1-\gamma}\|v_1-v_0\| \tag{11}$$

**Proof:** We saw in class that:

$$\forall l \, \|v_l - v^*\| \leq \frac{\gamma^l}{1-\gamma} \|v_1 - v_0\| \tag{12}$$

so,

$$\begin{aligned}
\|Q_l - Q^*\| &= \max_{(s,a)\in S\times A} \|Q_l(s,a) - Q^*(s,a)\| \\
&= \max_{(s,a)\in S\times A} \left\| R_M^a(s) + \sum_{j\in S} P_{sj}^a v_{l-1}(s) - R_M^a(s) - \sum_{j\in S} P_{sj}^a v^*(s) \right\| \\
&\leq \max_{(s,a)\in S\times A} \sum_{j\in S} P_{sj}^a \|v_{l-1}(s) - v^*(s)\| \\
&\leq \max_{a\in A} \sum_{j\in S} P_{sj}^a \|v_{l-1} - v^*\| \\
&\leq \max_{a\in A} \sum_{j\in S} P_{sj}^a \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\| \\
&= \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\| \max_{a\in A} \sum_{j\in S} P_{sj}^a \\
&= \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\|
\end{aligned}$$

and the bound is received. □

## 7.3 Putting It All Together

We've received:

$$\left\| \hat{Q}_l - Q^* \right\| \leq \left\| \hat{Q}_l - Q_l \right\| + \|Q_l - Q^*\| \tag{13}$$

where

- For any $w > 0$: $\Pr\left(\left\| \hat{Q}_l - Q_l \right\| \geq w \cdot \sum_{i=1}^{l} \gamma^i\right) \leq l \cdot |S| \cdot |A| \cdot 2e^{-2m_D w^2}$
- $\|Q_l - Q^*\| \leq \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\|$

*Hence we can sum up all of the above by:*

$$\begin{aligned}
&\Pr(\left\|\widehat{Q}_l - Q^*\right\| \geq w \cdot \sum_{i=1}^{l} \gamma^i + \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\|) \\
\leq\ & \Pr(\left\|\widehat{Q}_l - Q_l\right\| + \|Q_l - Q^*\| \geq w \cdot \sum_{i=1}^{l} \gamma^i + \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\|) \\
\leq\ & \Pr(\left\|\widehat{Q}_l - Q_l\right\| + \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\| \geq w \cdot \sum_{i=1}^{l} \gamma^i + \frac{\gamma^{l-1}}{1-\gamma} \|v_1 - v_0\|) \\
\leq\ & \Pr(\left\|\widehat{Q}_l - Q_l\right\| \geq w \cdot \sum_{i=1}^{l} \gamma^i) \\
\leq\ & l \cdot |S| \cdot |A| \cdot 2e^{-2m_D w^2}
\end{aligned}$$

## 7.4 Finding the Constants $m_D$ and $l_D$

For parameters $w,\ m_D\ and\ l_D$ s.t.:

1. $\varepsilon = w \cdot \sum_{i=1}^{l_D} \gamma^i + \frac{\gamma^{l_D-1}}{1-\gamma} \|v_1 - v_0\|$
2. $\delta = l_D \cdot |S| \cdot |A| \cdot 2e^{-2m_D w^2}$

we get, by the above result, that with probability at least $1-\delta$ we have reached an $\varepsilon - optimal$ policy i.e.: $\Pr(\left\|\widehat{Q}_{l_D} - Q^*\right\| < \varepsilon) > 1-\delta$
So, let's find such parameters $w,\ m_D\ and\ l_D$. Note that $m_D$ and $l_D$ must be natural numbers, but it suffices to constrain them to a *real* value, since taking an upper value won't change the asymptotic complexity.

