2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

A Potential-Based Method for Finite-Stage Markov Decision Processes

Qing-Shan Jia *Member, IEEE*

Abstract— **Finite-Stage Markov Decision Process (MDP) supplies a general framework for many practical problems when only the performance in a finite duration is of interest. Dynamic programming (DP) supplies a general way to find the optimal policies but is usually practically infeasible, due to the exponentially increasing policy space. Approximating the finite-stage MDP by an infinite-stage MDP reduces the search space but usually does not find the optimal stationary policy, due to the approximation error. We develop a method that finds the optimal stationary policies for the finite-stage MDP. The method is based on performance potentials, which can be estimated through sample paths and thus suits practical application.**

Index Terms— **Performance potentials, policy iteration, stationary policy, finite-stage Markov Decision Processes.**

I. Introduction

Markov decision process (MDP) and the associated dynamic programming (DP) methodology [1]–[3] provide a general framework for posing and analyzing problems of sequential decision making under uncertainty. Finite-Stage MDP provides such general framework for practical systems when only the performance in a finite time duration is of interest. There are generally two difficulties to apply DP to solve finite-stage MDP in practice: large search space and time-consuming simulation-based performance evaluation. The first one is known as *the curse of dimensionality*, because the size of the policy space increases exponentially fast when the number of stages and the size of the state space increase. The second one is associated with the complex nature of the system. With the development of human society, many man-made systems arise in our daily life. These systems not only follow physical laws but also follow man-made rules. Manufacturing systems, transportation systems, and communication networks are all such examples. These systems are called discrete event dynamic systems (DEDS). The only way to describe the detailed dynamic of DEDS is through simulation, which is time-consuming in general. This means the transition probability among states that are used in DP can only be estimated through simulation. These two challenges have attracted the interest of many researchers.

As a compromise to the first difficulty, we can focus on stationary policies only, which dramatically reduce the search space. Then an infinite-stage MDP could be used to approximate the finite-stage MDP, and then we can use policy iteration and value iteration to find the optimal stationary policies w.r.t. the infinite-stage criterion. We will refer to this as *the approximation approach*. However, this approach only solves the finite-stage MDP approximately. Because the finite-stage performance criterion is different from the infinite-stage performance criterion, the stationary policy thus obtained is usually not the optimal stationary policy w.r.t. the original finite-stage MDP. In this paper, we develop a method to obtain the optimal stationary policy for the finite-stage MDP. Because we focus on stationary policies only, this method requires much less memory space than DP. And because we do not change the performance criterion, the resulting stationary policy is better than the approximation approach. A detailed comparison is presented in Section V.

To deal with the second difficulty, approaches that combine the estimation of transition probabilities among states and policy iteration (or value iteration) are developed. The well-known results include the reinforcement learning [4], [5], the neuro-dynamic programming [6], the gradient-based policy iteration [7]–[11], the adaptive dynamic programming [12], the evolutionary policy iteration and the model reference adaptive search [13], just to name a few. Among these efforts, the potential-based policy iteration [14] can be applied on-line and estimates the potentials through sample paths. This is especially important when only the sample paths are available. However, the potential-based policy iteration is developed only for infinite-stage MDP. So far as we know, there are little study on how to extend the potential-based approach to finite-stage MDP. In this paper, we define performance potentials for finite-stage MDP, and establish the connection to potentials in infinite-stage MDP. The method we developed thus preserves the advantage of potential-based approaches, i.e., can be applied when only the simulation model is available.

The rest of this paper is organized as follows. First, we present the mathematical problem formulation of finite-stage MDP in Section II. We consider two objective functions that are frequently used in finite-stage MDP: The expected discounted reward and the expected total reward. Then in Section III and IV, we develop the potential-based policy iteration to optimize the expected discounted reward and the expected total reward, respectively. The method finds the globally optimal stationary policy when the actions at different states can be chosen independently. When actions at different states are chosen correlatively, we develop the gradient-based policy iteration. We discuss the relationship among the method developed in this paper, DP, and

This work was supported in part by NSFC Grant (Nos. 60704008 and 60736027), in part by the National New Faculty Funding for Universities with Doctoral Program (20070003110), and in part by the Programme of Introducing Talents of Discipline to Universities (the National 111 International Collaboration Project).

Qing-Shan Jia is with the Center for Intelligent and Networked Systems (CFINS), Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, P.R. China (phone: +86-10-62773006; fax: +86-10-62796115; e-mail: jiaqs@tsinghua.edu.cn).

the method that approximates the finite-stage MDP by an infinite-stage MDP in Section V. We briefly conclude in Section VI.

II. Problem Formulation

We consider an MDP [1]–[3] with finite state space $\mathcal{S} = \{1, 2, \dots M\}$, finite action space $\mathcal{A}$, and finite stage number N. Let X_n be the state of the system at stage n. If the system is at state i, an action can be chosen from the feasible action set $\mathcal{A}_i \subseteq \mathcal{A}$ and applied to the system. This action determines the transition probability matrix P. The system receives an immediate reward $f(i, \beta)$, when it is at state i with action β. A deterministic and stationary policy is a mapping from $\mathcal{S}$ to $\mathcal{A}$, denoted as $\mathcal{L} : \beta = \mathcal{L}(i)$. To simplify the discussion, we assume the system is ergodic under all stationary policies. Note that the ergodicity here refers to the long-term behavior of the system when a stationary policy is used, although we are interested in the finite-stage performance of this system in this paper. The expected discounted reward under policy $\mathcal{L}$ is defined as $\eta^{\mathcal{L}}_{N\alpha} = (\eta^{\mathcal{L}}_{N\alpha}(1), \eta^{\mathcal{L}}_{N\alpha}(2), \dots \eta^{\mathcal{L}}_{N\alpha}(M))^\tau$, where

$$\eta^{\mathcal{L}}_{N\alpha}(i) = (1-\alpha)\, E\left\{ \sum_{n=0}^{N-1} \alpha^n f\left(X_n, \mathcal{L}(X_n)\right) | X_0 = i \right\}, \tag{1}$$

