Published in final edited form as:
J Mach Learn Res. 2005 July ; 6: 1073–1097.

A Generalization Error for Q-Learning

Susan A. Murphy
Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, USA

Abstract

Planning problems that involve learning a policy from a single training set of finite horizon trajectories arise in both social science and medical fields. We consider Q-learning with function approximation for this setting and derive an upper bound on the generalization error. This upper bound is in terms of quantities minimized by a Q-learning algorithm, the complexity of the approximation space and an approximation term due to the mismatch between Q-learning and the goal of learning a policy that maximizes the value function.

Keywords

multistage decisions; dynamic programming; reinforcement learning; batch data

1. Introduction

In many areas of the medical and social sciences the following planning problem arises. A training set or batch of n trajectories of T +1-decision epochs is available for estimating a policy. A decision epoch at time t, $t = 0,1,...,T$, is composed of information observed at time t, O_t , an action taken at time t, A_t and a reward, R_t . For example there are currently a number of ongoing large clinical trials for chronic disorders in which, each time an individual relapses, the individual is re-randomized to one of several further treatments (Schneider et al., 2001; Fava et al., 2003; Thall et al., 2000). These are finite horizon problems with T generally quite small, $T = 2$–4, with known exploration policy. Scientists want to estimate the best “strategies,” i.e. policies, for managing the disorder. Alternately the training set of n trajectories may be historical; for example data in which clinicians and their patients are followed with i! nformation about disease process, treatment burden and treatment decisions recorded through time. Again the goal is to estimate the best policy for managing the disease. Alternately, consider either catalog merchandizing or charitable solicitation; information about the client, and whether or not a solicitation is made and/or the form of the solicitation is recorded through time (Simester, Peng and Tsitsiklis, 2003). The goal is to estimate the best policy for deciding which clients should receive a mailing and the form of the mailing. These latter planning problems can be viewed as infinite horizon problems but only T decision epochs per client are recorded. If T is large, the rewards are bounded and the dynamics are stationary Markovian then this finite horizon problem provides an approximation to the discounted infinite horizon problem (Kearns, Mansour and Ng, 2000).

These planning problems are characterized by unknown system dynamics and thus can also be viewed as learning problems as well. Note there is no access to a generative model nor an online simulation model nor the ability to conduct offline simulation. All that is available is the n trajectories of $T + 1$ decision epochs. One approach to learning a policy in this setting is Q-learning (Watkins, 1989) since the actions in the training set are chosen according to a (non-

SAMURPHY@UMICH.EDU.
Editor: Michael Littman

optimal) exploration policy; Q-learning is an off-policy method (Sutton and Barto, 1998). When the observables are vectors of continuous variables or are otherwise of high dimension, Q-learning must be combined with function approximation.

The contributions of this paper are as follows. First a version of Q-learning with function approximation, suitable for learning a policy with one training set of finite horizon trajectories and a large observation space, is introduced; this "batch" version of Q-learning processes the entire training set of trajectories prior to updating the approximations to the Q-functions. An incremental implementation of batch Q-learning results in one-step Q-learning with function approximation. Second performance guarantees for this version of Q-learning are provided. These performance guarantees do not assume assume that the system dynamics are Markovian. The performance guarantees are upper bounds on the average difference in value functions or more specifically the average generalization error. Here the generalization error for batch Q-learning is defined analogous to the generalization error in supervised learning (Schapire et al., 1998); it is the average diffe! rence in value when using the optimal policy as compared to using the greedy policy (from Q-learning) in generating a separate test set. The performance guarantees are analogous to performance guarantees available in supervised learning (Anthony and Bartlett, 1999).

The upper bounds on the average generalization error permit an additional contribution. These upper bounds illuminate the mismatch between Q-learning with function approximation and the goal of finding a policy maximizing the value function (see the remark following Lemma 2 and the third remark following Theorem 2). This mismatch occurs because the Q-learning algorithm with function approximation does not directly maximize the value function but rather this algorithm approximates the optimal Q-function within the constraints of the approximation space in a least squares sense; this point is discussed as some length in section 3 of Tsitsiklis and van Roy (1997).

In the process of providing an upper bound on the average generalization error, finite sample bounds on the difference in average values resulting from different policies are derived. There are three terms in the upper bounds. The first term is a function of the optimization criterion used in batch Q-learning, the second term is due to the complexity of the approximation space and the last term is an approximation error due to the above mentioned mismatch. The third term which is a function of the complexity of the approximation space is similar in form to generalization error bounds derived for supervised learning with neural networks as in Anthony and Bartlett (1999). From the work of Kearns, Mansour, and Ng (1999, 2000) and Peshkin and Shelton (2002), we expect and find as well here that the number of trajectories needed to guarantee a specified error level is exponential in the horizon time, T . The upper bound does not depend on the dimension of the observables O_t 's. This is in contrast to the results of Fiechter (1994Fiechter (1997) in which the upper bound on the average generalization error depends on the number of possible values for the observables.

A further contribution is that the upper bound on the average generalization error provides a mechanism for generalizing ideas from supervised learning to reinforcement learning. For example if the optimal Q-function belongs to the approximation space, then the upper bounds imply that batch Q-learning is a PAC reinforcement learning algorithm as in Feichter (1994 in Feichter (1997); see the first remark following Theorem 1. And second the upper bounds provide a starting point in using structural risk minimization for model selection (see the second remark after Theorem 1).

In Section 2, we review the definition of the value function and Q-function for a (possibly non-stationary, non-Markovian) finite horizon decision process. Next we review batch Q-learning with function approximation when the learning algorithm must use a training set of n

trajectories. In Section 5 we provide the two main results, both of which provide the number of trajectories needed to achieve a given error level with a specified level of certainty.

2. Preliminaries

In the following we use upper case letters, such as O and A, to denote random variables and lower case letters, such as o and a, to denote instantiates or values of the random variables. Each of the n trajectories is composed of the sequence $\{O_0, A_0, O_1, \ldots, A_T, O_{T+1}\}$ where T is a finite constant. Define $\mathbf{O}_t = \{O_0, \ldots, O_t\}$ and similarly for $\mathbf{A}_t$. Each action A_t takes values in finite, discrete action space A and O_t takes values in the observation space O. The observation space may be multidimensional and continuous. The arguments below will not require the Markovian assumption with the value of O_t equal to the state at time t. The rewards are $R_t = r_t(\mathbf{O}_t, \mathbf{A}_t, O_{t+1})$ for r_t a reward function and for each $0 \le t \le T$ (if the Markov assumption holds then replace $\mathbf{O}_t$ with O_t and $\mathbf{A}_t$ with A_t). We assume that the rewards are bounded, taking values in the interval [0,1].