¿From (1) we get:

$$\begin{aligned}
w &= \frac{\varepsilon - \frac{\gamma^{l_D-1}}{1-\gamma} \|v_1 - v_0\|}{\sum_{i=1}^{l_D} \gamma^i} \\
&= \frac{\varepsilon - \frac{\gamma^{l_D-1}}{1-\gamma} \|v_1 - v_0\|}{\frac{\gamma - \gamma^{l_D+1}}{1-\gamma}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(1-\gamma)\cdot(\varepsilon - \frac{\gamma^{l_D-1}}{1-\gamma}\|v_1 - v_0\|)}{\gamma - \gamma^{l_D+1}} \\
&= \frac{(1-\gamma)\varepsilon - (\gamma^{l_D-1}\|v_1 - v_0\|)}{\gamma - \gamma^{l_D+1}}
\end{aligned}
$$

In order to use Chernoff, we must have $w > 0$,

$$
\begin{aligned}
w > 0 \quad &\Leftrightarrow \quad \frac{(1-\gamma)\varepsilon - \gamma^{l_D-1}\|v_1 - v_0\|)}{\gamma - \gamma^{l_D+1}} > 0 \\
&\Leftrightarrow \quad (1-\gamma)\varepsilon - \gamma^{l_D-1}\|v_1 - v_0\|) > 0 \quad (\gamma - \gamma^{l_D} > 0 \ since \ 0 < \gamma < 1) \\
&\Leftrightarrow \quad (1-\gamma)\varepsilon > \gamma^{l_D-1}\|v_1 - v_0\| \\
&\Leftrightarrow \quad \frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|} > \gamma^{l_D-1} \\
&\Leftrightarrow \quad \ln\left(\frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|}\right) > (l_D - 1)\cdot \ln\gamma \\
&\Leftrightarrow \quad l_D > \frac{\ln\left(\frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|}\right)}{\ln\gamma} + 1 \quad (since \ \ln\gamma < 0) \\
&\Leftrightarrow \quad l_D > log_\gamma\left(\frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|}\right) + 1
\end{aligned}
$$

And we get the constraint:

$$
l_D > log_\gamma\left(\frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|}\right) + 1 \tag{14}
$$

This constraint is fulfilled by setting $l_D = log_\gamma(\frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|}) + f$, for *any* $f > 1$, and in particular, for $f = log_\gamma 2$ ($log_\gamma 2 > 1$ since $0 < \gamma < 1$), we get:

$$
l_D = \log_\gamma \frac{(1-\gamma)\varepsilon}{\|v_1 - v_0\|} + \log_\gamma 2 = \log_\gamma \frac{2(1-\gamma)\varepsilon}{\|v_1 - v_0\|} \tag{15}
$$

¿From (2) we get:

$$
\ln\delta = \ln(l_D \cdot |S| \cdot |A| \cdot 2) - 2m_D w^2 \ln e = \ln(2 l_D |S| \cdot |A|) - 2m_D w^2 \tag{16}
$$

Hence:

$$m_D = \frac{\ln(2l_D|S|\cdot|A|) - \ln\delta}{2w^2} = \frac{\ln(\frac{2l_D|S|\cdot|A|}{\delta})}{2w^2} \qquad (17)$$

By substituting with the $w$ received from (1), we get:

$$\begin{aligned}
m_D &= \frac{\ln(\frac{2l_D|S|\cdot|A|}{\delta})}{2w^2} \\
&= \frac{\ln(\frac{2l_D|S|\cdot|A|}{\delta})}{2(\frac{(1-\gamma)\varepsilon-\gamma^{l_D-1}\|v_1-v_0\|)}{\gamma-\gamma^{l_D+1}})^2} \\
&= \frac{(\gamma-\gamma^{l_D+1})^2\cdot\ln(\frac{2l_D|S|\cdot|A|}{\delta})}{2((1-\gamma)\varepsilon-\gamma^{l_D-1}\|v_1-v_0\|)^2} \\
& \mathit{Substitution}\ l_D\ \mathit{with\ the\ value\ we've\ received\ in}\ 15: \\
&= \frac{(\gamma-\gamma^{\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})+1})^2\cdot\ln(\frac{2\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})|S|\cdot|A|}{\delta})}{2((1-\gamma)\varepsilon-\gamma^{\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})-1}\|v_1-v_0\|)^2} \\
& \mathit{Note}\ \gamma^{l_D} = \gamma^{\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})} = \frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|} \\
&= \frac{(\gamma-\gamma\cdot\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})^2\cdot\ln(\frac{2\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})|S|\cdot|A|}{\delta})}{2((1-\gamma)\varepsilon-\gamma^{-1}\cdot\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|}\|v_1-v_0\|)^2} \\
&= \frac{(\gamma-\gamma\cdot\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})^2\cdot\ln(\frac{2|S|\cdot|A|\cdot\log_\gamma\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|}}{\delta})}{2((1-\gamma)\varepsilon-2\gamma^{-1}(1-\gamma)\varepsilon)} \\
&= \gamma\cdot\frac{(1-\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})^2\cdot\ln(\frac{2|S|\cdot|A|\cdot\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})}{\delta})}{2(1-\gamma)^2\varepsilon^2(1-\gamma^{-1})^2} \\
&= \gamma^3\cdot\frac{(1-\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})^2\cdot\ln(\frac{2|S|\cdot|A|\cdot\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})}{\delta})}{2(1-\gamma)^4\varepsilon^2}
\end{aligned}$$

