Constrained Markov Decision Process and Optimal Policies

Wei Huang, Jun Zhang

April 16, 2012

Contents

Abstract

In the course lectures, we have discussed a lot regarding unconstrained Markov Decision Process (MDP). The dynamic programming decomposition and optimal policies with MDP are also given. However, in this report we are going to discuss a different MDP model, which is constrained MDP. There are many realistic demand of studying constrained MDP. For instance, in the wireless sensors networks, each sensor need to decide whether or not (1 or 0) to report its observation to the sink node. The policy of choosing action at each sensor should not only be based on observations and past actions, but also left battary. In these kind of application scenarios with constraint, to derive the optimal policies, constraint should be put into consideration.

I Introduction

The material presented in this reported mainly from the [1] and the lecture notes of ECSE509.In the lectures, we have already discussed about infinite horizon MDP with average cost. The model about to be discussed would be the same, however with an extra constraint when deriving polcies. Some proofs in original paper which are almost the same as lecture notes will be omitted.

I.1 System Model

Assume there is a system, which has finite state space $S = \{0, 1, 2, \ldots, N\}$. An controller will decide the action A_t in each state based on the past observations and actions $H_{t-1} = (X_0, A_0, X_1, A_2, \ldots, X_{t-1}, A_{t-1})$. The action space is denoted as $\mathbf{A}$. We are interested in the controlled Markov process, that is state X_{t+1} depends only on X_t and A_t. That is,

$$P(X_{t+1} = y|H_{t-1}, X_t = x; A_t = a) = P(X_{t+1} = y|X_t = x; A_t = a) \tag{1}$$

At each epoch t, there is a incurred reward C_t depends on the state X_t and action A_t. Assume the system horizon is infinite and consider average cost. The optimal control policy

should attain the following equation.

$$R_x(u) = \liminf_n n^{-1} E_u\Big[\sum_{k=0}^{n-1} C(X_k, A_k)|X_0 = x\Big], \tag{2}$$

For the reward, it is more common to adopt supremum, however in this report we are interested in the worst-performance reward.

Aide from the reward, the system will incur a cost denoted as $D(X_k, A_k)$.

$$K_x(u) = \limsup_n n^{-1} E_u\Big[\sum_{k=0}^{n-1} D(X_k, A_k)|X_0 = x\Big], \tag{3}$$

The objective here is even under worst situation, the system cost should be less than a prefixed value α. That is,

$$K_x(u) \leq \alpha \tag{4}$$

for any x.

The reason that an extra constraint is put here is sometimes we want to guarantee even in worst situation, the system can be controlled to work properly. Take the wireless sensor networks example agian, we want to maximize the battery life of each node, so the sink can have observations from as many sensor as possible.

Assume $\mathbf{U}_0$ is a subset of $\mathbf{U}$, whose control policies are able to meet the constraint (4). We want a policy from $\mathbf{U}_0$ can attain following equation.

$$R_x = \sup_{u \in \mathbf{U}_0} R_x(\mathbf{u}) \tag{5}$$

A policy which can attain R_x and at the same time satifying (4) is named as optimal policy.

There are three different kind of control polcies. A general control policy space $\mathbf{U}$, which depends all past observations and control actions, $\mathbf{u} = \{u_0, u_1, \ldots\} \in \mathbf{U}$ and $u_t = g_t(H_{t-1}, X_t = x)$.

As we are studying controlled MDP, optimal policies have been proved in the class to depend only on previous state and action. This kind of Markov policies are belonging to

a space **F**, which is a subspace of **U**. Any policy $\mathbf{f} = \{f_0, f_1, \ldots\} \in \mathbf{F}$ can be represented as $f_t = g_t(X_{t-1}, A_{t-1})$.

A more restricted simple or non-randomized stationary policy is represented as **G**, which is the subspace of both **F** and **U**. G can be characterized by a simple mapping $g : S \to \mathbf{A}$ and $g(x)$ acquires meaning as an element of **A** and is viewed as a deterministic vector.

In the later proof we need a mixed policy, whose space is denoted as $\mathbf{F}_m$. A mixed policy $\mathbf{f}_q$ is a stationary policy that randomize between two simple policies $\mathbf{g}_1$ and $\mathbf{g}_2$. $\mathbf{f}_q = q\mathbf{g}_1 + (1-q)\mathbf{g}_2$, with $q \in [0, 1]$.

I.2 Preliminaries and Unconstrained Problem

In this section, some important results and assumptions from lecture notes related with average cost MDP are listed here.

Assumption 1: Here we assume that process $\{X_n\}$ is irreducible and only have one recurrent class. Then we will have the hitting time for a particular state y starting from state $x \in S$ is finite.

$$\sup_{x \in S} \sup_{\mathbf{g} \in \mathbf{G}} E_g(T_x) < \infty \tag{6}$$

Remark It has been proved in the paper that above equation will also hold in the case $\mathbf{f} \in \mathbf{F}$

Assumption 2: Under all Markov policies **g**, Markov Chain $\mathbf{P}^{\mathbf{g}}$ is irreduciable and non-periodic and steady state $\mathbf{P}^*(\mathbf{g})$ equal to

$$\mathbf{P}^*(\mathbf{g}) = \lim_n \mathbf{P}_0 \mathbf{P}^n \tag{7}$$

where $\mathbf{P}_0$ is the initial state.