α is a discount factor, $0 < \alpha < 1$, and τ represents the transpose. The discount factor describes that the performance at early stages are of more interest than the performance at late stages. α has practical meaning, say the interest rate, especially when the performance is economy related. The goal of the MDP is to find a stationary policy $\mathcal{L}^i$ such that its performance in Equation (1) is the minimum[1] among all stationary policies, when the initial state is i. Note that when the initial state is different, the corresponding optimal stationary policy may also change, i.e., $\mathcal{L}^i \neq \mathcal{L}^j$, for $i \neq j$. We hold an assumption as in the standard MDP formulation that the action can be chosen independently at each state. This will be called the *independent action assumption* in the following discussion. This ensures that by focusing on the deterministic stationary policies we can find the aforementioned optimal stationary policy. When this assumption does not hold, the actions at different states may be correlated, and then the optimal stationary policy may be obtained only at random stationary policies.

Another frequently used criterion in finite-stage MDP is the expected total reward criterion. The performance measure is $\eta_{NT} = (\eta_{NT}(1), \eta_{NT}(2), \dots \eta_{NT}(M))^\tau$, where $\eta_{NT}(i) = E\left\{ \sum_{n=0}^{N-1} f(X_n) | X_0 = i \right\}, i = 1, 2, \dots M.$

III. Expected Discounted Reward Criterion

In the following, we first review the definition of performance potentials g_α in infinite-stage MDP. The system performance measure in the infinite-stage MDP (i.e., the discounted performance criterion[2]) can be written as a function of performance potentials [15]. We then develop a relationship between the expected discounted reward criterion $\eta_{N\alpha}$ in the finite-stage MDP and the discounted performance criterion in the infinite-stage MDP. Then we write $\eta_{N\alpha}$ as a function of g_α. After that we can generalize the potential-based policy iteration in the infinite-stage MDP to the finite-stage MDP straightforwardly.

To simplify the notation, let $f^{\mathcal{L}}(i) = f(i, \mathcal{L}(i)), f^{\mathcal{L}} = (f^{\mathcal{L}}(1), \dots f^{\mathcal{L}}(M))^\tau$. When there is no confusion, we omit the superscript $*^{\mathcal{L}}$ that denotes the quantities associated with policy $\mathcal{L}$. For example $f^{\mathcal{L}}(i)$ may be written as $f(i)$ in some places. The discounted Poisson equation is defined as

$$(I - \alpha P + \alpha e\pi)\, g_\alpha = f,$$

where I is a unit matrix, $e = (1, \dots 1)^\tau$ is an M-by-1 column vector, $\pi = (\pi(1), \dots \pi(M))$ is the row vector of the steady state probabilities, and g_α is called the α-potential (discounted performance potential) [15], [16]. When $\alpha = 1$, it is the standard Poisson equation, its solution is simply called the potential, which is the same as the relative cost in [1] or the bias in [3]. Since we assume the Markov chain is ergodic under all stationary policies, $I - \alpha P + \alpha e\pi$ is invertible [15], and we have

$$g_\alpha = (I - \alpha P + \alpha e\pi)^{-1} f. \tag{2}$$

In particular, when $\alpha = 1$, the matrix $(I - P + e\pi)^{-1}$ is called the fundamental matrix in [17]. Since the constant part of all the potentials can be ignored when $0 < \alpha < 1$, it is also shown that

$$g_\alpha(i) = \lim_{N\to\infty} E\left\{ \sum_{n=0}^{N-1} \alpha^n f(X_n) | X_0 = i \right\}$$

[18]–[20].

The performance measure in the infinite-stage MDP is the discounted performance criterion, $\eta_\alpha = (\eta_\alpha(1), \dots \eta_\alpha(M))^\tau$, where

$$\eta_\alpha(i) = (1-\alpha)\, E\left\{ \sum_{n=0}^{\infty} \alpha^n f(X_n) | X_0 = i \right\}.$$

The performance measure can be written as a function of the α-potentials [15], [20] as follows

$$\eta_\alpha = (1-\alpha)\, g_\alpha + \alpha\eta e, \tag{3}$$

where $\eta = \pi f$ is the average-cost performance in the infinite-stage MDP. This is the foundation of the potential-based policy iteration in the infinite-stage MDP. We find that there is a strong relationship between $\eta_{N\alpha}$ and η_α as shown in Lemma 1.

Lemma 1: $\eta_{N\alpha} = \left(I - \alpha^N P^N\right) \eta_\alpha$.

[1]Without loss of generality, we consider minimization problems in this paper. The discussion can be easily applied to maximization problems.

[2]The average performance criterion corresponds to the case with the discount factor α equals to 1 [15].