Now, let's bound the number of calls to $PS(M)$ i.e. let's bound $m_D \times l_D$:

$$l_D \times m_D = [\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})] \times [\frac{\gamma^3\cdot(1-\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})^2\cdot\ln(\frac{2|S|\cdot|A|\cdot\log_\gamma(\frac{2(1-\gamma)\varepsilon}{\|v_1-v_0\|})}{\delta})}{2(1-\gamma)^4\varepsilon^2}]$$

$$
\begin{aligned}
&= O([\log_\gamma \varepsilon] \times [\frac{(1-\varepsilon)^2 \cdot \ln\left(\frac{|S|\cdot|A|\cdot\log_\gamma\varepsilon}{\delta}\right)}{\varepsilon^2}]) \\
&= O([\frac{\ln\varepsilon}{\ln\gamma}] \times [\frac{(1-\varepsilon)^2 \cdot \ln\left(\frac{|S|\cdot|A|\cdot\frac{\ln\varepsilon}{\ln\gamma}}{\delta}\right)}{\varepsilon^2}]) \\
&\quad \mathit{Since}\ \ln\gamma < 0 : \\
&= O([\ln\frac{1}{\varepsilon}] \times [\frac{(1-\varepsilon)^2 \cdot \ln\left(\frac{|S|\cdot|A|\cdot\ln\frac{1}{\varepsilon}}{\delta}\right)}{\varepsilon^2}]) \\
&\quad \mathit{w.l.g\ we\ can\ assume\ that}\ \varepsilon < 1,\ \mathit{and\ hence} : \\
&= O([\ln\frac{1}{\varepsilon}] \times [\frac{\ln\left(\frac{|S|\cdot|A|\cdot\ln\frac{1}{\varepsilon}}{\delta}\right)}{\varepsilon^2}]) \\
&= O([\ln\frac{1}{\varepsilon}] \times [\frac{1}{\varepsilon^2}\cdot(\ln(\frac{|S|\cdot|A|}{\delta}) + \ln\ln\frac{1}{\varepsilon})]) \\
&= O(\frac{1}{\varepsilon^2}\cdot\ln(\frac{|S|\cdot|A|}{\delta})\cdot\ln\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2}\cdot\ln\frac{1}{\varepsilon}\cdot\ln\ln\frac{1}{\varepsilon})
\end{aligned}
$$

So the number of calls to $PS(M)$ that insures with probability at least $1-\delta$ that the received policy in the $Phased - Q - Learning$ algorithm is within $\varepsilon$ distance from the optimal policy is $O(\frac{1}{\varepsilon^2}\cdot\ln\frac{|S|\cdot|A|}{\delta}\cdot\ln\frac{1}{\varepsilon}+\frac{1}{\varepsilon^2}\cdot\ln\frac{1}{\varepsilon}\cdot\ln\ln\frac{1}{\varepsilon})$

□

# 8 Proof for The Indirect Algorithm

## 8.1 Notations

- $\widehat{P}^a_{sj}$ is the *estimation* of the transition probability from states $s$ to $j$ by action $a$, calculated by the *IndirectAlgorithm*: $\widehat{P}^a_{sj} = \frac{\#(s \to_a j)}{m_I}$.
- $Q_l(s,a)$ is the value of state $s$ and action $a$ received from $l$ iterations of the $VI$ algorithm, using the real transition probabilities $(P^a_{sj})$.
- $\widehat{Q}_l(s,a)$ is the value of state $s$ and action $a$ received from $l$ iterations of the $VI$ algorithm, using the $\widehat{P}^a_{sj}$ estimation of the transition probabilities.