Remark Although we are going to discuess only non-periodic case. If in the periodic case, we apply the Cesaro means as following

$$\hat{\mathbf{P}}^* = \frac{1}{n} \sum_{n=0}^{\infty} \mathbf{P}^n \tag{8}$$

According to [3] Theroem A.2, for irreducible transition matrix, $\hat{\mathbf{P}}^*$ will also reach a unique positive stationary solution, whose each row π is also indentical and meets $\pi\mathbf{e} = 1$. So any results in this report based on non-periodic assumption will also apply for periodic case.

Lemma 1. **F** *and* **G** *are sequentially compact. Both* **P** *and* $\mathbf{P}^*$ *are continuous functions on* **F**.

Proof. Since **A** *is compact, therefore* **F** *are tight.Since* **G** *belong to* **F**. **G** *is closed and* **A** *is compact, so* **G** *is also sequentially compact. According to weak convergence of probability measures,* **P** *is continuous functions on* **F**.

Assume $\mathbf{P}^*$ *is not continuous function on* **F**. *Assume for some* $\mathbf{f}_n \to \mathbf{f}_0$ *there is a subsequence* $\mathbf{f}_m$ *belonging to* $\mathbf{f}_n$, *which cannot obtain* $\mathbf{f}_m \to \mathbf{f}_0$. *However, according to the properties of* $\mathbf{P}^*$, *we have* $\mathbf{P}^*(\mathbf{f_m})\mathbf{P}(\mathbf{f_m}) = \mathbf{P}^*(\mathbf{f_m})$, *which implies* $\mathbf{P}^*\mathbf{P}(\mathbf{f_0}) = \mathbf{P}^*$. *Recall that there is only one recurrent class. Here, we have two and we got a contradiction. Hence,* $\mathbf{P}^*$ *is also continuous function on* **F**

Theorem 2. *Assume* **S** *be finite and* **A** *compact. Further assume state* 0 *is accessible from each* $x \in \mathbf{S}$. *Then we will have*

$$\sup_{x\in\mathbf{S}} \sup_{\mathbf{g}\in\mathbf{G}} E_{\mathbf{g}}(T_x) < \infty \tag{9}$$

and

$$\sup_{x\in\mathbf{S}} \sup_{\mathbf{f}\in\mathbf{F}} E_{\mathbf{f}}(T_x) < \infty \tag{10}$$

Proof.

$$
\begin{aligned}
E_{\mathbf{f}}(T_x) &= \sigma[knP(T_x = kn)) = \sigma[P(T_x > k(n-1)) - P(T_x > kn)]kn \\
\sup E_{\mathbf{f}}(T_x) &= \sigma[\sup(P(T_x > k(n-1))) - \sup(P(T_x > kn))]kn \\
&\quad Since \sup(P(T_x > kn)) \leq \beta^k \\
&\leq \sigma[\beta^{k-1} - \beta^k]kn \\
&\leq n(\beta^0 + \beta^1 + \ldots + \beta^k) \\
&\leq n(1-\beta^k)/(1-\beta) \;\; since\; n(1-\beta^k) < n
\end{aligned}
\tag{11}
$$

Thus we have

$$\sup E_{\mathbf{f}}(T_x) \leq \frac{n(1-\beta^k)}{(1-\beta)} < \frac{n}{1-\beta} < \infty$$

Theorem 3. *Suppose there exists a scalar c and a bounded vector* $\boldsymbol{h}$ *such that the DPE*

$$c + h(x) = \sup_{a \in \boldsymbol{A}}[C(x,a) + \sum_{y \in S} P_{xy}(a)h(y)] \tag{12}$$

is satisfied for each $x \in S$*. Then any policy* $g \in G$ *specified by*

$$g(x) = \arg\sup_{a \in \boldsymbol{A}}[C(x,a) + \sum_{y \in S} P_{xy}(a)h(y)] \tag{13}$$

attains

$$J = \sup_{u \in \boldsymbol{U}} R_x(\boldsymbol{u}) \tag{14}$$

The proof of this theorem is the same as we did for unconstrained MDP, hence is omitted here.