We assume the trajectories are sampled at random according to a fixed distribution denoted by P. Thus the trajectories are generated by one fixed distribution. This distribution is composed of the unknown distribution of each O_t conditional on $(\mathbf{O}_{t-1}, \mathbf{A}_{t-1})$ (call these unknown conditional densities $\{f_0, \ldots f_T\}$) and an exploration policy for generating the actions. Denote the exploration policy by $\mathbf{p}_T = \{p_0, \ldots, p_T\}$ where the probability that action a is taken given history $\{\mathbf{O}_t, \mathbf{A}_{t-1}\}$ is $p_t(a|\mathbf{O}_t, \mathbf{A}_{t-1})$ (if the Markov assumption holds then, as before, replace $\mathbf{O}_t$ with O_t and $\mathbf{A}_{t-1}$ with A_{t-1}.) We assume that $p_t(a|\mathbf{o}_t, \mathbf{a}_{t-1}) > 0$ for each action $a \in$ A and for each possible value $(\mathbf{o}_t, \mathbf{a}_{t-1})$; that is, at each time all actions are possible. Then the likelihood (under P) of the trajectory, $\{o_0, a_0, o_1, \ldots, a_T, o_{T+1}\}$ is

$$f_0(o_0)p_0(a_0 \mid o_0)\prod_{t=1}^{T} f_t(o_t \mid \mathbf{o}_{t-1}, \mathbf{a}_{t-1})p_t(a_t \mid \mathbf{o}_t, \mathbf{a}_{t-1}) f_{T+1}(o_{T+1} \mid \mathbf{o}_T, \mathbf{a}_T). \tag{1}$$

Denote expectations with respect to the distribution P by an E.

Define a deterministic, but possibly non-stationary and non-Markovian, policy, π, as a sequence of decision rules, $\{\pi_1, \ldots, \pi_T\}$, where the output of the time t decision rule, $\pi_t(\mathbf{o}_t, \mathbf{a}_{t-1})$, is an action. Let the distribution P_π denote the distribution of a trajectory whereby the policy π is used to generate the actions. Then the likelihood (under P_π) of the trajectory $\{o_0, a_0, o_1, \ldots, a_T, o_{T+1}\}$ is

$$f_0(o_0)1_{a_0=\pi_0(o_0)}\prod_{j=1}^{T} f_j(o_j \mid \mathbf{o}_{j-1}, \mathbf{a}_{j-1})1_{a_j=\pi_j(\mathbf{o}_j, \mathbf{a}_{j-1})} f_{T+1}(o_{T+1} \mid \mathbf{o}_T, \mathbf{a}_T) \tag{2}$$

where for a predicate W, 1_W is 1 if W is true and is 0 otherwise. Denote expectations with respect to the distribution P_π by an E_π.

Note that since (1) and (2) differ only in regard to the policy for generating actions, an expectation with respect to either P or P_π that does not involve integration over the policy results in the same quantity. For example, $E[R_t|\mathbf{O}_t, \mathbf{A}_t] = E_\pi[R_t|\mathbf{O}_t, \mathbf{A}_t]$, for any policy π.

Let Π be the collection of all policies. In a finite horizon planning problem (permitting non-stationary, non-Markovian policies) the goal is to estimate a policy that maximizes $E_\pi[\sum_{j=1}^{T} R_j \mid O_0 = o_0]$ over $\pi \in \Pi$. If the system dynamics are Markovian and each $r_j(\mathbf{o}_j, \mathbf{a}_j, o_{j+1}) = \gamma j_{r(}o_j, a_j, o_{j+1})$ for r a bounded reward function and $\gamma \in (0,1)$ a discount factor, then this finite horizon problem provides an approximation to the discounted infinite horizon problem (Kearns Mansour and Ng, 2000) for T large.

Given a policy, π, the value function for an observation, o_0, is

$$V_\pi(o_0) = E_\pi\left[\sum_{j=1}^{T} R_j \,\middle|\, O_0 = o_0\right].$$

The t-value function for policy π is the value of the rewards summed from time t on and is

$$V_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) = E_\pi\left[\sum_{j=t}^{T} R_j \,\middle|\, \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_{t-1} = \boldsymbol{a}_{t-1}\right].$$

If the Markovian assumption holds then $(\mathbf{o}_t , \mathbf{a}_{t-1})$ in the definition of $V_{\pi,t}$ is replaced by o_t . Note that the time 0 value function is simply the value function ($V_{\pi,0} = V_\pi$). For convenience, set $V_{\pi,T+1} = 0$. Then the value functions satisfy the following relationship:

$$V_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) = E_\pi[R_t + V_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_{t-1} = \boldsymbol{a}_{t-1}]$$

for $t = 0,...,T$. The time t Q-function for policy π is

$$Q_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_t) = E[R_t + V_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_t = \boldsymbol{a}_t].$$

(The subscript, π, can be omitted as this expectation is with respect to the distribution of O_{t+1} given $(\mathbf{O}_t , \mathbf{A}_t), f_{t+1}$; this conditional distribution does not depend on the policy.) In Section 4 we express the difference in value functions for policy $\boldsymbol{\pi}$ and policy π in terms of the advantages (as defined in Baird, 1993). The time t advantage is

$$\mu_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_t) = Q_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_t) - V_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}).$$

The advantage can be interpreted as the gain in performance obtained by following action a_t at time t and thereafter policy π as compared to following policy π from time t on.

The optimal value function $V^*(o)$ for an observation o is

$$V^*(o) = \max_{\pi\in\Pi} V_\pi(o)$$

and the optimal t-value function for history $(\mathbf{o}_t , \mathbf{a}_{t-1})$ is

$$V_t^*(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) = \max_{\pi\in\Pi} V_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}).$$

As is well-known, the optimal value functions satisfy the Bellman equations (Bellman, 1957)

$$V_t^*(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) = \max_{a_t\in A} E[R_t + V_{t+1}^*(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_t = \boldsymbol{a}_t].$$

Optimal, deterministic, time t decision rules must satisfy

$$\pi_t^*(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t\in A} E[R_t + V_{t+1}^*(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_t = \boldsymbol{a}_t].$$

The optimal time t Q-function is

$$Q_t^*(\boldsymbol{o}_t, \boldsymbol{a}_t) = E[R_t + V_{t+1}^*(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t = \boldsymbol{o}_t, \boldsymbol{A}_t = \boldsymbol{a}_t],$$

and thus the optimal time t advantage, which is given by

$$\mu_t^*(\boldsymbol{o}_t, \boldsymbol{a}_t) = Q_t^*(\boldsymbol{o}_t, \boldsymbol{a}_t) - V_t^*(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}),$$

is always nonpositive and furthermore it is maximized in a_t at $a_t = \pi_t^*(\mathbf{o}_t, \mathbf{a}_{t-1})$.

3. Batch Q-Learning

We consider a version of Q-learning for use in learning a non-stationary, non-Markovian policy with one training set of finite horizon trajectories. The term "batch" Q-learning is used to emphasize that learning occurs only after the collection of the training set. The Q-functions are estimated using an approximator (i.e. neural networks, decision-trees etc.) (Bertsekas and Tsitsiklis, 1996; Tsitsiklis and van Roy, 1997) and then the estimated decision rules are the argmax of the estimated Q functions. Let Q_t be the approximation space for the tth Q-function, e.g. $\mathrm{Q}_t = \{Q_t(\mathbf{o}_t, \mathbf{a}_t; \theta) : \theta \in \theta\}$; θ is a vector of parameters taking values in a parameter space θ which is a subset of a Euclidean space. For convenience set Q_{T+1} equal to zero and write $\mathrm{E}_n f$ for the expectation of an arbitrary function, f, of a trajectory with respect to the probability obtained by choosing a trajectory uniformly from the training set of n trajectories (for example, $E_n[f(O_t)] = 1 \big/ n \sum_{i=1}^n f(O_{it})$ for O_{it} the tth observation in the ith trajectory). In batch Q-learning using dynamic programming and function approximation solve the following backwards through time t =!$T,T-1,...,1$ to obtain

$$\theta_t \in \arg\min_{\theta} E_n\left[R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}; \theta_{t+1}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t; \theta)\right]2. \tag{3}$$