**Claim 8.1** *For any $w > 0$, $(s,a) \in S \times A$*

$$\Pr\left(\left\|\sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j)\right\| \geq w\right) \leq 2e^{-2\cdot m_I \cdot w^2} \tag{18}$$

**Proof:** The samples are independent and identically distributed; they are bounded since the immediate return is bounded and $\gamma < 1$; and:

$$\begin{aligned} E[\sum_{j\in S} \widehat{P}^a_{sj} \cdot v_k(j)] &= \sum_{j\in S} E[\widehat{P}^a_{sj} \cdot v_k(j)] \\ &= \sum_{j\in S} E[\frac{\#(s \to_a j)}{m_I}] \cdot v_k(j) \\ &= \sum_{j\in S} P^a_{sj} \cdot v_k(j) \end{aligned}$$

therefore, by Chernoff-inequality, the claim holds:

$$\begin{aligned} &\Pr\left(\left\|\sum_{j\in S} \widehat{P}^a_{sj} \cdot v_k(j) - \sum_{j\in S} P^a_{sj} \cdot v_k(j)\right\| \geq w\right) \\ = \ &\Pr\left(\left\|\sum_{j\in S} \widehat{P}^a_{sj} \cdot v_k(j) - E[\sum_{j\in S} \widehat{P}^a_{sj} \cdot v_k(j)]\right\| \geq w\right) \leq 2e^{-2\cdot m_I \cdot w^2} \end{aligned}$$

□

**Claim 8.2** *Bounding $\|\widehat{v}_l - v_l\|$:*

1. $\|\widehat{v}_l - v_l\| \leq \left\|\widehat{Q}_l - Q_l\right\|$
2. $\sum_{j\in S} \widehat{P}^a_{sj} = 1$
3. $\|\widehat{v}_l - v_l\| \leq \max_{s\in S,\ a\in A,\ 0\leq k\leq l} \left\|\sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j)\right\| \cdot \sum_{i=1}^{l} \gamma^i$

**Proof:**

- First, let's see that $\|\widehat{v}_l - v_l\| \leq \left\|\widehat{Q}_l - Q_l\right\|$:

$$\begin{aligned}
\|\widehat{v}_l - v_l\| &= \max_{s\in S} \|\widehat{v}_l(s) - v_l(s)\| \\
&= \max_{s\in S} \left\|\max_{a\in A} \widehat{Q}_l(s,a) - \max_{a\in A} Q_l(s,a)\right\| \\
&\leq \max_{s\in S} \max_{a\in A} \left\|\widehat{Q}_l(s,a) - Q_l(s,a)\right\| \\
&\leq \max_{(s,a)\in S\times A} \left\|\widehat{Q}_l(s,a) - Q_l(s,a)\right\| \\
&= \left\|\widehat{Q}_l - Q_l\right\|
\end{aligned}$$

- Secondly, $\sum_{j\in S} \widehat{P}^a_{sj} = 1$, since:

$$\begin{aligned}
\sum_{j\in S} \widehat{P}^a_{sj} &= \sum_{j\in S} \frac{\#(s \to_a j)}{m_I} \\
&= \frac{1}{m_I} \sum_{j\in S} \#(s \to_a j) \\
&= \frac{1}{m_I} \cdot m_I \quad (on\ each\ sampling\ there's\ some\ returned\ state\ j) \\
&= 1
\end{aligned}$$