Write equation (13) into vector form and set right side equal to its supremum, we will have

$$J\mathbf{e} + \mathbf{h}(\hat{\mathbf{g}}) = \mathbf{C}(\hat{\mathbf{g}}) + \mathbf{P}(\hat{\mathbf{g}})\mathbf{h}(\hat{\mathbf{g}}) \tag{15}$$

Premultiply above equation by $\mathbf{P}^*(\hat{\mathbf{g}})$, we will get

$$
\begin{aligned}
\mathbf{P}^*(\hat{\mathbf{g}})J\mathbf{e} + \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{h}(\hat{\mathbf{g}}) &= \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{C}(\hat{\mathbf{g}}) + \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{P}(\hat{\mathbf{g}})\mathbf{h}(\hat{\mathbf{g}}) \\
\mathbf{P}^*(\hat{\mathbf{g}})J\mathbf{e} &= \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{C}(\hat{\mathbf{g}}) \\
J\mathbf{e} &= \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{C}(\hat{\mathbf{g}})
\end{aligned}
\tag{16}
$$

The second equation is due to $\mathbf{P}^*(\hat{\mathbf{g}})\mathbf{h}(\hat{\mathbf{g}}) = \mathbf{P}^*(\hat{\mathbf{g}})\mathbf{P}(\hat{\mathbf{g}})\mathbf{h}(\hat{\mathbf{g}})$. The last equation is because $\mathbf{P}^*(\hat{\mathbf{g}})$ is the steady state, each row of which is euqal and satisfies $\pi\mathbf{e} = \mathbf{1}$. These results are from the lecture notes.

Lemma 4. *Let the DPE be satisfied. Then the* $\boldsymbol{h}$ *appearing in (13) is a constant vector.*

II Lagrange Formulation

We first restate the problem as following.

$$R_x(\mathbf{u}) = \liminf_n n^{-1} E_{\mathbf{u}}\Big[\sum_{k=0}^{n-1} C(X_k, A_k)|X_0 = x\Big],\ s.t\ \ K_x(\mathbf{u}) \le \alpha \tag{17}$$

Under the help of Lagrange multiplier, above question (17) is possible be translated into an unconstrained dynamic programming equation with a parameter λ.

$$J_x^\lambda(\mathbf{u}) = \liminf_n n^{-1} E_{\mathbf{u}}\Big[\sum_{k=0}^{n-1} B^\lambda(X_k, A_k)|X_0 = x\Big], \tag{18}$$

where

$$B^\lambda(x, a) = C(x, a) - \lambda D(x, a) \tag{19}$$

Based on assumptions, results of previous section and lecture notes regarding uncontrainted DPE, above problem would have at least one solution $\mathbf{g}^\lambda \in \hat{\mathbf{G}}^\lambda$. Therefore, the supremum $J^\lambda = \sup_{\mathbf{u}} J_x^\lambda(\mathbf{u})$ is attained by at least one $\mathbf{g}^\lambda$. Also, according to the accessibility hypothesis, the initial state has no impact here. Before trying to derive the optimal policy regarding this new DPE, there are some necessary properties of J^λ need to be proved.

Define a new notation $\hat{\mathbf{G}}^\lambda$ which denote those $\mathbf{g} \in \mathbf{G}$ satisfying the constrained DPE with parameter λ. Then let,

$$\hat{\mathbf{G}}_\eta = \cup_{\lambda \le \eta}(\lambda, \hat{\mathbf{G}}) \tag{20}$$

Lemma 5. *J^λ, R^λ, and K^λ are all monotone non-increading in λ, where $J^\lambda = J^\lambda(\boldsymbol{g}^\lambda)$, $R^\lambda = R^\lambda(\boldsymbol{g}^\lambda)$ and $K^\lambda = K^\lambda(\boldsymbol{g}^\lambda)$.*

Proof:

$$\begin{aligned} J^{\lambda+\eta}(\boldsymbol{g}^{\lambda}) - J^{\lambda}(\boldsymbol{g}^{\lambda}) &\leq J^{\lambda+\eta}(\boldsymbol{g}^{\lambda+\eta}) - J^{\lambda}(\boldsymbol{g}^{\lambda}) \\ &\leq J^{\lambda+\eta}(\boldsymbol{g}^{\lambda+\eta}) - J^{\lambda}(\boldsymbol{g}^{\lambda+\eta}) \\ &\leq 0 \end{aligned} \tag{21}$$

The first inequality holds because $\boldsymbol{g}^{\lambda+\eta}$ is the optimal solution of supremum $J^{\lambda+\eta}$.

The second inequality holds because $J^{\lambda}(\boldsymbol{g}^{\lambda+\eta}) \leq J^{\lambda}(\boldsymbol{g}^{\lambda})$. The third inequality is derived as following:

$$\begin{aligned} J^{\lambda+\eta}(\boldsymbol{g}^{\lambda+\eta}) - J^{\lambda}(\boldsymbol{g}^{\lambda+\eta}) &= \liminf_{n} n^{-1} E_{\boldsymbol{u}=\boldsymbol{g}^{\lambda+\eta}}\Big[\sum_{k=0}^{n-1} -\eta D(X_k, A_k) | X_0 = x\Big] \\ &= -\eta K^{\lambda+\eta} \\ &\leq 0 \end{aligned} \tag{22}$$

where the last inequality is from the positiveness of $D(x, a)$. Hence, $J^{\lambda+\eta}(\boldsymbol{g}^{\lambda+\eta}) \leq J^{\lambda}(\boldsymbol{g}^{\lambda})$ and so J^{λ} is monotome non-increasing in λ.