Suppose that Q-functions are approximated by linear combinations of p features ($\mathrm{Q}_t = \{\theta^T q_t(\mathbf{o}_t, \mathbf{a}_t) : \theta \in \mathbf{R}^p\}$) then to achieve (3) solve backwards through time, $t = T, T-1,...,0$,

$$0 = E_n\left[\left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}; \theta_{t+1}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t; \theta_t)\right) q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)^T\right] \tag{4}$$

for θ_t

An incremental implementation with updates between trajectories of (3) and (4) results in one-step Q-learning (Sutton and Barto, 1998, pg. 148, put γ = 1, assume the Markov property and no need for function approximation). This is not surprising as Q-learning can be viewed as approximating least squares value iteration (Tsitsiklis and van Roy, 1996). To see the connection consider the following generic derivation of an incremental update. Denote the ith example in a training set by X_i. Define $\hat{\theta}^n$ to be a solution of $\sum_{i=1}^n f(X_i, \theta) = 0$ for f a given p dimensional vector of functions and each integer n. Using a Taylor series, expand $\sum_{i=1}^{n+1} f(X_i, \hat{\theta}^{(n+1)})$ in $\hat{\theta}^{(n+1)}$ about $\hat{\theta}^{(n)}$ to obtain a between-example update to $\hat{\theta}^{(n)}$:

$$\hat{\theta}^{(n+1)} \leftarrow \hat{\theta}^{(n)} + \frac{1}{n+1}\left(E_{n+1}\left(-\frac{\partial f(X, \hat{\theta}^n)}{\partial \hat{\theta}^n}\right)\right)^{-1} f(X_{n+1}, \hat{\theta}^n).$$

Replace $\frac{1}{n+1}\left(E_{n+1}\left(-\frac{\partial f(X, \hat{\theta}^n)}{\partial \hat{\theta}^n}\right)\right)^{-1}$ by a step-size α_n ($\alpha_n \rightarrow 0$ as $n \rightarrow \infty$) to obtain a general formula for the incremental implementation. Now consider an incremental implementation of (4) for each $t = 0,...,T$. Then for each t, $X = (\mathbf{O}_{t+1}, \mathbf{A}_t)$, $\theta = \theta_t$ and

$$f(X, \theta_t) = \left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}; \hat{\theta}_{t+1}^{(n+1)}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t; \theta_t)\right) q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)^T$$

is a vector of dimension p. The incremental update is

$$\hat{\theta}_t^{(n+1)} \leftarrow \hat{\theta}_t^{(n)} + \alpha_n \left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}; \hat{\theta}_{t+1}^{(n+1)}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t; \hat{\theta}_t^{(n)})\right) q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)^T$$

for $t = 0,...,T$. This is the one-step update of Sutton and Barto (1998, pg. 148) with $\gamma = 1$ and generalized to permit function approximation and nonstationary Q-functions and is analogous to the TD(0) update of Tsitsiklis and van Roy (1997) permitting non-Markovian, nonstationary value functions.

Denote the estimator of the optimal Q-functions based on the training data by $\hat{Q}_t$ for $t = 0,...,T$ (for simplicity, is omitted). The estimated policy, $\hat{\pi}$, satisfies $\hat{\pi}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \hat{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$ for each t. Note that members of the approximation space Q_t need not be "Q-functions" for any policy. For example the Q-functions corresponding to the use of a policy π ($Q_{\pi,t}$, $t = 0,...,T$) must satisfy

$$E[R_t + V_{\pi, t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t) \mid \boldsymbol{O}_t, \boldsymbol{A}_t] = Q_{\pi, t}(\boldsymbol{O}_t, \boldsymbol{A}_t)$$

where $V_{\pi,t+1}(\mathbf{O}_{t+1}, \mathbf{A}_t) = Q_{\pi,t+1}(\mathbf{O}_{t+1}, \mathbf{A}_t, a_{t+1})$ with a_{t+1} set equal to $\pi_{t+1}(\mathbf{O}_{t+1}, \mathbf{A}_t)$. Q-learning does not impose this restriction on $\{\hat{Q}_t, t = 0, \ldots, T\}$; indeed it may be that no member of the approximation space can satisfy this restriction. None-the-less we refer to the $\hat{Q}_t$'s as estimated Q-functions. Note also that the approximation for the Q-functions combined with the definition of the estimated decision rules as the argmax of the estimated Q functions places implicit restrictions on the set of policies that will be considered. In effect the space of interesting polices is no longer Π but rather $\Pi\mathrm{Q} = \{\pi_\theta, \theta \in \theta\}$ where $\pi_\theta = \{\pi_{1,\theta},..., \pi_{T,\theta}\}$ and where each $\pi_{t,\theta}(\mathbf{o}_t , \mathbf{a}_{t-1}) \in \arg\max_{at} Q_t(\mathbf{o}_t , \mathbf{a}_t ; \theta)$for some $Q_t \in \mathrm{Q}_t$.

4. Generalization Error

Define the generalization error of a policy π at an observation o_0 as the average difference between the optimal value function and the value function resulting from the use of policy π in generating a separate test set. The generalization error of policy π at observation o_0 can be written as

$$V^*(o_0) - V_\pi(o_0) = -E_\pi\left[\sum_{t=0}^{T} \mu_t^*(\boldsymbol{O}_t, \boldsymbol{A}_t) \mid O_0 = o_0\right] \quad (5)$$

where E_π denotes the expectation using the likelihood (2). So the generalization error can be expressed in terms of the optimal advantages evaluated at actions determined by policy π; that is when each $A_t = \pi_t(\mathbf{O}_t , \mathbf{A}_{t-1})$. Thus the closer each optimal advantage, μt *($\mathbf{O}_t$, $\mathbf{A}_t$) for $A_t = \pi_t(\mathbf{O}_t , \mathbf{A}_{t-1})$ is to zero, the smaller the generalization error. Recall that the optimal advantage, μt *($\mathbf{O}_t$, $\mathbf{A}_t$), is zero when $A_t = \pi_t$ *($\mathbf{O}_t$, $\mathbf{A}_{t-1}$). The display in (5) follows from Kakade's (ch. 5, 2003) expression for the difference between the value functions for two policies.

Lemma 1

Given policies $\boldsymbol{\pi}$ and π,

$$V_{\boldsymbol{\pi}}(o_0) - V_{\boldsymbol{\pi}}(o_0) = -E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T}\mu_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, O_0 = o_0\right].$$

Set $\boldsymbol{\pi}$ = π* to obtain (5). An alternate to Kakade’s (2003) proof is as follows.