Proof: By definition,

$$\begin{aligned}
&\eta_\alpha(i)\\
&=(1-\alpha)\,E\left\{\sum_{n=0}^{\infty}\alpha^n f(X_n)|X_0=i\right\}\\
&=(1-\alpha)\,E\left\{\sum_{n=0}^{N-1}\alpha^n f(X_n)+\sum_{n=N}^{\infty}\alpha^n f(X_n)|X_0=i\right\}\\
&=\eta_{N\alpha}(i)\\
&+(1-\alpha)\sum_{j=1}^{M}E\left\{\sum_{n=N}^{\infty}\alpha^n f(X_n)|X_0=i,X_N=j\right\}\\
&\times P(X_N=j|X_0=i)\\
&=\eta_{N\alpha}(i)+\sum_{j=1}^{M}\alpha^N\eta_\alpha(j)\,P^N(i,j). \qquad (4)
\end{aligned}$$

The first line in Equation (4) is by definition. The second line follows from the property of expectation. In matrix form, we have $\eta_\alpha=\eta_{N\alpha}+\alpha^N P^N\eta_\alpha$, so $\eta_{N\alpha}=\left(I-\alpha^N P^N\right)\eta_\alpha$. ■

Using Lemma 1 and Equation (3), we can write $\eta_{N\alpha}$ as a function of g_α,

$$\eta_{N\alpha}=\left(I-\alpha^N P^N\right)\left((1-\alpha)\,g_\alpha+\alpha\eta e\right). \qquad (5)$$

Equation (5) establishes the relationship between the discounted performance criterion in the finite-stage MDP and the α-potentials in the infinite-stage MDP. The two MDPs adopt a same stationary policy.

We now analyze how the change from policy $\mathcal{L}$ to $\mathcal{L}'$ affects the system performances. Following Equation (5) we have

$$\begin{aligned}
\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}=&\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)\left((1-\alpha)\,g_\alpha^{\mathcal{L}'}+\alpha\eta^{\mathcal{L}'}e\right)\\
&-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)\left((1-\alpha)\,g_\alpha^{\mathcal{L}}+\alpha\eta^{\mathcal{L}}e\right). \qquad (6)
\end{aligned}$$

Suppose $\mathcal{L}$ is the currently adopted policy, and $\mathcal{L}'$ can be any other policy candidate. Note that both potentials $g_\alpha^{\mathcal{L}}$ and $g_\alpha^{\mathcal{L}'}$ appear in Equation (6). Since in practice we usually have thousands of hundreds of policy candidates $\mathcal{L}'$, it is practically infeasible to calculate the potential $g_\alpha^{\mathcal{L}'}$ for each $\mathcal{L}'$. So, we develop the following Lemma 2 to describe the difference $\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}$ as a function of only $g_\alpha^{\mathcal{L}}$, the potentials of the currently adopted policy $\mathcal{L}$.

Lemma 2:

$$\begin{aligned}
&\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}=\\
&\left(I-\alpha P^{\mathcal{L}'}\right)^{-1}(1-\alpha)\times\left(\alpha Q\left(I-\alpha^N(P^{\mathcal{L}})^N\right)g_\alpha^{\mathcal{L}}\right.\\
&\left.+\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)f^{\mathcal{L}'}-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)f^{\mathcal{L}}\right), \qquad (7)
\end{aligned}$$

where $Q=P^{\mathcal{L}'}-P^{\mathcal{L}}$.

Proof: $\forall 0<\alpha<1$, for $\mathcal{L}$ and $\mathcal{L}'$,

$$\begin{aligned}
\eta_\alpha&=(1-\alpha)\sum_{n=0}^{\infty}\alpha^n P^n f=(1-\alpha)\left(I+\sum_{n=1}^{\infty}\alpha^n P^n\right)f\\
&=(1-\alpha)\,f+\alpha P\eta_\alpha. \qquad (8)
\end{aligned}$$

By Lemma 1, Equation (8), and the definition of $\eta_{N\alpha}^{\mathcal{L}}$ and $\eta_{N\alpha}^{\mathcal{L}'}$, we have

$$\begin{aligned}
&\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}\\
&=\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)\eta_\alpha^{\mathcal{L}'}-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)\eta_\alpha^{\mathcal{L}}\\
&=\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)\left((1-\alpha)\,f^{\mathcal{L}'}+\alpha P^{\mathcal{L}'}\eta_\alpha^{\mathcal{L}'}\right)\\
&-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)\left((1-\alpha)\,f^{\mathcal{L}}+\alpha P^{\mathcal{L}}\eta_\alpha^{\mathcal{L}}\right)\\
&=\alpha\left(P^{\mathcal{L}'}\left(\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}\right)+\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\eta_{N\alpha}^{\mathcal{L}}\right)\\
&+(1-\alpha)\left(\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)f^{\mathcal{L}'}-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)f^{\mathcal{L}}\right), \qquad (9)
\end{aligned}$$

where the first equality follows from Lemma 1, the second equality follows from Equation (8), and the last equality follows from the definition of $\eta_{N\alpha}^{\mathcal{L}}$ and $\eta_{N\alpha}^{\mathcal{L}'}$. Combining the term $\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}$ in the right hand side of Equation (8) to the left hand side and rearranging the equation, we have

$$\begin{aligned}
&\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}\\
&=\left(I-\alpha P^{\mathcal{L}'}\right)^{-1}\left(\alpha\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\eta_{N\alpha}^{\mathcal{L}}\right.\\
&\left.+(1-\alpha)\left(\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)f^{\mathcal{L}'}-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)f^{\mathcal{L}}\right)\right). \qquad (10)
\end{aligned}$$