$$\begin{aligned} J^{\lambda+\eta}(\boldsymbol{g}^{\lambda}) - J^{\lambda}(\boldsymbol{g}^{\lambda}) &= \liminf_{n} n^{-1} E_{\boldsymbol{u}}\Big[\sum_{k=0}^{n-1} -\eta D(X_k, A_k) | X_0 = x\Big] \\ &= -\eta K^{\lambda} \\ &\leq 0 \end{aligned} \tag{23}$$

Based on (21)-(23), we can conclude that K^{λ} is also monotone non-increasing on λ because,

$$\begin{aligned} -\eta K^{\lambda} &\leq -\eta K^{\lambda+\eta} \\ K^{\lambda+\eta} &\leq K^{\lambda} \end{aligned} \tag{24}$$

For R_{λ}, assume it is not monotone non-increasing. Then $R^{\lambda} < R^{\lambda+\eta}$. Based on equation (24), following inequations holds.

$$\begin{aligned} R^{\lambda} - \lambda K^{\lambda} &< R^{\lambda+\eta} - \lambda K^{\lambda+\eta} \\ J^{\lambda}(\boldsymbol{g}^{\lambda}) &< J^{\lambda}(\boldsymbol{g}^{\lambda+\eta}) \end{aligned} \tag{25}$$

It is a obvious contradiction of last inequality, since $J^{\lambda}(\boldsymbol{g}^{\lambda+\eta}) \leq J^{\lambda}(\boldsymbol{g}^{\lambda})$.

Lemma 6. *J^λ is uniformly absolutely continuous with*

$$-K^\lambda \leq \Big(\frac{dJ^\lambda}{d\lambda}\Big)^+ \leq -\lim_{\eta \downarrow 0} K^{\lambda+\eta} \tag{26}$$

Also, the derivative

$$\frac{dJ^\lambda}{d\lambda} = -K^\lambda \tag{27}$$

exists for almost all $\lambda \geq 0$.

Proof:

Based on the equations (21) and (24), we will have

$$|J^{\lambda+\eta} - J^\lambda| \leq \eta K^\lambda \leq \eta K^0$$

As K^λ is monotone non-increasing in λ. Also due to $D(x, a)$ is a continuous function on a compact set, so K^0 is also bounded. As K^λ is continuous almost everywhere. From equation (21) we have $-\eta K(\lambda) \leq J^{\lambda+\eta} - J^\lambda \leq -\eta K^{\lambda+\eta} \leq 0$, divide it by η, then K^λ possesses limits from the right and obtains an equality in (26) for almost all λ. The right derivative should coincide with ordinary derivative by the absolute continuity. Hence we can have equ.(26). Then according to squeeze theorem, we can get equ.(27) from equ.(26).

We know in the normal constrained optimization problem, the reason we introduce Lagrange multiplier is to let the optimiztion problem meet the constraint. Here we define a γ with the similar functionality.

$$\gamma = \inf\{\lambda : K^\lambda \leq \alpha\} \tag{28}$$

Lemma 7. *Let $K(\boldsymbol{g}) \leq \alpha$ for some $\boldsymbol{g} \in \boldsymbol{G}$. Then $\gamma \leq \infty$.*

The proof in original paper is not quite clear. Here we give a more detailed proof, which is still based on the contradiction.

Assume the claim is false, so that $\gamma = \infty$ and for any λ we will have $K^\lambda > \alpha$. By definition

$$\begin{aligned} J^\lambda \triangleq J^\lambda(\boldsymbol{g}^\lambda) &= R^\lambda - \lambda K^\lambda \\ &< R^\lambda - \lambda\alpha, \ as \ K^\lambda > \alpha \\ &< R^0 - \lambda\alpha, \ as \ R^\lambda \ is \ non - increasing \end{aligned} \tag{29}$$

Also, by the assumption of the theorem, $K(\boldsymbol{g}) \leq \alpha$ for some $\boldsymbol{g} \in \boldsymbol{G}$. Therefore, $\exists\delta > 0$, such that $K(\boldsymbol{g}) = \alpha - \delta$ for this $\boldsymbol{g}$. Then we have

$$\begin{aligned} J(\boldsymbol{g}) &= R(\boldsymbol{g}) - \lambda K(\boldsymbol{g}) \\ &= R(\boldsymbol{g}) - \lambda(\alpha - \delta) \\ &= R(\boldsymbol{g}) - \lambda\alpha + \lambda\delta) \end{aligned} \tag{30}$$

Then, for any λ

$$\begin{aligned} J(\boldsymbol{g}) - J(\boldsymbol{g}^\lambda) &= R(\boldsymbol{g}) - \lambda\alpha + \lambda\delta - (R^\lambda - \lambda K^\lambda) \\ &= [R(\boldsymbol{g}) - R(\boldsymbol{g}^\lambda)] + [\lambda(K^\lambda - \alpha)] + \lambda\delta \end{aligned} \tag{31}$$

Case I:If the first term in above equation is postive, then $J(\boldsymbol{g}) - J(\boldsymbol{g}^\lambda) > 0$ for all λ.