Proof. First note

$$V_{\boldsymbol{\pi}}(o_0) = E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \,\middle|\, O_0 = o_0\right] = E_{\boldsymbol{\pi}}\left[E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \,\middle|\, \boldsymbol{O}_T, \boldsymbol{A}_T\right] \,\middle|\, O_0 = o_0\right]. \tag{6}$$

And $E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \mid \boldsymbol{O}_T, \boldsymbol{A}_T\right]$ is the expectation with respect to the distribution of O_{T+1} given the history ($\mathbf{O}_T$, $\mathbf{A}_T$); this is the density f_{T+1} from Section 2 and f_{T+1} is independent of the policy used to choose the actions. Thus we may subscript E by either π or $\boldsymbol{\pi}$ without changing the expectation; $E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \mid \boldsymbol{O}_T, \boldsymbol{A}_T\right] = E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \mid \boldsymbol{O}_T, \boldsymbol{A}_T\right] = \sum_{t=0}^{T-1} R_t + Q_{\boldsymbol{\pi},T}(\boldsymbol{O}_T, \boldsymbol{A}_T).$ The conditional expectation can be written in a telescoping sum as

$$\begin{aligned} E_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} R_t \,\middle|\, \boldsymbol{O}_T, \boldsymbol{A}_T\right] &= \sum_{t=0}^{T} Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t) - V_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}) \\ &+ \sum_{t=1}^{T} R_{t-1} + V_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}) - Q_{\boldsymbol{\pi},t-1}(\boldsymbol{O}_{t-1}, \boldsymbol{A}_{t-1}) \\ &+ V_{\boldsymbol{\pi},0}(O_0) \end{aligned}$$

The first sum is the sum of the advantages. The second sum is a sum of temporal-difference errors; integrating the temporal-difference error with respect to the conditional distribution of O_t given ($\mathbf{O}_{t-1}$, $\mathbf{A}_{t-1}$), denoted by f_t in Section 2, we obtain zero,

$$E\left[R_{t-1} + V_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}) \,\middle|\, \boldsymbol{O}_{t-1}, \boldsymbol{A}_{t-1}\right] = Q_{\boldsymbol{\pi},t-1}(\boldsymbol{O}_{t-1}, \boldsymbol{A}_{t-1})$$

(as before E denotes expectation with respect to (1); recall that expectations that do not integrate over the policy can be written either with an E or an E_π). Substitute the telescoping sum into (6) and note that $V_{\boldsymbol{\pi},0}(O_0) = V_{\boldsymbol{\pi}}(O_0)$ to obtain the result. ▪

In the following Lemma the difference between value functions corresponding to two policies, $\boldsymbol{\pi}$ and π, is expressed in terms of both the L_1 and L_2 distances between the optimal Q-functions and *any* functions $\{Q_0, Q_1, \ldots, Q_T\}$ satisfying π_t ($\mathbf{o}_t$, $\mathbf{a}_{t-1}$) ∈ arg max$_{at}$ Q_t ($\mathbf{o}_t$, $\mathbf{a}_t$), t = 0,...,T and *any* functions $\left\{\mathcal{Q}_0, \mathcal{Q}_1, \ldots, \mathcal{Q}_T\right\}$ satisfying $\boldsymbol{\pi}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \mathcal{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$, t = 0,...,T. We assume that there exists a positive constant, L for which p_t (a_t|$\mathbf{o}_t$, $\mathbf{a}_{t-1}$) ≥ L^{-1} for each t and all pairs ($\mathbf{o}_t$, $\mathbf{a}_{t-1}$); if the stochastic decision rule, p_t, were uniform then L would be the size of the action space.

Lemma 2

For all functions, Q_t satisfying π_t($\mathbf{o}_t$, $\mathbf{a}_{t-1}$) ∈ arg max$_{at}$ Q_t($\mathbf{o}_t$, $\mathbf{a}_t$), t = 0,...,T, and all functions $\mathcal{Q}_t$ satisfying $\boldsymbol{\pi}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \mathcal{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$, t = 0,...,T we have,

$$\left| V_{\pi}(o_0) - V_{\pi}(o_0) \right| \le \sum_{t=0}^{T} 2L^{\;t+1} E\left[\left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) \right| O_0 = o_0 \right]$$

$$+ \sum_{t=0}^{T} 2L^{\;t+1} E\left[\left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_{\pi,t}(\boldsymbol{O}_t, \boldsymbol{A}_t) \right| O_0 = o \right]$$

and

$$\left| V_{\pi}(o_0) - V_{\pi}(o_0) \right| \le \sum_{t=0}^{T} 2L^{\;(t+1)/2} \sqrt{E\left[(Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t))^2 \,\middle|\, O_0 = o_0 \right]}$$

$$+ \sum_{t=0}^{T} 2L^{\;(t+1)/2} \sqrt{E\left[(\mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_{\pi,t}(\boldsymbol{O}_t, \boldsymbol{A}_t))2 \,\middle|\, O_0 = o_0 \right]},$$

where E denotes expectation with respect to the distribution generating the training sample (1).

Remark:

1. Note that in general arg $\max_{at} Q_{\pi,t}(\boldsymbol{o}_t, \boldsymbol{a}_t)$ may not be π_t thus we can not choose $\mathcal{Q}_t = Q_{\pi,t}$. However if $\pi = \pi^*$ then we can choose $\mathcal{Q}_t = Q_t^*(= Q_{\pi^*,t}\,\text{by definition})$ and the second term in both upper bounds is equal to zero.

2. This result can be used to emphasize one aspect of the mismatch between estimating the optimal Q function and the goal of learning a policy that maximizes the value function. Suppose $\mathcal{Q}_t = Q_t^*,\ \pi = \pi^*$. The generalization error is

$$V^*(o_0) - V_{\pi}(o_0) \le \sum_{t=0}^{T} 2L^{\;(t+1)/2} \sqrt{E\left[(Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_t^*(\boldsymbol{O}_t, \boldsymbol{A}_t))^2 \,\middle|\, O_0 = o_0 \right]}$$

for Q_t any function satisfying $\pi_t(\mathbf{o}_t, \mathbf{a}_{t-1}) \in$ arg $\max_{at} Q_t(\mathbf{o}_t, \mathbf{a}_t)$. Absent restrictions on the Q_t s, this inequality cannot be improved in the sense that choosing each $Q_t = Q_t^*$ and $\pi_t = \pi_t^*$ yields 0 on both sides of the inequality. However an inequality in the opposite direction is not possible, since as was seen in Lemma 1, $V^*(o_0) - V_\pi(o_0)$ involves the Q functions only through the advantages (see also (7) below with $\pi = \pi^*$). Thus for the difference in value functions to be small, the average difference between $Q_t(\mathbf{o}_t, \mathbf{a}_t) - \max_{at} Q_t(\mathbf{o}_t, \mathbf{a}_t)$ and $Q^*_t(\mathbf{o}_t, \mathbf{a}_t) - \max_{at} Q_t^*(\mathbf{o}_t, \mathbf{a}_t)$ must be small; this does not require that the average difference between $Q_t(\mathbf{o}_t, \mathbf{a}_t)$ and $Q_t^*(\mathbf{o}_t, \mathbf{a}_t)$ is small. The mismatch is not unexpected. For example, Baxter and Bartlett (2001) provide an example in which the approximation space for the value function includes a value function for which the greedy policy is optimal, yet the greedy policy found by temporal difference learning (TD(1!)) function performs very poorly.