We also have

$$\begin{aligned}
&\alpha\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\eta_{N\alpha}^{\mathcal{L}}\\
&=\alpha\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\left(I-\alpha^N(P^{\mathcal{L}})^N\right)\eta_\alpha^{\mathcal{L}}\\
&=\alpha\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\left(I-\alpha^N(P^{\mathcal{L}})^N\right)\left((1-\alpha)\,g_\alpha^{\mathcal{L}}+\alpha\eta^{\mathcal{L}}e\right)\\
&=\alpha\left(P^{\mathcal{L}'}-P^{\mathcal{L}}\right)\left(I-\alpha^N(P^{\mathcal{L}})^N\right)(1-\alpha)\,g_\alpha^{\mathcal{L}}, \qquad (11)
\end{aligned}$$

where the first equality follows from the definition of $\eta_{N\alpha}^{\mathcal{L}}$, the second equality follows from Equation (3), and the last equality follows from the fact that

$$(I-\alpha^N(P^{\mathcal{L}})^N)\alpha\eta^{\mathcal{L}}e=\alpha\eta^{\mathcal{L}}(e-\alpha^N e)=\alpha\eta^{\mathcal{L}}(1-\alpha^N)e,$$

and

$$(P^{\mathcal{L}'}-P^{\mathcal{L}})e=e-e=0,$$

noting that $P^{\mathcal{L}}e=e, P^{\mathcal{L}'}e=e$, and $\eta^{\mathcal{L}}$ is a scalar.

Following Equations (12) and (11), we have

$$\begin{aligned}
&\eta_{N\alpha}^{\mathcal{L}'}-\eta_{N\alpha}^{\mathcal{L}}\\
&=\left(I-\alpha P^{\mathcal{L}'}\right)^{-1}(1-\alpha)\left(\alpha Q\left(I-\alpha^N(P^{\mathcal{L}})^N\right)g_\alpha^{\mathcal{L}}\right.\\
&\left.+\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)f^{\mathcal{L}'}-\left(I-\alpha^N(P^{\mathcal{L}})^N\right)f^{\mathcal{L}}\right). \qquad (12)
\end{aligned}$$

■

Using Lemma 2, we can develop a necessary and sufficient condition for the globally optimal stationary policy for the finite-stage MDP.

Theorem 1: In a finite-stage MDP with expected discounted reward performance criterion, a policy $\mathcal{L}^i$ is the

optimal stationary policy for initial state i if and only if

$$\left\{\alpha P^{\mathcal{L}'}\left(I-\alpha^N(P^{\mathcal{L}^i})^N\right)g_\alpha^{\mathcal{L}^i}+\left(I-\alpha^N(P^{\mathcal{L}'})^N\right)f^{\mathcal{L}'}\right\}_i$$
$$\geq\left\{\alpha P^{\mathcal{L}^i}\left(I-\alpha^N(P^{\mathcal{L}^i})^N\right)g_\alpha^{\mathcal{L}^i}+\left(I-\alpha^N(P^{\mathcal{L}^i})^N\right)f^{\mathcal{L}^i}\right\}_i$$

for all $\mathcal{L}'\in\mathcal{E}$, where $\mathcal{E}:\mathcal{S}\to\mathcal{A}$ is the set of all stationary policies, and $\{a\}_i$ is the i-th element of vector a.

Proof: For $0<\alpha<1$,

$$\lim_{N\to\infty}\left(\alpha P^{\mathcal{L}'}\right)^N=0.$$

Hence we have [17]

$$\left(I-\alpha P^{\mathcal{L}'}\right)^{-1}=I+\alpha P^{\mathcal{L}'}+\alpha^2(P^{\mathcal{L}'})^2+\cdots. \quad (13)$$

Since the Markov Chain is ergodic and the states are finite, it is positive recurrent [21] and every item in $\left(I-\alpha P^{\mathcal{L}'}\right)^{-1}$ is positive. Then following from Lemma 2, if the condition in Theorem 1 is met, no other stationary policy can decrease $\eta_{N\alpha}^{\mathcal{L}^i}(i)$. Thus the current policy $\mathcal{L}^i$ is the optimal one. This proves the sufficient part. If the policy $\mathcal{L}^i$ is the optimal one for initial state i, no other policy can decrease $\eta_{N\alpha}^{\mathcal{L}^i}(i)$. This proves the necessary part. Thus this is a necessary and sufficient condition for globally optimal stationary policy. ■

Theorem 1 indicates a way to improve the policy and tells when to stop the improvement. We now present the *potential-based policy iteration* for finite-stage MDP. The iteration process is as follows. Let $\mathcal{L}_k^i$ be the policy at the kth iteration. In the $(k+1)$th iteration, we select

$$\mathcal{L}_{k+1}^i=\arg\min_{\mathcal{L}}\left\{\alpha P^{\mathcal{L}}\left(I-\alpha^N(P^{\mathcal{L}})^N\right)g_\alpha^{\mathcal{L}_k^i}+\left(I-\alpha^N(P^{\mathcal{L}})^N\right)f^{\mathcal{L}}\right\}_i. \quad (14)$$

Since $\eta_{N\alpha}(i)$ is decreased at each iteration and both the state and the action sets are finite, the optimal stationary policy can be reached in finite iterations.