Case II: If the first term is negative, then pick λ such that $\lambda \geq \frac{R(\boldsymbol{g}^\lambda - R(\boldsymbol{g}))}{\delta}$, then we also have $J(\boldsymbol{g}) - J(\boldsymbol{g}^\lambda) > 0$. What we can conclude from Case I and Case II is as long as λ is sufficiently large $\exists\lambda$ such that $J(\boldsymbol{g}) - J(\boldsymbol{g}^\lambda) > 0$, which is cleary a contradiction to the fact that

$$\boldsymbol{g}^\lambda = \arg\max_{\boldsymbol{g}\in\boldsymbol{G}}\{J(\boldsymbol{g})\}$$

. Therefore, we conclude that Lemma is correct.

Lemma 8. *$R(\boldsymbol{g})$ and $K(\boldsymbol{g})$ are continuous on G and $J^\lambda(\boldsymbol{g})$ is continuous on $\boldsymbol{R}^+ \times \boldsymbol{G}$.*

Proof: $C(x,a)$, $D(x,a)$ and $\boldsymbol{P}^*$ are continuous, which implies $\mathbf{P}^*(\mathbf{g_n})\mathbf{C}(\mathbf{g_n}) \to \mathbf{P}^*(\mathbf{g_n})\mathbf{C}(\mathbf{g_n})$ and $\mathbf{P}^*(\mathbf{g_n})\mathbf{D}(\mathbf{g_n}) \to \mathbf{P}^*(\mathbf{g_n})\mathbf{D}(\mathbf{g_n})$. Since $\mathbf{R} = \mathbf{P}^*(\mathbf{g_n})\mathbf{C}(\mathbf{g_n})$, $\mathbf{K} = \mathbf{P}^*(\mathbf{g_n})\mathbf{D}(\mathbf{g_n})$ and $J^\lambda(\boldsymbol{g}) = R(\boldsymbol{g}) - \lambda K(\boldsymbol{g})$, $\mathbf{J}^\lambda(\mathbf{g_n}) \to \mathbf{J}^\lambda(\mathbf{g_0})$. Therefore, $\mathbf{J}^\lambda(\mathbf{g})$ is continuous on $\mathbf{R}^+ \times \mathbf{G}$.

Theorem 9. *For any η, space $\hat{\boldsymbol{G}}_\eta$ is compact.*

Proof: Since $\hat{\boldsymbol{G}}_\eta$ is a subspace of $([0,\eta]\times \boldsymbol{G})$, it is already totally bounded. What we still need to show is it is closed. Define a sequence $\lambda_n \to \lambda_0$, and assume $\boldsymbol{g}^{\lambda_n} \to \boldsymbol{g}_0$, with $\boldsymbol{g}^{\lambda_n} \in \hat{\boldsymbol{G}}^{\lambda_n}$. We must show that $\boldsymbol{g}_0 \in \hat{\boldsymbol{G}}^{\lambda_0}$. From the lecture notes, given an average cost infinite horzion MDP whose cost function is $B(x,a)$, we have

$$J^{\lambda_n}\boldsymbol{e} + \boldsymbol{h}^{\lambda_n} = \boldsymbol{B}^{\lambda_n}(\boldsymbol{g}^{\lambda_n}) + \boldsymbol{P}(\boldsymbol{g}^{\lambda_n})\boldsymbol{h}^{\lambda_n} \tag{32}$$

Each term in above equation converges to their limit since they are all continuous , that is

$$J^{\lambda_0}\boldsymbol{e} + \boldsymbol{h}^{\lambda_0} = \boldsymbol{B}^{\lambda_0}(\boldsymbol{g}^{\lambda_0}) + \boldsymbol{P}(\boldsymbol{g}^{\lambda_0})\boldsymbol{h} \tag{33}$$

Note that the continuity of $\boldsymbol{h}$ is from directly of equation Lemma 4. What still remains to show is that right hand side of maximal. First fix x and define a function $f(n,a)$ as follows

$$f(n,a) \triangleq C(x,a) - \lambda_n D(x,a) + \sum_{y\in\mathbf{S}} P_{xy}(a)h^{\lambda_n}(y) \tag{34}$$

Then the x coordinate on the right side of (33) reads $\lim_n \sup_a f(n,a)$. However, $f(n,a)$ converges uniformly in n with respect to $\mathbf{A}$. We also see that $f(n,)$ is uniformly continuous, and not that $\{\boldsymbol{g}_n(x)\}$ converges. These facts enable us to conclude that

$$\lim_n \sup_{a\in\mathbf{A}} f(n,a) = \sup_{a\in\mathbf{A}} \lim_n f(n,a) \tag{35}$$

which shows that the right side of (33) is the supremum over $\mathbf{A}$ for each $x \in \mathbf{S}$

III The Optimal Policy

In previous sections, some important properties regarding the newly defined DPE has been shown and proved. These properties are necesary to derive the optimal policy for the constrained dynamic programming problem.