Proof—Define $\mu_t(\mathbf{o}_t, \mathbf{a}_t) = Q_t(\mathbf{o}_t, \mathbf{a}_t) - \max_{at} Q_t(\mathbf{o}_t, \mathbf{a}_t)$ for each t; note that $\mu_t(\mathbf{o}_t, \mathbf{a}_t)$ evaluated at $a_t = \pi_t(\mathbf{o}_t, \mathbf{a}_{t-1})$ is zero. Start with the result of Lemma 1. Then note the difference between the value functions can be expressed as

$$V_{\pi}(o_0) - V_{\pi}(o_0) = \sum_{t=0}^{T} E_{\pi}\left[\mu_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - \mu_{\pi,t}(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, O_0 = o_0 \right]. \qquad (7)$$

since P_π puts $a_t = \pi_t(\mathbf{o}_t, \mathbf{a}_{t-1})$ and $\mu_t(\mathbf{o}_t, \mathbf{a}_t) = 0$ for this value of a_t. When it is clear from the context μ_t ($\mu_{\pi,t}$) is used as abbreviation for $\mu_t(\mathbf{O}_t, \mathbf{A}_t)$ ($\mu_{\pi,t}(\boldsymbol{O}_t, \boldsymbol{A}_t)$) in the following. Also $Q_{\pi,t}(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)$ with a_t replaced by $\pi_t(\mathbf{O}_t, \mathbf{A}_{t-1})$ is written as $Q_{\pi,t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \pi_t)$. Consider the absolute value of the tth integrand in (7):

$$
\begin{aligned}
&\left|\mu_t - \mu_{\boldsymbol{\pi},t}\right| \\
&= \left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - \max_{a_t} Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t) + Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t) \right| \\
&\le \left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t) \right| + \left| \max_{a_t} Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t) \right| .
\end{aligned}
$$

Since $\max_{at}\ \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t)$and for any functions, h and h', $|\max_{at} h(a_t) - \max_{at} h'(a_t)| \le \max_{at} |h(a_t) - h'(a_t)|$,

$$
\begin{aligned}
&\left| \max_{a_t} Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t) \right| \\
&\le \max_{a_t} \left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| \\
&+ \left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, \boldsymbol{\pi}_t) \right| .
\end{aligned}
$$

We obtain $\left|\mu_t - \mu_{\boldsymbol{\pi},t}\right|$

$$
\begin{aligned}
&\le 2\max_{a_t} \left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| + 2\max_{a_t} \left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| \\
&\le 2L \sum_{a_t} \left| Q_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| p_t(a_t \mid \boldsymbol{O}_t, \boldsymbol{A}_{t-1}) \\
&+2L \sum_{a_t} \left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| p_t(a_t \mid \boldsymbol{O}_t, \boldsymbol{A}_{t-1}).
\end{aligned}
\tag{8}
$$

Insert the above into (7) and use Lemma A1 to obtain $\left| V_{\boldsymbol{\pi}}(o_0) - V_{\pi}(o_0) \right|$

$$
\begin{aligned}
&\le 2L \sum_{t=0}^{T} E_{\pi}\left[\sum_{a_t} \left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| p_t(a_t \mid \boldsymbol{O}_t, \boldsymbol{A}_{t-1}) \middle| O_0 = o_0 \right] \\
&+ E_{\pi}\left[\sum_{a_t} \left| \mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_{t-1}, a_t) \right| p_t(a_t \mid \boldsymbol{O}_t, \boldsymbol{A}_{t-1}) \middle| O_0 = o_0 \right] \\
&= 2L \sum_{t=0}^{T} E\left[\left(\prod_{\ell=0}^{t-1} \frac{1_{A_\ell = \pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell\ \boldsymbol{A}_{\ell-1})} \right) \left| Q_t - \mathcal{Q}_t \right| \middle| O_0 = o_0 \right] \\
&+ E\left[\left(\prod_{\ell=0}^{t-1} \frac{1_{A_\ell = \pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell\ \boldsymbol{A}_{\ell-1})} \right) \left| \mathcal{Q}_t - Q_{\boldsymbol{\pi},t} \right| \middle| O_0 = o_0 \right] \\
&\le 2\sum_{t=0}^{T} L^{t+1} E\left[\left| Q_t - \mathcal{Q}_t \right| \middle| O_0 = o_0 \right] + 2\sum_{t=0}^{T} L^{t+1} E\left[\left| \mathcal{Q}_t - Q_{\boldsymbol{\pi},t} \right| \middle| O_0 = o_0 \right]
\end{aligned}
$$

(Q_t, $Q_{\boldsymbol{\pi},t}$ is used as abbreviation for $Q_t(\mathbf{O}_t, \mathbf{A}_t)$, respectively $Q_{\boldsymbol{\pi},t}(\mathbf{O}_t, \mathbf{A}_t)$). This completes the proof of the first result.

Start from (8) and use Hölder's inequality to obtain, $\left| V_{\boldsymbol{\pi}}(o_0) - V_{\pi}(o_0) \right|$

$$\le 2\sum_{t=0}^{T} E_\pi\left[\max_{a_t} \left| Q_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right| \,\middle|\, O_0 = o_0\right]$$

$$+E_\pi\left[\max_{a_t} \left| \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - Q_{\pi,t}(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right| \,\middle|\, O_0 = o_0\right]$$

$$\le 2\sum_{t=0}^{T} \sqrt{E_\pi\left[\max_{a_t} \left| Q_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right|^2 \,\middle|\, O_0 = o_0\right]}$$

$$+\sqrt{E_\pi\left[\max_{a_t} \left| \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - Q_{\pi,t}(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right|^2 \,\middle|\, O_0 = o_0\right]}$$

$$\le 2\sum_{t=0}^{T} \sqrt{L\ E_\pi\left[\sum_{a_t} \left| Q_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right|^2 P_t(a_t \mid \mathbf{O}_t, \mathbf{A}_{t-1}) \,\middle|\, O_0 = o_0\right]}$$

$$+2\sum_{t=0}^{T} \sqrt{L\ E_\pi\left[\sum_{a_t} \left| \mathcal{Q}_t(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t) - Q_{\pi,t}(\mathbf{O}_t, \mathbf{A}_{t-1}, a_t)\right|^2 P_t(a_t \mid \mathbf{O}_t, \mathbf{A}_{t-1}) \,\middle|\, O_0 = o_0\right]}.$$

Now use Lemma A1 and the lower bound on the p_t's to obtain the result,

$$\left| V_\pi(o_0) - V_\pi(o_0)\right| \le 2\sum_{t=0}^{T} \sqrt{LE\left[\prod_{\ell=0}^{t-1} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \mathbf{O}_\ell, \mathbf{A}_{\ell-1})}(Q_t - \mathcal{Q}_t)^2 \,\middle|\, O_0 = o_0\right]}$$

$$+\sqrt{LE\left[\prod_{\ell=0}^{t-1} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \mathbf{O}_\ell, \mathbf{A}_{\ell-1})}(\mathcal{Q}_t - Q_{\pi,t})^2 \,\middle|\, O_0 = o_0\right]}$$

$$\le 2L^{(t+1)/2}\sum_{t=0}^{T} \sqrt{E\left[(Q_t - \mathcal{Q}_t)^2 \,\middle|\, O_0 = o_0\right]}$$

$$+\sqrt{E\left[(\mathcal{Q}_t - Q_{\pi,t})^2 \,\middle|\, O_0 = o_0\right]}.$$

■

5. Finite Sample Upper Bounds on the Average Generalization Error

Traditionally the performance of a policy π is evaluated in terms of maximum generalization error: $\max_o [V^*(o) - V_\pi(o)]$ (Bertsekas and Tsitsiklis, 1996). However here we consider an average generalization error as in Kakade (2003) (see also Fiechter, 1997; Kearns, Mansour and Ng, 2000; Peshkin and Shelton, 2002); that is $\int_o [V^*(o) - V_\pi(o)]\, dF(o)$ for a specified distribution F on the observation space. The choice of F with density $f = f_0$ (f_0 is the density of O_0 in likelihoods (1) and (2)) is particularly appealing in the development of a policy in many medical and social science applications. In these cases, f_0 represents the distribution of initial observations corresponding to a particular population of subjects. The goal is to produce a good policy for this population of subjects. In general as in Kakade (2003) F may be chosen to incorporate domain knowledge concerning the steady state dis! tribution of a good policy. If only a training set of trajectories is available for learning and we are unwilling to assume that the system dynamics are Markovian, then the choice of F is constrained by the following consideration. If the distribution of O_0 in the training set (f_0) assigns mass zero to an observation o', then the training data will not be able to tell us anything about $V_\pi(o')$. Similarly if f_0 assigns a very small positive mass to o' then only an exceptionally large training set will permit an accurate estimate of $V_\pi(o')$. Of course this will not be a problem for the average generalization error, as long as F also assigns very low mass to o'. Consequently in our construction of the finite sample error bounds for the *average* generalization error, we will only consider distributions F for which the density of F, say f, satisfies $\sup_o \left| \frac{f(o)}{f_0(o)} \right| \le M$for some finite constant M. In this case the average generalization error is bounded above by

$$\begin{aligned} \int V^*(o) - V_\pi(o)\, dF(o) &\le ME\left[V^*(O_0) - V_\pi(O_0) \right] \\ &= -ME_\pi\left[\sum_{t=0}^{T} \mu_t^*(\boldsymbol{O}_t, \boldsymbol{A}_t) \right]. \end{aligned}$$

The second line is a consequence of (5) and the fact that the distribution of O_0 is the same under likelihoods (1) and (2).