- Thirdly, let's prove the bound for $\|\widehat{v}_l - v_l\|$ by induction:
For $l = 0$, $\|\widehat{v}_0 - v_0\| = 0$ and the bound holds.
Let assume the bound holds for $l-1$, and prove it for $l$:

$$\begin{aligned}
\|\widehat{v}_l - v_l\| &\leq \left\|\widehat{Q}_l - Q_l\right\| \\
&= \max_{(s,a)\in S\times A} \left\{ \left\| R_M{}^a(s) + \gamma \sum_{j\in S} \widehat{P}^a_{sj} \widehat{v}_{l-1}(j) - R_M{}^a(s) - \gamma \sum_{j\in S} P^a_{sj} v_{l-1}(j) \right\| \right\} \\
&\leq \max_{(s,a)\in S\times A} \{\gamma \left\| \sum_{j\in S} \widehat{P}^a_{sj} \widehat{v}_{l-1}(j) - \sum_{j\in S} \widehat{P}^a_{sj} v_{l-1}(j) \right\| + \\
&\qquad \gamma \left\| \sum_{j\in S} \widehat{P}^a_{sj} v_{l-1}(j) - \sum_{j\in S} P^a_{sj} v_{l-1}(j) \right\| \}
\end{aligned}$$

$$
\begin{aligned}
\leq\; & \max_{(s,a)\in S\times A}\{\gamma\sum_{j\in S}\widehat{P}^a_{sj}\left\|\widehat{v}_{l-1}(j)-v_{l-1}(j)\right\|+\\
& \gamma\left\|\sum_{j\in S}\widehat{P}^a_{sj}v_{l-1}(j)-\sum_{j\in S}P^a_{sj}v_{l-1}(j)\right\|\}\\
\leq\; & \max_{(s,a)\in S\times A}\left\{\gamma\sum_{j\in S}\widehat{P}^a_{sj}\left\|\widehat{v}_{l-1}-v_{l-1}\right\|+\gamma\left\|\sum_{j\in S}\widehat{P}^a_{sj}v_{l-1}(j)-\sum_{j\in S}P^a_{sj}v_{l-1}(j)\right\|\right\}\\
\leq\; & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\widehat{v}_{l-1}-v_{l-1}\right\|\sum_{j\in S}\widehat{P}^a_{sj}+\gamma\left\|\sum_{j\in S}\widehat{P}^a_{sj}v_{l-1}(j)-\sum_{j\in S}P^a_{sj}v_{l-1}(j)\right\|\right\}\\
=\; & \max_{(s,a)\in S\times A}\left\{\gamma\left\|\widehat{v}_{l-1}-v_{l-1}\right\|+\gamma\left\|\sum_{j\in S}\widehat{P}^a_{sj}v_{l-1}(j)-\sum_{j\in S}P^a_{sj}v_{l-1}(j)\right\|\right\}\\
& By\ induction's\ assumption:\\
\leq\; & \max_{(s,a)\in S\times A}\{\gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l}\left\{\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b{}_{tj}v_k(j)\right\|\cdot\sum_{i=1}^{l-1}\gamma^i\right\}+\\
& \gamma\left\|\sum_{j\in S}\widehat{P}^a_{sj}v_{l-1}(j)-\sum_{j\in S}P^a_{sj}v_{l-1}(j)\right\|\}\\
\leq\; & \max_{(s,a)\in S\times A}\{\gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l}\left\{\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b{}_{tj}v_k(j)\right\|\cdot\sum_{i=1}^{l-1}\gamma^i\right\}+\\
& \gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l-1}\left\{\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b_{tj}v_k(j)\right\|\right\}\}\\
=\; & \gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l}\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b{}_{tj}v_k(j)\right\|\cdot\sum_{i=1}^{l-1}\gamma^i+\\
& \gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l-1}\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b_{tj}v_k(j)\right\|\\
=\; & \max_{(t,b)\in S\times A,\ 0\leq k\leq l}\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b{}_{tj}v_k(j)\right\|\cdot\sum_{i=2}^{l}\gamma^i+\\
& \gamma\max_{(t,b)\in S\times A,\ 0\leq k\leq l}\left\|\sum_{j\in S}\widehat{P}^b_{tj}v_k(j)-\sum_{j\in S}P^b_{tj}v_k(j)\right\|
\end{aligned}
$$

$$\leq \max_{(t,b)\in S\times A,\ 0\leq k\leq l} \left\| \sum_{j\in S} \widehat{P}^b_{tj} v_k(j) - \sum_{j\in S} P^b{}_{tj} v_k(j) \right\| \cdot \sum_{i=1}^{l} \gamma^i$$