Consider the policies in the general policy space U. There might be some policies are optimal to the unconstrained problem however fail to satisfy the constraint $K_x(u) \le \alpha$. There might be also policies inside space U, which meets the constraint howover they might not

be able to attain the largest reward J. An intuitive guess about the optimal policy would be certain randomization between these two policies, as continuity of J^λ and compactness of $\hat{G}_\eta$ have been proved respectively. An important assumption is made here before deriving optimal policies.

Assumption 3: Assume there at least exist policies $\mathbf{g}^0$, which is an unconstrained supremum and defined in the following way. $\mathbf{g}^0$ fails to meet the constraint.

$$\begin{aligned} \mathbf{g}^0 &= \arg\sup_{\mathbf{g}\in\mathbf{G}} R(\mathbf{g}) \\ K(\mathbf{g}^0) &> \alpha \end{aligned} \tag{36}$$

Suppose further there exists a $\mathbf{g} \in \mathbf{G}$ such that

$$K(\mathbf{g}) < \alpha \tag{37}$$

Theorem 10. *Suppose that for some $\lambda \geq 0$ and some $\mathbf{f} \in \mathbf{F}$ we have $K(\mathbf{f}) = \alpha$ and $J^\lambda(\mathbf{f}) = \mathbf{J}^\lambda$ for all $x \in S$. Therefore*

$$R(\boldsymbol{f}) \geq R_x(\boldsymbol{u}) + \lambda[\alpha - K_x(\boldsymbol{u})] \tag{38}$$

Proof. As $J^\lambda(\boldsymbol{f}) = \sup_{\boldsymbol{u}} J_x^\lambda(\boldsymbol{u})$, we have $J^\lambda(\boldsymbol{f}) \geq J_x^\lambda(\boldsymbol{u})$ for all $\boldsymbol{u}$ and x. Since $J^\lambda = R(\boldsymbol{f}) - \lambda K(\boldsymbol{f}) = R(\boldsymbol{f}) - \lambda\alpha$ and $J_x^\lambda(\boldsymbol{u}) = R_x(\boldsymbol{u}) - \lambda K_x(\boldsymbol{u})$, it is easy to attain

$$R(\boldsymbol{f}) \geq R_x(\boldsymbol{u}) + \lambda[\alpha - K_x(\boldsymbol{u})] \tag{39}$$

If $\mathbf{u} \in \mathbf{U_0}$, where U_0 is the policy space attains the constraint $K_x(\boldsymbol{u}) \leq \alpha$, so the second term of right-hand side is positive. Therefore, $R(\boldsymbol{f}) \geq R_x(\boldsymbol{u})$ for each $\mathbf{u} \in \mathbf{U_0}$ and $x \in \boldsymbol{S}$.

Remark Theorem 10 indicates that if such λ and corresponding $\mathbf{g}^\lambda$ exist, the optmial policy in fact is a simple policy belonging to $\mathbf{G}$. The remaining question now is what will happen when there is no such λ exists? Following theorem will discuss such case and show that even in this case it is possible for us to construct a mixed policy $\mathbf{f_q}$, which will randomize between two simple policies.

Theorem 11. *If Assumption 3 made in the beginning of this section holds, there exists a constrained optimal policy in mixed policy space* $\mathbf{F_m}$.

Proof:

Case I: if $K^\lambda = \alpha$ *for some* λ, *any corresponding* $\mathbf{g}^\lambda \in \hat{\mathbf{G}}$ *satisfies the conditions of Theorem 10, and is therefore optimal policy. Note that* $\hat{\mathbf{G}}$ *is the policy space for DPE (18).*

Case II: Suppose no such λ *as the above exists. Since* K^λ *is non-increasing and* $\gamma \in (0, \infty)$, *we have*

$$\begin{aligned} \lim_{\lambda \uparrow \gamma} K^\lambda &= \alpha^0 \\ \lim_{\lambda \downarrow \gamma} K^\lambda &= \alpha_0 \end{aligned} \tag{40}$$

in which $\alpha_0 < \alpha < \alpha^0$. *Let* $\{\lambda_n^+\}$ *be a sequence that increases to* γ, *along which the corresponding* $\mathbf{g}^{\lambda_\mathbf{n}^+} \in \hat{\mathbf{G}}$ *converges, since* $\boldsymbol{G}$ *is compact. Theorem 9 has proved that* $\hat{G}_\gamma$ *is also compact. therefore* $\bar{\mathbf{g}} = \lim \mathbf{g}^{\lambda_\mathbf{n}^+} \in \hat{\mathbf{G}}_\gamma$ *and* $K(\bar{g}) = \alpha^0$. *In similar the decreasing sequence* $\{\lambda_n^-\}$ *yields* $\underline{g} = \lim g^{\lambda_n^-} \in \hat{G}_\gamma$ *with* $K(\bar{\boldsymbol{g}}) = \alpha_0$. *Recall in the beginning we define a mixed policy space* $\mathbf{F_m}$, *whose element is named as* $\mathbf{f_q}$.