Note that the potential-based policy iteration in finite-stage MDP finds the optimal stationary policy w.r.t. $\eta_{N\alpha}$. The potential-based policy iteration in infinite-stage MDP finds the optimal stationary policy w.r.t. η_α. When the number of stages is infinite, the two methods find the same stationary policy, because $\lim_{N\to\infty}\eta_{N\alpha}=\eta_\alpha$. But when the number of stages is small, the potential-based policy iteration in infinite-stage MDP usually finds unsatisfactory policies, due to the difference between $\eta_{N\alpha}$ and η_α. Thus the policy iteration developed above always finds better stationary policy w.r.t. $\eta_{N\alpha}$.

In above discussion, we use the *independent action assumption*. However, in some systems this assumption does not hold. In the rest of this section, we consider the case when actions at different states may be correlated. First we develop a formula to calculate the gradient of system performance w.r.t. perturbations in policies. This formula only utilizes the information of the currently adopted policy, so it can be applied even when the actions at different states are correlated. Then we develop the gradient-based policy iteration for finite-stage MDP.

Suppose that $P^{\mathcal{L}}$ changes to $P(\delta)=P^{\mathcal{L}}+\delta Q=\delta P^{\mathcal{L}'}+(1-\delta)P^{\mathcal{L}}$, and $f^{\mathcal{L}}$ changes to $f(\delta)=f^{\mathcal{L}}+\delta h$ with $\delta\in[0,1]$, where $h=f^{\mathcal{L}'}-f^{\mathcal{L}}$. This corresponds to a randomized policy which applies policy $\mathcal{L}'$ with probability δ and policy $\mathcal{L}$ with probability $1-\delta$. The performance will change to $\eta_{N\alpha}(\delta)$. The derivative of $\eta_{N\alpha}$ in the direction of Q is denoted as $\frac{d\eta_{N\alpha}}{d\delta}$. Then we can write $\frac{d\eta_{N\alpha}}{d\delta}$ as a function of $g_\alpha^{\mathcal{L}}$.

Lemma 3: Assume $\eta_{N\alpha}^{\mathcal{L}}$ is analytical for all $\delta\in[0,1]$, then we have

$$\frac{d\eta_{N\alpha}}{d\delta}=\left(I-\alpha P^{\mathcal{L}}\right)^{-1}(1-\alpha)\left(\alpha Q\left(I-\alpha^N(P^{\mathcal{L}})^N\right)g_\alpha^{\mathcal{L}}\right.$$
$$-\alpha^N\left((P^{\mathcal{L}})^{N-1}Q+(P^{\mathcal{L}})^{N-2}QP^{\mathcal{L}}+\cdots+Q(P^{\mathcal{L}})^{N-1}\right)f^{\mathcal{L}}$$
$$\left.+\left(I-\alpha^N(P^{\mathcal{L}})^N\right)h\right).$$

Proof: Just note that

$$\frac{d\eta_{N\alpha}}{d\delta}=\lim_{\delta\to0}\frac{\eta_{N\alpha}(\delta)-\eta_{N\alpha}}{\delta}.$$

The rest follows from Equation (12). ■

Lemma 3 only requires the information of the currently adopted policy and the difference Q and h to calculate $\frac{d\eta_{N\alpha}}{d\delta}$. If we select

$$\mathcal{L}_{k+1}^i=\arg\min_{\mathcal{L}}\left\{\alpha Q\left(I-\alpha^N(P^{\mathcal{L}_k^i})^N\right)g_\alpha^{\mathcal{L}_k^i}\right.$$
$$-\alpha^N\left((P^{\mathcal{L}_k^i})^{N-1}Q+(P^{\mathcal{L}_k^i})^{N-2}QP^{\mathcal{L}_k^i}+\cdots\right.$$
$$\left.\left.+Q(P^{\mathcal{L}_k^i})^{N-1}\right)f^{\mathcal{L}_k^i}+\left(I-\alpha^N(P^{\mathcal{L}_k^i})^N\right)h\right\}, \quad (15)$$

where the minimum is taken over all feasible policies w.r.t. the correlation constraint among the actions at different states, this is the *gradient-based policy iteration* for finite-stage MDP. Unfortunately, even if $\frac{d\eta_{N\alpha}}{d\delta}<0$, we may not have $\eta_{N\alpha}^{\mathcal{L}'}<\eta_{N\alpha}^{\mathcal{L}}$, due to the difference between the gradient and the difference. Furthermore, in MDP with correlated actions, the optimal system performance may be obtained only at random stationary policies. So the stochastic approximation may be used together with gradient-based policy iteration. In infinite-stage MDP with expected discounted reward criterion, the combination of stochastic approximation and gradient estimation techniques can find the local optimal policy with probability 1 [9]. For both finite-stage MDP and infinite-stage MDP there are many open problems in this direction.

IV. Expected Total Reward Criterion

In this section, we first establish the relationship between the expected total reward criterion η_{NT} and the expected discounted reward criterion $\eta_{N\alpha}$. Then we develop formulas to describe the performance difference $\eta_{NT}^{\mathcal{L}'}-\eta_{NT}^{\mathcal{L}}$ and performance derivative $\frac{d\eta_{NT}}{d\delta}$ as a function of performance potentials $g_\alpha^{\mathcal{L}}$. Similar to Section III, under the independent action assumption, we develop the necessary and sufficient condition for the globally optimal stationary policy w.r.t. the expected total reward criterion. The potential-based policy iteration can then be developed to find the globally optimal stationary policy. For problems with correlated actions, we develop the gradient-based policy iteration.