In the following theorem a non-asymptotic upper bound on the average generalization error is provided; this upper bound depends on the number of trajectories in the training set (n), the performance of the approximation on the training set, the complexity of the approximation space and of course on the confidence (δ) and accuracy (ε) demanded. The batch Q-learning algorithm minimizes quadratic forms (see (3)); thus we represent the performance of functions $\{Q_0, Q_1, \ldots, Q_T\}$ on the training set by these quadratic forms,

$$Err_{n, Q_{t+1}}(Q_t) = E_n\left[R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t) \right]2$$

for each j (recall Q_{T+1} is set to zero and E_n represents the expectation with respect to the probability obtained by choosing a trajectory uniformly from the training set).

The complexity of each Q_t space can be represented by it's covering number (Anthony and Bartlett, 1999, pg 148). Suppose F is a class of functions from a space, X, to $\mathbb{R}$. For a sequence $x = (x_1, \ldots, x_n) \in X^n$, define $\mathrm{F}_{|x}$ to be a subset of $\mathbb{R}^n$ given by $\mathrm{F}_{|x} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathrm{F}\}$. Define the metric dp on $\mathbb{R}^n$ by $d_p(z, y) = \left(1 / n\sum_{i=1}^{n} |z_i - y_i|^p\right)^{1/p}$ for p a positive integer (for p=∞, define $d_\infty(z, y) = \max_{i=1}^{n} |z_i - y_i|$). Then $\mathrm{N}(\varepsilon, \mathrm{F}_{|x}, d_p)$ is defined as the minimum cardinality of an ε-covering of $\mathrm{F}_{|x}$ with respect to the metric d_p. Next given ε>0, positive integer n, metric d_p and function class, F, the covering number for F is defined as

$$N_p(\varepsilon, F, n) = \max\left\{ N(\varepsilon, F_{|x}, d_p) : x \in X^n \right\}.$$

In the following theorem, $\mathrm{F} = \{\max_{a_{t+1}} Q_{t+1}(\mathbf{o}_{t+1}, \mathbf{a}_t) - Q_t(\mathbf{o}_t, \mathbf{a}_t) : Q_t \in \mathrm{Q}_t, t = 0, \ldots, T\}$ and $(x)^+$ is x if $x > 0$ and zero otherwise.

Theorem 1

Assume that the functions in Q_t, $t \in 0, \ldots, T$ are uniformly bounded. Suppose that there exists a positive constant, say L, for which $p_t(a_t | \mathbf{o}_t, \mathbf{a}_t) \ge L^{-1}$ for all $(\mathbf{o}_t, \mathbf{a}_t)$ pairs, $0 \ge t \ge T$. Then for $\varepsilon > 0$ and with probability at least 1–δ, over the random choice of the training set, every choice of functions, $Q_j \in Q_j, j = 0, \ldots, T$ with associated policy π defined by $\pi_j(\mathbf{o}_j, \mathbf{a}_{j-1}) = \arg\max_{aj} Q_j(\mathbf{o}_j, \mathbf{a}_{j-1})$ and every choice of functions, $\mathcal{Q}_j \in \mathrm{Q}_j,\ j = 0, \ldots, T$ with associated policy π defined by $\pi_j(\boldsymbol{o}_j, \boldsymbol{a}_{j-1}) = \arg\max_{a_j} \mathcal{Q}_j(\boldsymbol{o}_j, \boldsymbol{a}_j)$ the following bound is satisfied,

$$\begin{aligned}&\int \left| V_{\pi}(o) - V_{\pi}(o) \right| dF(o)\\ &\leq 6ML^{1/2} \sum_{t=0}^{T} \left[\sum_{i=t}^{T} (16)^{i-t} L^{i} (Err_{n,Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(\tilde{Q}_i)) + \right]^{1/2}\\ &+12ML^{1/2}\varepsilon\\ &+6\mathrm{ML}^{1/2} \sum_{t=0}^{T} \sum_{i=t}^{T} (16)^{(i-t)/2} L^{i/2} \sqrt{E\left[\tilde{Q}_i(\boldsymbol{O}_i, \boldsymbol{A}_i) - Q_{\pi,i}(\boldsymbol{O}_i, \boldsymbol{A}_i) \right]^2}.\end{aligned}$$

for n satisfying

$$4(T+1)N_1\left(\frac{\varepsilon^2}{32M'(16L)^{(T+2)}}, F, 2n \right) \exp\left\{ \frac{\varepsilon^4 n}{32(M')^2(16L)^{2(T+2)}} \right\} \leq \delta \tag{9}$$

and where M' is a uniform upper bound on the absolute value of $f \in$ F and E represents the expectation with respect to the distribution (1) generating the training set.

Remarks:

1. Suppose that $Q^*_t \in Q_t$ for each select t. Select $\tilde{Q}_t = Q_t^*$ and $\hat{Q}_t = \arg\min_{Q_t \in Q_t} Err_{n,\hat{Q}_{t+1}}(Q_t)$, $t = T, T-1, \ldots, 0$ (recall Q_{T+1}, $\hat{Q}_{T+1}$ are identically zero). Then with probability greater than $1 - \delta$, we obtain,

$$\int V^*(o) - V_{\hat{\pi}}(o) dF(o) \leq 12ML^{1/2}\varepsilon \tag{10}$$

for all n satisfying (9). Thus, as long as the covering numbers for each Q_t and thus for F do not grow too fast, estimating each Q_t by minimizing $Err_{n,\hat{Q}_{t+1}}(Q_t)$ yields a policy that consistently achieves the optimal value. Suppose the approximation spaces Q_t , $t = 0,\ldots,T$ are feed-forward neural networks as in remark 4 below. In this case the training set size n sufficient for (10) to hold need only be polynomial in (1/δ, 1/ε) and batch Q-learning is a probably approximate correct (PAC) reinforcement learning algorithm as defined by Fiechter (1997). As shown by Fiechter (1997) this algorithm can be converted to an efficient on-line reinforcement learning algorithm (here the word on-line implies updating the policy between trajectories).