Let

$$\boldsymbol{f}_q = q\underline{\boldsymbol{g}} + (1-q)\bar{\boldsymbol{g}} \tag{41}$$

Whether this $\mathbf{f_q}$ ***satisfies the conditions in Theroem 10 and therefore is an optimal policy?***

To answer above question, what need to be shown is whether able to find a $q \in [0, 1]$, *such that* $J^\gamma = J^\gamma(\boldsymbol{f}_q)$ *and* $K(\boldsymbol{f}_q) = \alpha$.

Since J^λ *has been proved to be contious. Hence,* $J^\gamma = J^\gamma(\boldsymbol{g}^\gamma) = J^\gamma(\underline{\boldsymbol{g}}) = J^\gamma(\bar{\boldsymbol{g}})$. *It is also straightforward that* $\boldsymbol{B}^\gamma(\boldsymbol{f}_q) = q\boldsymbol{B}^\gamma(\underline{\boldsymbol{g}}) + (1-q)\boldsymbol{B}^\gamma(\bar{\boldsymbol{g}})$ *and* $\boldsymbol{P}(\boldsymbol{f}_q) = q\boldsymbol{P}(\underline{\boldsymbol{g}}) + (1-q)\boldsymbol{P}(\bar{\boldsymbol{g}})$. *Moreover, (4) has shown that* $\boldsymbol{h}^\gamma$ *is the same for* $\underline{\boldsymbol{g}}$ *and* $\bar{\boldsymbol{g}}$. *Therefore, we have*

$$\begin{aligned} J^\gamma \boldsymbol{e} + \boldsymbol{h}^\gamma &= \boldsymbol{B}^\gamma(\underline{\boldsymbol{g}}) + \boldsymbol{P}(\underline{\boldsymbol{g}})\boldsymbol{h}^\gamma \\ J^\gamma \boldsymbol{e} + \boldsymbol{h}^\gamma &= \boldsymbol{B}^\gamma(\bar{\mathbf{g}}) + \boldsymbol{P}(\bar{\mathbf{g}})\boldsymbol{h}^\gamma \end{aligned} \tag{42}$$

Premultiplying the first term in (42) by q and second term y $1-q$ *and then adding them together,*

we will get

$$J^{\gamma}\boldsymbol{e} + \boldsymbol{h}^{\gamma} = \boldsymbol{B}^{\gamma}(\mathbf{f_q}) + \boldsymbol{P}(\mathbf{f_q})\boldsymbol{h}^{\gamma} \tag{43}$$

Premultiplying both side by $\mathbf{P}^*(\mathbf{f_q})$*, we can attain following equation*

$$\boldsymbol{P}^*(f_q)J^{\gamma}\boldsymbol{e} + \boldsymbol{P}^*(f_q)\boldsymbol{h}^{\gamma} = \boldsymbol{P}^*(f_q)\boldsymbol{B}^{\gamma}(f_q) + \boldsymbol{P}^*(f_q)\boldsymbol{P}(f_q)\boldsymbol{h}^{\gamma} \tag{44}$$

since $\boldsymbol{P}^*(\boldsymbol{g}) = \lim_n \boldsymbol{P}_0\boldsymbol{P}^n$*, then* $\boldsymbol{P}^*(f_q)\boldsymbol{h}^{\gamma} = \boldsymbol{P}^*(f_q)\boldsymbol{P}(f_q)\boldsymbol{h}^{\gamma}$ *Then above equation can be simplified as*

$$\begin{aligned} \boldsymbol{P}^*(f_q)J^{\gamma}\boldsymbol{e} &= \boldsymbol{P}^*(f_q)\boldsymbol{B}^{\gamma}(f_q) \\ J^{\gamma} &= \boldsymbol{P}^*(f_q)\boldsymbol{B}^{\gamma}(\mathbf{f_q}) \\ &= J^{\gamma}(\mathbf{f_q}) \end{aligned} \tag{45}$$

The above second equation is because $\mathbf{P}^*$ *is the steady state, each row of which is euqal and satisfies* $\pi\boldsymbol{e} = 1$*. Last equation is due to similar argument used when we are deriving equation (16). Therefore, we have shown* $J^{\gamma} = J^{\gamma}(\mathbf{f_q})$*, and now what remains to be shown is whether we can get* $K(f_q) = \alpha$*.*

Since $\boldsymbol{f}_q$ *is continuous on* q*,* $\boldsymbol{D}^{\gamma}(f_q) = q\boldsymbol{D}^{\gamma}(\underline{g}) + (1-q)\boldsymbol{D}^{\gamma}(\bar{g})$*,* $\boldsymbol{P}^*(f_q)$ *and* $\boldsymbol{P}(f_q)$ *are all continuous. Recall that*

$$K(f_q)\boldsymbol{e} = \boldsymbol{P}^*(f_q)\boldsymbol{D}(f_q) \tag{46}$$

As both $\boldsymbol{P}^*(f_q)$ *and* $\boldsymbol{D}(f_q)$ *are continuous,* $K(f_q)$ *is also continuous. If we choose* $q = 0$*,* $\boldsymbol{K}(f_q) = \alpha_0 < \alpha$ *and* $q = 1$*,* $K(f_q) = \alpha^0 < \alpha$*. Therefore, we can find a* $q \in (0,1)$*, such that* $K(f_q) = \alpha$*.*