First, we present a relationship between η_{NT} and $\eta_{N\alpha}$. Unfortunately we cannot optimize η_{NT} directly from $\eta_{N\alpha}$, as shown in Lemma 4.

Lemma 4: $\lim_{\alpha\to 1^-} \eta_{N\alpha} = 0$.

Proof: Lemma 4 is a direct result of Lemma 1 and Equation (8). Just note that [20]

$$(I-\alpha P)^{-1} = (I-\alpha P+\alpha e\pi)^{-1} + \frac{\alpha}{1-\alpha} e\pi.$$

■

To relate η_{NT} to $\eta_{N\alpha}$, we introduce $\eta_{NT\alpha} = (\eta_{NT\alpha}(1), \dots \eta_{NT\alpha}(M))^\tau$, where

$$\eta_{NT\alpha}(i) = E\left\{\sum_{n=0}^{N-1} \alpha^n f(X_n) \,|X_0 = i\right\}, i = 1, 2, \dots M.$$

By definition, $\eta_{NT\alpha} = \frac{\eta_{N\alpha}}{1-\alpha}$. And we also have

Lemma 5: $\lim_{\alpha\to 1^-} \eta_{NT\alpha} = \eta_{NT}$.

So we first describe $\eta^{\mathcal{L}'}_{NT\alpha} - \eta^{\mathcal{L}}_{NT\alpha}$ as a function of $g^{\mathcal{L}}_\alpha$, and then use Lemma 5 to describe $\eta^{\mathcal{L}'}_{NT} - \eta^{\mathcal{L}}_{NT}$ as a function of $g^{\mathcal{L}}_\alpha$.

Lemma 6:

$$\eta^{\mathcal{L}'}_{NT\alpha} - \eta^{\mathcal{L}}_{NT\alpha} = \left(I-\alpha P^{\mathcal{L}'}\right)^{-1} \left(\alpha Q\left(I-\alpha^N (P^{\mathcal{L}})^N\right) g^{\mathcal{L}}_\alpha \right.$$
$$\left. +\alpha^N \left(\left(I-\alpha^N (P^{\mathcal{L}'})^N\right) f^{\mathcal{L}'} - \left(I-\alpha^N (P^{\mathcal{L}})^N\right) f^{\mathcal{L}}\right)\right).$$

Lemma 7:

$$\eta^{\mathcal{L}'}_{NT} - \eta^{\mathcal{L}}_{NT} = \lim_{\alpha\to 1^-} \left(\eta^{\mathcal{L}'}_{NT\alpha} - \eta^{\mathcal{L}}_{NT\alpha}\right)$$
$$= -e\pi^{\mathcal{L}'} \left(Q\left(I-(N+1)(P^{\mathcal{L}})^N\right.\right.$$
$$\left.\left. -\left(I-(P^{\mathcal{L}})^N\right)\left(I-P^{\mathcal{L}}+e\pi^{\mathcal{L}}\right)^{-1}\left(e\pi^{\mathcal{L}}-P^{\mathcal{L}}\right)\right) g^{\mathcal{L}}_1$$
$$+N\left(\left(I-2(P^{\mathcal{L}'})^N\right) f^{\mathcal{L}'} - \left(I-2(P^{\mathcal{L}})^N\right) f^{\mathcal{L}}\right),$$

where $g^{\mathcal{L}}_1 = \left(I-P^{\mathcal{L}}+e\pi^{\mathcal{L}}\right)^{-1} f^{\mathcal{L}}$.

Based on Lemma 7 and the independent action assumption, we develop the necessary and sufficient condition for globally optimal stationary policy w.r.t. the expected total reward criterion.

Theorem 2: Under the independent action assumption, in finite-stage MDP with expected total reward performance criterion, a policy $\mathcal{L}^i$ is the optimal stationary policy for initial state i if and only if

$$\left\{P^{\mathcal{L}'}\left(I-(N+1)(P^{\mathcal{L}^i})^N\right.\right.$$
$$\left. -\left(I-(P^{\mathcal{L}^i})^N\right)\left(I-P^{\mathcal{L}^i}+e\pi^{\mathcal{L}^i}\right)^{-1}\left(e\pi^{\mathcal{L}^i}-P^{\mathcal{L}^i}\right)\right) g_1^{\mathcal{L}^i}$$
$$\left. +N\left(I-2(P^{\mathcal{L}'})^N\right) f^{\mathcal{L}'}\right\}_i \le$$
$$\left\{P^{\mathcal{L}^i}\left(I-(N+1)(P^{\mathcal{L}^i})^N\right.\right.$$
$$\left. -\left(I-(P^{\mathcal{L}^i})^N\right)\left(I-P^{\mathcal{L}^i}+e\pi^{\mathcal{L}^i}\right)^{-1}\left(e\pi^{\mathcal{L}^i}-P^{\mathcal{L}^i}\right)\right) g_1^{\mathcal{L}^i}$$
$$\left. +N\left(I-2(P^{\mathcal{L}^i})^N\right) f^{\mathcal{L}^i}\right\}_i,$$

for all $\mathcal{L}' \in \mathcal{E}$.

TABLE I

THE COMPARISON OF COMPUTATIONAL COMPLEXITY PER ITERATION FOR IS AND FS AND TOTAL FOR DP.