$\square$

**Claim 8.3** *As a conclusion from claims 8.1 and 8.2 we get that for every $w > 0$:*

$$\Pr(\|\widehat{v}_l - v_l\| \geq w) \leq l \cdot |S| \cdot |A| \cdot 2e^{-2m_I(\frac{w(1-\gamma)}{\gamma(1-\gamma^l)})^2} \tag{19}$$

**Proof:**

$$\Pr(\|\widehat{v}_l - v_l\| \geq w) \leq \Pr(\max_{(s,a)\in S\times A,\ 0\leq k\leq l} \left\| \sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j) \right\| \cdot \sum_{i=1}^{l} \gamma^i \geq w)$$

$$\begin{aligned}
&= \Pr(\exists (s,a) \in S \times A,\ 0 \leq k \leq l \left\| \sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j) \right\| \cdot \sum_{i=1}^{l} \gamma^i \geq w) \\
&\leq \sum_{(s,a)\in S\times A,\ 0\leq k\leq l} \Pr(\left\| \sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j) \right\| \cdot \sum_{i=1}^{l} \gamma^i \geq w) \\
&= \sum_{(s,a)\in S\times A,\ 0\leq k\leq l} \Pr(\left\| \sum_{j\in S} \widehat{P}^a_{sj} v_k(j) - \sum_{j\in S} P^a_{sj} v_k(j) \right\| \geq \frac{w}{\sum_{i=1}^{l} \gamma^i}) \\
&\quad By\ 8.1: \\
&\leq \sum_{(s,a)\in S\times A,\ 0\leq k\leq l} 2e^{-2\cdot m_I \cdot (\frac{w}{\sum_{i=1}^{l} \gamma^i})^2} \\
&= l \cdot |S| \cdot |A| \cdot 2e^{-2\cdot m_I \cdot (\frac{w}{\frac{\gamma - \gamma^{l+1}}{1-\gamma}})^2} \\
&= l \cdot |S| \cdot |A| \cdot 2e^{-2\cdot m_I \cdot (\frac{(1-\gamma)w}{\gamma(1-\gamma^l)})^2}
\end{aligned}$$

$\square$

**Claim 8.4** *As we've seen in class:*

$$\|v_l - v^*\| \leq \frac{\gamma^l}{1-\gamma} \tag{20}$$

**Claim 8.5** *For every* $l > \frac{\ln \varepsilon(1-\gamma)}{\ln \gamma}$,

$$\Pr(\|\widehat{v}_l - v^*\| \geq \varepsilon) \leq l \cdot \ |S| \cdot |A| \cdot 2e^{-2 \cdot m_I \cdot (\frac{(1-\gamma)(\varepsilon \frac{\gamma^l}{1-\gamma})}{\gamma(1-\gamma^l)})^2} \tag{21}$$

**Proof:**

$$\begin{aligned}
\Pr(\|\widehat{v}_l - v^*\| \geq \varepsilon) &\leq \Pr(\|\widehat{v}_l - v_l\| + \|v_l - v^*\| \geq \varepsilon) \\
&\leq \Pr(\|\widehat{v}_l - v_l\| + \frac{\gamma^l}{1-\gamma} \geq \varepsilon) \\
&= \Pr(\|\widehat{v}_l - v_l\| \geq \varepsilon - \frac{\gamma^l}{1-\gamma}) \\
&\quad \textit{as long as}: \ \varepsilon - \frac{\gamma^l}{1-\gamma} > 0: \\
&\leq l \cdot \ |S| \cdot |A| \cdot 2e^{-2 \cdot m_I \cdot (\frac{(1-\gamma)(\varepsilon - \frac{\gamma^l}{1-\gamma})}{\gamma(1-\gamma^l)})^2}
\end{aligned}$$