2. Even when Q_t* does not belong to Q_t we can add the optimal Q function at each time, t, to the approximation space, Q_t with a cost of no more than an increase of 1 to the covering number $N_1\left(\frac{\varepsilon^2}{32M'(16L)^{(T+2)}}, F, 2n \right)$. If we do this the result continues to hold when we set $\boldsymbol{\pi}$ to an optimal policy $\boldsymbol{\pi}^*$ and set $\tilde{Q}_t = Q_t^*$ for each t; the generalization error is

$$\begin{aligned}&\int V^*(o) - V_{\pi}(o) dF(o) \leq 6ML^{1/2} \sum_{t=0}^{T} \left[\sum_{i=t}^{T} (16)^{i-t} L^{i} Err_{n,Q_{i+1}}(Q_i) \right]^{1/2}\\ &+12ML^{1/2}\varepsilon\end{aligned}$$

for all n satisfying (9). This upper bound is consistent with the practice of using a policy $\hat{\boldsymbol{\pi}}$ for which $\hat{\boldsymbol{\pi}}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \hat{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$ and $\hat{Q}_t \in \arg\min_{Q_t \in Q_t} Err_{n,\hat{Q}_{t+1}}(Q_t)$. Given that the covering numbers for the approximation space can be expressed in a sufficiently simple form (as in remark 4 below), this upper bound can be used to carry out model selection using structural risk minimization (Vapnik, 1982). That is, one might consider a variety of approximation spaces and use structural risk minimization to use the training data to choose which approximation space is best. The resulting upper bound on the average generalization error can be found by using the above result and Lemma 15.5 of Anthony and Bartlett (1999).

3. The restriction on n in (9) is due to the complexity associated with the approximation space (e.g. the Q_t ’s). The restriction is crude; to see this, note that if there were only a finite number of functions in F then n need only satisfy

$$2(T+1)\mid F\mid \exp\left\{\frac{2\varepsilon^4 n}{(3M^{\prime})^2(16L)^{2(T+2)}}\right\}=\delta$$

(use Hoeffding’s inequality; see Anthony and Bartlett, pg 361, 1999) and thus for a given (ε, &δ) we may set the number of trajectories in the training set n equal to $\frac{(3M^{\prime})^2(16L)^{2(T+2)}}{2\varepsilon^4}\ \ln\left(\frac{2(T+1)\mid F\mid}{\delta}\right)$.This complexity term appears similar to that achieved by learning algorithms (e.g. see Anthony and Bartlett, 1999, pg. 21) or in reinforcement learning (e.g. Peshkin and Shelton, 2002) however note that n is of the order ε^{-4} rather than the usual ε^{-2}. The ε^{-4} term (instead of ε^{-2}) is attributable to the fact that $Err_{Q_{t+1}}(Q_t)$ is not only a function of Q_t but also of Q_{t+1}. However further assumptions on the approximation space permit an improved result. See Theorem 2 below for one possible refinement. Note the needed training set size n depends exponentially on the horizon time T but not on the dimension of the observation space. Thi! s is not unexpected as the upper bounds on the generalization error of both Kearns, Mansour and Ng (2000) and Peshkin and Shelton’s (2002) policy search methods (the latter using a training set and importance sampling weights) also depend exponentially on the horizon time.

4. When F is infinite, we use covering numbers for the approximation space Q_t and then appeal to Lemma A2 in the appendix to derive a covering number for F; this results in

$$N_1(\varepsilon,\ F,\ n)\leq(T+1)\max_{t=0,\ldots,T} N_1\left(\frac{\varepsilon}{2\mid A\mid},\ Q_t,\ \mid A\mid n\right)^2.$$

One possible approximation space is based on feed-forward neural networks. From Anthony and Bartlett (1999) we have that if each Q_j is the class of functions computed by a feed-forward network with W weights and k computation units arranged in L layers and each computation unit has a fixed piecewise-polynomial activation function with q pieces and degree no more than ℓ, then $N_1(\varepsilon,\ Q_t,\ n)\leq e(d+1)\left(\frac{2eM^{\prime}}{\varepsilon}\right)^d$ where $d = 2(W + 1)(L + 1) \log_2(4(W + 1)(L + 1) q(k + 1)/\ln 2) + 2(W + 1)(L + 1)^2 \log_2(\ell + 1) + 2(L + 1)$. To see this combine Anthony and Bartlett’s Theorems 8.8, 14.1 and 18.4. They provide covering numbers for functions computed by other types of neural networks as well. A particularly simple neural network is an affine combination of a given set of p input features; i.e. $f(x)=\omega_0+\sum_{i=1}^{p-1}\omega_i x_i$ for $(1, x)$ a vector of p real valued features and each $\omega_i \in \mathbb{R}$. Suppose each Q_t is a class of functions computed by this network. Then Theorems 11.6 and 18.4 of Anthony and Bartlett imply that $N_1(\varepsilon,\ Q_t,\ n)\leq e(p+1)\left(\frac{2eM^{\prime}}{\varepsilon}\right)^p$. In this case

$$n\geq\frac{32(M^{\prime})^4(16L)^{2(T+2)}}{\varepsilon^4}\ \log\left(\frac{4(T+1)^2\,e^2(p+1)^2(128e\mid A\mid(M^{\prime})^2(16L)^{(T+2)})^{2p}}{\delta\varepsilon^4 p}\right).$$

This number will be large for any reasonable accuracy, ε and confidence, δ.

Proof of Theorem 1—An upper bound on the average difference in value functions can be obtained from Lemma 2 by using Jensen's inequality and the assumption that the the density of $F\ (f)$ satisfies $\sup_o \left| \frac{f(o)}{f_0(o)} \right| \le M$ for some finite constant M:

$$\int \left| V_{\boldsymbol{\pi}}(o) - V_{\boldsymbol{\pi}}(o) \right| dF(o) \le M \sum_{t=0}^{T} 2L^{(t+1)/2} \sqrt{E\left[Q_t(\boldsymbol{O}_i, \boldsymbol{A}_i) - \mathcal{Q}_t(\boldsymbol{O}_i, \boldsymbol{A}_i) \right]^2}$$
$$+ M \sum_{t=0}^{T} 2L^{(t+1)/2} \sqrt{E\left[\mathcal{Q}_t - Q_{\boldsymbol{\pi},t} \right]^2} \qquad (11)$$

where $\mathcal{Q}_t - Q_{\boldsymbol{\pi},t}$ is used as abbreviation for $\mathcal{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t)$, respectively $Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t)$. In the following an upper bound on each $E\left[Q_t - \mathcal{Q}_t \right]^2$ is constructed.

The performance of the approximation on an infinite training set can be represented by

$$Err_{Q_{t+1}}(Q_t) = E\left[R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t \right]^2$$

for each t (recall $Q_{T+1} = 0$, also we abbreviate $Q_t\ (\mathbf{O}_t, \mathbf{A}_t\ )$ by Q_t whenever no confusion may arise). The errors, *Err*'s, can be used to provide an upper bound on the L_2 norms on the Q-functions by the following argument. Consider $Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\mathcal{Q}_t)$ for each t. Within each of these quadratic forms add and subtract

$$Q_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \pi_{t+1}) - Q_{\pi,t} - E\left[\max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \pi_{t+1}) \,\middle|\, \boldsymbol{O}_t, \boldsymbol{A}_t \right].$$

In the above $Q_{\boldsymbol{\pi},t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \boldsymbol{\pi}_{t+1})$ is defined as $Q_{\boldsymbol{\pi},t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1})$ with a_{t+1} replaced by $\boldsymbol{\pi}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t)$. Expand each quadratic form and use the fact that $E\left[R_t + Q_{\boldsymbol{\pi},t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \boldsymbol{\pi}_{t+1}) \,\middle|\, \boldsymbol{O}_t, \boldsymbol{A}_t \right] = Q_{\boldsymbol{\pi},t}$. Cancelling common terms yields

$$E\left[Q_{\pi,t} - Q_t + E\left[\max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \pi_{t+1}) \,\middle|\, \boldsymbol{O}_t, \boldsymbol{A}_t \right] \right] 2$$
$$- E\left[Q_{\pi,t} - \mathcal{Q}_t + E\left[\max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \pi_{t+1}) \,\middle|\, \boldsymbol{O}_t, \boldsymbol{A}_t \right] \right] 2.$$