	DP	IS		FS
		PI	VI	
time	$O(N\lvert\mathcal{A}\rvert\lvert\mathcal{S}\rvert^2)$	$O(\lvert\mathcal{S}\rvert^3)$	$O(\lvert\mathcal{A}\rvert\lvert\mathcal{S}\rvert^2)$	$O(\lvert\mathcal{S}\rvert^3)$
space	$O(N\lvert\mathcal{S}\rvert)$	$O(\lvert\mathcal{S}\rvert)$	$O(\lvert\mathcal{S}\rvert)$	$O(\lvert\mathcal{S}\rvert)$
$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{A}\rvert^{N\lvert\mathcal{S}\rvert}$	$\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}$	$\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}$	$\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}$

The proof of Theorem 2 is similar to the proof of Theorem 1. Theorem 2 indicates a way to do *potential-based policy iteration*: In the $(k+1)$th iteration, we select

$$\mathcal{L}^i_{k+1} = \arg\max_{\mathcal{L}} \left\{P^{\mathcal{L}}\left(I-(N+1)(P^{\mathcal{L}^i_k})^N\right.\right.$$
$$\left. -\left(I-(P^{\mathcal{L}^i_k})^N\right)\left(I-P^{\mathcal{L}^i_k}+e\pi^{\mathcal{L}^i_k}\right)^{-1}$$
$$\left.\times\left(e\pi^{\mathcal{L}^i_k}-P^{\mathcal{L}^i_k}\right)\right) g_1^{\mathcal{L}^i_k} + N\left(I-2(P^{\mathcal{L}})^N\right) f^{\mathcal{L}}\Big\}, \tag{16}$$

where the maximum is taken componentwisely. When actions at different states are correlated, we have

Lemma 8: Assume $\eta_{NT\alpha}$ is analytical for all $\delta \in [0,1]$, then we have

$$\frac{d\eta_{NT\alpha}}{d\delta} = \left(I-\alpha P^{\mathcal{L}}\right)^{-1}\left(\alpha Q\left(I-\alpha^N (P^{\mathcal{L}})^N\right) g^{\mathcal{L}}_\alpha\right.$$
$$-\alpha^N\left((P^{\mathcal{L}})^{N-1}Q + (P^{\mathcal{L}})^{N-2}QP^{\mathcal{L}} + \dots\right.$$
$$\left.\left.+Q(P^{\mathcal{L}})^{N-1}\right) f^{\mathcal{L}} + \left(I-\alpha^N (P^{\mathcal{L}})^N\right) h\right).$$

Lemma 9: Assume $\eta_{NT\alpha}$ is analytical for all $\delta \in [0,1]$, then we have

$$\frac{d\eta_{NT}}{d\delta} = -e\pi^{\mathcal{L}}\left(Q\left(I-(N+1)(P^{\mathcal{L}})^N - \left(I-(P^{\mathcal{L}})^N\right)\right.\right.$$
$$\left.\times\left(I-P^{\mathcal{L}}+e\pi^{\mathcal{L}}\right)^{-1}\left(e\pi^{\mathcal{L}}-P^{\mathcal{L}}\right)\right) g^{\mathcal{L}}_1$$
$$-N\left((P^{\mathcal{L}})^{N-1}Q + \dots + Q(P^{\mathcal{L}})^{N-1}\right) f^{\mathcal{L}}$$
$$\left.-N\alpha^{N-1}(P^{\mathcal{L}})^N h\right).$$

Note that Lemma 9 only requires the information of the currently adopted policy and the difference Q and h to calculate $\frac{d\eta_{NT}}{d\delta}$. If we select

$$\mathcal{L}_{k+1} = \arg\max_{\mathcal{L}} \left\{P^{\mathcal{L}}\left(I-(N+1)(P^{\mathcal{L}_k})^N - \left(I-(P^{\mathcal{L}_k})^N\right)\right.\right.$$
$$\left.\times\left(I-P^{\mathcal{L}_k}+e\pi^{\mathcal{L}_k}\right)^{-1}\left(e\pi^{\mathcal{L}_k}-P^{\mathcal{L}_k}\right)\right) g_1^{\mathcal{L}_k}$$
$$-N\left((P^{\mathcal{L}_k})^{N-1}Q + \dots + Q(P^{\mathcal{L}_k})^{N-1}\right) f^{\mathcal{L}_k}$$
$$\left.-N\alpha^{N-1}(P^{\mathcal{L}_k})^N h\right\}, \tag{17}$$

where the maximum is also taken componentwisely. This is the *gradient-based policy iteration* for expected total reward criterion in finite-stage MDP.

V. DISCUSSIONS

As mentioned in Section I, there are usually three methods for finite-stage MDP, namely DP, approximating the finite-stage MDP by an infinite-stage MDP (IS for short), and the method developed in this paper (FS for short). We first compare the computational complexity of the three methods, as shown in Table I.