Let's see when does $\varepsilon - \frac{\gamma^l}{1-\gamma} > 0$:

$$\begin{aligned}
\varepsilon - \frac{\gamma^l}{1-\gamma} > 0 \quad &\Leftrightarrow \quad \varepsilon \cdot (1-\gamma) > \gamma^l \\
&\Leftrightarrow \quad \ln(\varepsilon \cdot (1-\gamma)) > l \cdot \ln \gamma \\
&\Leftrightarrow \quad l > \frac{\ln(\varepsilon \cdot (1-\gamma))}{\ln \gamma} \\
&\quad\quad (\textit{Since } \ln \gamma < 1 \textit{ due to the fact that } 0 < \gamma < 1)
\end{aligned}$$

And this is the claim's assumption, so the proof is complete. □

## 8.2 Bounding the Number of Calls to $PS(M)$

We've seen:
For every $l_I > \frac{\ln \varepsilon(1-\gamma)}{\ln \gamma}$,

$$\Pr(\|\widehat{v}_{l_I} - v^*\| \geq \varepsilon) \leq l_I \cdot \ |S| \cdot |A| \cdot 2e^{-2 \cdot m_I \cdot \frac{(1-\gamma)(\varepsilon - \frac{\gamma^{l_I}}{1-\gamma})}{\gamma(1-\gamma^{l_I})}^2} \tag{22}$$

Let's find $m_I$ and $l_I$ such that

$$\Pr(\|\widehat{v}_{l_I} - v^*\| \geq \varepsilon) \leq \delta \tag{23}$$

By the previous claims, it suffices to find $m_I$ and $l_I$ s.t.:

1. $l_I > \frac{\ln(\varepsilon(1-\gamma))}{\ln\gamma}$
2. $\delta = l_I \cdot \ |S| \cdot |A| \cdot 2e^{-2 \cdot m_I \cdot (\frac{(1-\gamma)(\varepsilon - \frac{\gamma^{l_I}}{1-\gamma})}{\gamma(1-\gamma^{l_I})})^2}$

Note that $log_\gamma 2 > 0$ (since $0 < \gamma < 1$). Then we can set (taking upper value if necessary):

$$\begin{aligned} l_I &= \frac{\ln(\varepsilon(1-\gamma))}{\ln\gamma} + log_\gamma 2 \\ &= \log_\gamma(\varepsilon(1-\gamma)) + log_\gamma 2 \\ &= \log_\gamma(2\varepsilon(1-\gamma)) \end{aligned}$$

The number of calls to $PS(M)$ is $m_I$:

$$\begin{aligned}
&\quad\textit{From the second constraint we get}: \\
m_I &= (\ln(2 \cdot l_I \cdot |S| \cdot |A|) - \ln\delta) \cdot \frac{\gamma^2(1-\gamma^{l_I})^2}{2(1-\gamma)^2(\varepsilon - \frac{\gamma^{l_I}}{1-\gamma})^2} \\
&= (\ln(2 \cdot \log_\gamma(2\varepsilon(1-\gamma)) \cdot |S| \cdot |A|) - \ln\delta) \cdot \frac{\gamma^2(1-\gamma^{l_I})^2}{2(1-\gamma)^2(\varepsilon - \frac{\gamma^{l_I}}{1-\gamma})^2} \\
&= \ln(\frac{2 \cdot \log_\gamma(2\varepsilon(1-\gamma)) \cdot |S| \cdot |A|}{\delta}) \cdot \frac{\gamma^2}{2(1-\gamma)^2} \cdot (\frac{1-\gamma^{l_I}}{\varepsilon - \frac{\gamma^{l_I}}{1-\gamma}})^2 \\
&\quad\textit{By substituting } l_I \textit{ with the expression we've received for } l_I: \\
&= \ln(\frac{2 \cdot \log_\gamma(2\varepsilon(1-\gamma)) \cdot |S| \cdot |A|}{\delta}) \cdot \frac{\gamma^2}{2(1-\gamma)^2} \cdot (\frac{1-\gamma^{\log_\gamma(2\varepsilon(1-\gamma))}}{\varepsilon - \frac{\gamma^{\log_\gamma(2\varepsilon(1-\gamma))}}{1-\gamma}})^2 \\
&= \ln(\frac{2 \cdot \log_\gamma(2\varepsilon(1-\gamma)) \cdot |S| \cdot |A|}{\delta}) \cdot \frac{\gamma^2}{2(1-\gamma)^2} \cdot \frac{(1-(2\varepsilon(1-\gamma)))^2}{(\varepsilon - \frac{(2\varepsilon(1-\gamma))}{1-\gamma})^2}
\end{aligned}$$