Remark *Above theorem has told as even in reality we can not obtain an obtimal* λ*, the optimal policy is nice enough to be a convex combination of two simple policy. However, in the reality Assumption 3 we made in the beginning of this section might not hold. One possible situation is that the unconstrained policy which can obtain supremum of R, in the meanwhile it can also meet the constraint (*$\gamma = 0$*). The other situation is that all element of policy space unable to meet the constraint (necessary but not sufficient that* $\gamma = \infty$*). The theorem derived here is not applicable for above two situations. However, these two situations are also not interesting to us.*

III.1 A Practical Approach

In previous section, we have proved that it is possible to find a simple policy or a mixed policy to satisfy the constraint of system. However, a more interesting question is how to obtain an optimal policy when we trying to solve a constrained MDP in reality. In [2], the authors provided an apporach which can help us have a understanding about how to solve this kind of constrained MDP. First find the optimal λ, where Q-learning algorithm is used for a feasible α. The iteration algorithm is

$$\lambda_{k+1} = \lambda_k + (K(\mathbf{g}^{\lambda}) - \alpha) \tag{47}$$

where k is the interation number and value iteration is used to solve the J^{λ} and then we get optimal λ^*.

Now, following below procedure, a mixed policy g^* can be obtained.

- a) Perturb λ^* by δ, $\lambda^- = \lambda^* - \delta$ and $\lambda^+ = \lambda^* + \delta$
- b) Use value iteration algorithm find $\mathbf{g}^{\lambda^-}$ and $\mathbf{g}^{\lambda^+}$
- c) Apply policies $\mathbf{g}^{\lambda^-}$ and $\mathbf{g}^{\lambda^+}$, we can obtain $K(\mathbf{g}^{\lambda^-})$ and $K(\mathbf{g}^{\lambda^+})$. Find q, by solving $qK(\mathbf{g}^{\lambda^-}) + (1-q)K(\mathbf{g}^{\lambda^+}) = \alpha$
- d) Using new policy $\mathbf{g}^* = q\mathbf{g}^{\lambda^+} + (1-q)\mathbf{g}^{\lambda^-}$

IV Summary

In this report, we mainly studied the ideas in [1]. To solve the Markov decision process with constraint, Lagrange mutiplier has been introduced to reduce the problem to an unconstrained optimization parameterized by λ. Assume finite $\{\mathbf{S}\}$, compact $\{\mathbf{A}\}$, continuity of probabilities and an accessibility condition, these lead to the existence of an optimal policy.

The optimal policy is always stationary, either non-randomized stationary or consist of a mix of two non-randomized policies.

V Appendix

There are a few confusing statements and typos in the paper. We have either elaborated or corrected them in the above sections.

- In the statement of Theorem 2.5, inside the equation there is a typo $\mathbf{f} \in \mathbf{F}$ instead of $\mathbf{f} \in \mathbf{S}$
- Equation 2.7 and 2.8, in the sum part there should be N instead of n, as the state space is $[1, 2, \ldots N]$.
- Equation 3.13 of original paper the right hand side should be $\sup_{a \in \mathbf{A}} \lim_n f(n, a)$, instead of $sum_{a \in \mathbf{A}} \lim_n f(n, a)$
- Lemma 3.3 is not very clear. A more rigorous prove has been given as Lemma 5 in this report.
- Between equation 4.9 and 4.10, it should be $q \in [0, 1]$ instead of $\gamma \in [0, 1]$
- Between equation 4.10 and 4.11, there is a typo. The correct equation should be $\mathbf{B}^\gamma(\mathbf{f}_q) = q\mathbf{B}^\gamma(\underline{\mathbf{g}}) + (1 - q)\mathbf{B}^\gamma(\bar{\mathbf{g}})$
- Inside the line below equation 5.1, it should be $K^\lambda = \alpha$ instead of $J^\lambda = \alpha$.
- Also Theorem 2, Lemma 7 in the original proof is kind of unclear. Many proofs present here are more detailed and straightforward compared to the original proofs in the [1].

References

[1] Beutler, Frederick J. Ross, Keith W., "Optimal policies for controlled Markov chains with a constraint", Journal of Mathematical Analysis and Applications 112(1): 236-252

[2] Sun, C; Navaro, E.S; Wong, V.W.S; (2008): A Constrained MDP Based Vertical Hand-off Decision Algorithm for 4G Wireless Networks, IEEE International Conference on Communications, 19-23 May 2008, page: 2169-2174, Beijing.

[3] Martin L. Puterman, "Markov Decision Process Discrete Stochastic Dynamic Programming", John Wiley Sons Inc, 1994.