In Table I, PI and VI are short for policy iteration and value iteration, which are well-known methods for infinite-stage MDP. DP explores all the stationary and non-stationary policies. All the other methods only explore the stationary policies. The computational complexity of DP, PI, and VI are well-known results [3]. For the method developed in this paper, we use Eq. (14) as an example. Suppose $g_\alpha^{\mathcal{L}_k^i}$ is obtained by solving the Poisson equation using matrix inversion. Note that the minimization in Eq. (14) is taken over all policies. In practice, we can replace this by m randomly sampled policies, where m is a predetermined value. A small m may lead to a large number of iterations before the iterations stops. Then Eq. (14) has a time complexity of $O(|\mathcal{S}|^3)$ and space complexity of $O(|\mathcal{S}|)$.[3] So in each iteration, FS has a time complexity similar to PI. Since it is difficult to quantify the number of iterations thus needed in PI, VI, and FS, we do not know which method has the minimal time complexity in total. The space complexity of PI, VI, and FS is independent from N, and thus much smaller than DP when N is large.

We then compare the performance of the resulting policies of these methods. DP finds the optimal policy w.r.t. $\eta_{N\alpha}$. PI and VI find the optimal stationary policy w.r.t. η_α. And FS finds the optimal stationary policy w.r.t. $\eta_{N\alpha}$. So, we have $\eta_{N\alpha}^{\mathcal{L}_{\text{DP}}} \leq \eta_{N\alpha}^{\mathcal{L}_{\text{FS}}} \leq \eta_{N\alpha}^{\mathcal{L}_{\text{IS}}}$. So, the method developed in this paper has a computational complexity similar to policy iteration but finds better stationary policies for finite-stage MDP. It should be considered when the computational complexity (say the space complexity) of DP is not affordable in an application.

VI. Conclusion

In this paper, we generalize the potential-based method [15] to find the optimal stationary policies for the finite-stage MDP. By focusing on stationary policies, the method can be applied when DP is practically infeasible due to large search space and simulation-based performance evaluation. Two most frequently used criteria are considered: The expected discounted reward and the expected total reward. Under the independent action assumption the necessary and sufficient condition (Theorem 1 and 2) for globally optimal stationary policy is developed, together with the potential-based policy iteration (Equation (14) and (16)). The proposed method always finds better policies than directly applying the potential-based policy iteration for infinite-stage MDP [15], since the impact of the finite stage objective functions is considered in our method. When the actions at different states are correlated, we develop the gradient-based policy iteration (Equation (15) and (17)). Note that since the performance potential g_α used in this paper is exactly the same as the potential defined in infinite-stage MDP [15], all the algorithms to estimate potentials in infinite-stage MDP through sample path [22] can be applied here.

As the further research directions, we can try to extend the method developed in this paper to Borel case (i.e., the state space and the action space are Borel spaces) [23]. Another direction is to develop the potential-based policy iteration to find the optimal non-stationary policy in finite-stage MDP.

[3]Note that this is for calculating the optimal stationary policy of one initial state. If we want to calculate the optimal stationary policies for each initial state, the time and space complexity of Eq. (14) should be $O(|\mathcal{S}|^4)$ and $O(|\mathcal{S}|^2)$, respectively.

Acknowledgment

The author would like to thank Prof. X. R. Cao, Prof. Y. L. Lin, Prof. Q. C. Zhao, Mr. Y. Cao, and nine anonymous referees for the constructive comments on early versions of this manuscript. Of course, all remaining errors are solely the responsibility of the author.

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

[2] E. A. Feinberg and A. Shwartz, *Handbook of Markov Decision Processes Methods and Applications*. Boston, MA: Kluwer Academic Publishers, 2002.

[3] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley and Sons, Inc., 1994.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[5] A. Barto and S. Mahadevan, "Recent advances in hirerarchical reinforcement learning," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 13, pp. 41–77, 2003, special Issue on Reinforcement Learning.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[7] Y. C. Ho and X. R. Cao, *Perturbation Analysis of Discrete Event Systems*. Boston, MA: Kluwer Academic Publisher, 1991.

[8] E. K. P. Chong and P. J. Ramadages, "Stochastic optimization of regenerative systems using infinitesimal perturbation analysis," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1400–1410, July 1994.

[9] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of markov reward processes," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 191–209, 2001.

[10] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[11] X. R. Cao, "A basic formula for online policy gradient algorithms," *IEEE Trans. Automat. Contr.*, vol. 50, no. 5, pp. 696–699, May 2005.

[12] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York, NY: Wiley-Interscience, 2007.

[13] H. S. Chang, M. C. Fu, J. Hu, and S. I. Marcus, *Simulation-Based Algorithms for Markov Decision Processes*. New York, NY: Springer, 2007.

[14] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. New York, NY: Springer, 2007.

[15] ——, "From perturbation analysis to markov decision processes and reinforcement learning," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 13, no. 1-2, pp. 9–39, Jan.-Apr. 2003.

[16] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1975.

[17] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. Princeton, NJ: D. Van Nostrand Com. Inc., 1960.

[18] X. R. Cao and H. F. Chen, "Perturbation realization, potentials, and sensitivity analysis of markov processes," *IEEE Trans. Automat. Contr.*, vol. 42, no. 10, Oct. 1997.

[19] X. R. Cao, "The relations among potentials, perturbation analysis, and markov decision processes," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 8, pp. 71–87, 1998.

[20] ——, "A unified approach to markov decision problems and performance sensitivity analysis," *Automatica*, vol. 36, pp. 771–774, 2000.

[21] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. New York, NY: Academic Press, Inc., 1975.

[22] X. R. Cao and Y. W. Wan, "Algorithms for sensitivity analysis of markov systems through potentials and perturbation realization," *IEEE Trans. Contr. Syst. Technol.*, vol. 6, no. 4, July 1998.

[23] O. Hernandez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. New York, NY: Springer, 1996.