$$
\begin{aligned}
&= \ln\left(\frac{2\cdot\log_\gamma(2\varepsilon(1-\gamma))\cdot|S|\cdot|A|}{\delta}\right)\cdot\frac{\gamma^2}{2(1-\gamma)^2}\cdot\frac{(1-(2\varepsilon(1-\gamma)))^2}{\varepsilon^2(1-\frac{(2(1-\gamma))}{1-\gamma})^2}\\
&= \ln\left(\frac{2\cdot\log_\gamma(2\varepsilon(1-\gamma))\cdot|S|\cdot|A|}{\delta}\right)\cdot\frac{\gamma^2}{2(1-\gamma)^2}\cdot\frac{(1-(2\varepsilon(1-\gamma)))^2}{\varepsilon^2(1-2)^2}\\
&= \ln\left(\frac{2\cdot\log_\gamma(2\varepsilon(1-\gamma))\cdot|S|\cdot|A|}{\delta}\right)\cdot\frac{\gamma^2}{2(1-\gamma)^2}\cdot\frac{(1-(2\varepsilon(1-\gamma)))^2}{\varepsilon^2}\\
&= O((\ln\frac{\log_\gamma\varepsilon\cdot|S|\cdot|A|}{\delta})\cdot(\frac{1}{\varepsilon^2})\cdot(1-2\varepsilon(1-\gamma))^2)\\
&\quad Since\ we're\ interested\ in\ asymptotic\ measure\ (as\ \varepsilon\ and\ \delta\ go\ to\ zero),\\
&\quad we\ can\ assume\ that\ \varepsilon<2(1-\gamma)\ (because\ it's\ true\ for\ \varepsilon\ small\ enough),\\
&\quad therefore:\\
&= O((\ln\frac{\log_\gamma\varepsilon\cdot|S|\cdot|A|}{\delta})\cdot(\frac{1}{\varepsilon^2}))\\
&\quad Since\ \ln\gamma<0,\ we\ have:\ log_\gamma\varepsilon=\frac{\ln\varepsilon}{\ln\gamma}=O(\ln\frac{1}{\varepsilon}),\ therefore:\\
&= O(\ln\left(\frac{\ln\frac{1}{\varepsilon}\cdot|S|\cdot|A|}{\delta}\right)\cdot(\frac{1}{\varepsilon^2}))\\
&= O((\ln\left(\frac{|S|\cdot|A|}{\delta}\right)+\ln\ln\frac{1}{\varepsilon})\cdot(\frac{1}{\varepsilon^2}))\\
&= O(\frac{1}{\varepsilon^2}\cdot\ln\left(\frac{|S|\cdot|A|}{\delta}\right)+\frac{1}{\varepsilon^2}\cdot\ln\ln\frac{1}{\varepsilon})
\end{aligned}
$$

Here, again, $m_I$ and $l_I$ must be natural numbers, but taking their upper value won't change the asymptotic complexity, so, the number of calls to $PS(M)$ by the indirect algorithm is:
$O(\frac{1}{\varepsilon^2}\cdot\ln(\frac{|S|\cdot|A|}{\delta})+\frac{1}{\varepsilon^2}\cdot\ln\ln\frac{1}{\varepsilon})$

# 9 Conclusions

We saw that both Phased-Q-Learning and the indirect algorithm enjoy a rather rapid convergence to the optimal policy as a function of the number of observed transitions. Both have roughly the same sampling complexity, with a slight advantage to the indirect algorithm. This advantage is rather surprising since this sampling complexity is not enough to enable the *construction* of a good model of the given MDP.