Add and subtract $\mathcal{Q}_t$ in the first quadratic form and expand. This yields

$$Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\mathcal{Q}_t) =$$
$$E[\mathcal{Q}_t - Q_t]^2 + 2E[\mathcal{Q}_t - Q_t][\mathcal{Q}_t - Q_{\pi,t}]$$
$$+ 2E\left[\left(\mathcal{Q}_t - Q_t \right) \left(\max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \mathcal{Q}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) \right) \right]$$
$$+ 2E\left[\left(\mathcal{Q}_t - Q_t \right) \left(\max_{a_{t+1}} \mathcal{Q}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_{\pi,t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \boldsymbol{\pi}_{t+1}) \right) \right]. \qquad (12)$$

Using the arguments similar to those used around Equation (8) and using the fact that $(x + y)^2 \le 2x^2 + 2y^2$ we obtain,

$$
\begin{aligned}
&Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\tilde{Q}_t) \geq E\left[Q_t - \tilde{Q}_t\right]^2 \\
&- 4\left(E\left[Q_t - \tilde{Q}_t\right]^2\left(E\left[\tilde{Q}_t - Q_{\pi,t}\right]^2 + LE\left[Q_{t+1} - \tilde{Q}_{t+1}\right]^2 + LE\left[\tilde{Q}_{t+1} - Q_{\pi,t+1}\right]^2\right)\right)^{1/2}.
\end{aligned}
$$

Using this inequality we can now derive an upper bound on each $E\left[Q_t - \tilde{Q}_t\right]^2$ in terms of the Err’s and the $E\left[\tilde{Q}_{t+1} - Q_{\pi,t+1}\right]$’s. Define

$$m_t = L^{-(T-t)}E\left[\tilde{Q}_t - Q_{\pi,t}\right]^2 \text{ and } b_t = L^{-(T-t)}E\left[Q_t - \tilde{Q}_t\right]^2$$

and

$$e_t = L^{-(T-t)}\left(Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\tilde{Q}_t)\right)$$

for $t \leq T$ and $b_{T+1} = m_{T+1} = e_{T+1} = 0$. We obtain

$$e_t \geq b_t - 4\sqrt{b_t(m_t + b_{t+1} + m_{t+1})}.$$

Completing the square, reordering terms, squaring once again and using the inequality $(x + y)^2 \leq 2x^2 + 2y^2$ yields $b_t \leq 16(b_{t+1} + m_t + m_{t+1}) + 2e_t$ for $t \leq T$. We obtain

$$b_{T-t} \leq 2\sum_{i=0}^{t}(16)^i e_{T-t+i} + \sum_{i=1}^{t}(16)^i(16+1)m_{T-t+i} + 16m_{T-t}.$$

Inserting the definitions of b_{T-t}, e_{T-t+i} and reordering, yields

$$
\begin{aligned}
E\left[Q_t - \tilde{Q}_t\right]^2 &\leq 2\sum_{i=t}^{T}(16L)^{i-t}\left(Err_{Q_{i+1}}(Q_i) - Err_{Q_{i+1}}(\tilde{Q}_i)\right) \\
&+ \sum_{i=t+1}^{T}(16)^{i-t}(16+1)L^{T-t}m_i + L^{T-t}m_t.
\end{aligned}
\tag{13}
$$

As an aside we can start from (12) and derive the upper bound,

$$
\begin{aligned}
&Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\tilde{Q}_t) \leq E\left[Q_t - \tilde{Q}_t\right]^2 \\
&+4L^{(T-t)}\sqrt{L^{-(T-t)}E\left[Q_t - \tilde{Q}_t\right]^2\left(m_t + L^{-(T-t-1)}E\left[Q_{t+1} - \tilde{Q}_{t+1}\right]^2 + m_{t+1}\right)}.
\end{aligned}
$$

This combined with (13) implies that minimizing each $Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\tilde{Q}_t)$ in Q_t is equivalent to minimizing each $E\left[Q_t - \tilde{Q}_t\right]^2$ in Q_t modulo the approximation terms m_t for $t = 0,\ldots, T$.

Returning to the proof next note that

$$
\begin{aligned}
&Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(\tilde{Q}_t) \leq \left|Err_{Q_{t+1}}(Q_t) - Err_{n,Q_{t+1}}(Q_t)\right| \\
&+\left|Err_{Q_{t+1}}(\tilde{Q}_t) - Err_{n,Q_{t+1}}(\tilde{Q}_t)\right| \\
&+\left(Err_{n,Q_{t+1}}(Q_t) - Err_{n,Q_{t+1}}(\tilde{Q}_t)\right)+
\end{aligned}
$$

where $(x)^+$ is equal to x if $x \geq 0$ and is equal to 0 otherwise. Note that if each Q_t minimizes $Err_{n,Qt+1}$ as in (3) then the third term is zero. Substituting into (13), we obtain

$$
\begin{aligned}
&E\left[Q_t - \tilde{Q}_t\right]^2 \leq 2\sum_{i=t}^{T}(16L\ )^{i-t}\Big(\left|\ Err_{Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(\tilde{Q}_i)\right| \\
&+ \left|\ Err_{Q_{i+1}}(\tilde{Q}_i) - Err_{n,Q_{i+1}}(\tilde{Q}_i)\right| \\
&+ \left(Err_{n,Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(\tilde{Q}_i)\right)+\Big) \\
&+ \sum_{i=t+1}^{T}(16)^{i-t}(16+1)L\ ^{i-t}E\left[\tilde{Q}_i - Q_{\pi,i}\right]^2 + E\left[\tilde{Q}_t - Q_{\pi,t}\right]^2.
\end{aligned}
$$

Combine this inequality with (11); simplify the sums and use the fact that for x, y both nonnegative $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ to obtain $\int\left|\ V_{\pi}(o) - \ V_{\pi}(o)\right| dF(o)$

$$
\begin{aligned}
&\leq 6ML\ ^{1/2}\sum_{t=0}^{T}\left[\sum_{i=t}^{T}(16)^{i-t}L\ ^{i}\left(Err_{n,Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(\tilde{Q}_i)\right)+\right]^{1/2} \\
&+12ML\ ^{1/2}(16L\ )^{(T+2)/2}\sqrt{\max_{t}\ \sup_{Q_t, Q_{t+1}}\left|\ Err_{Q_{t+1}}(Q_t) - Err_{n,Q_{t+1}}(Q_t)\right|} \\
&+6ML\ ^{1/2}\sum_{t=0}^{T}\sum_{i=t}^{T}(16)^{(i-t)/2}L\ ^{i/2}\sqrt{E\left[\tilde{Q}_i - Q_{\pi,i}\right]^2}.
\end{aligned}
$$

All that remains is to provide an upper bound on

$$
P\left[\bigcup_{i=0}^{T}\left\{\text{for some}_{Q_t \in Q_t,\ t=0,\dots,T}\left|\ Err_{Q_{t+1}}(Q_i) - Err_{n,Q_{i+1}}(Q_i)\right)\right| > \varepsilon'\right\}\right].
$$

This probability is in turn bounded above by

$$
\sum_{i=0}^{T}P\left[\text{for some}_{Q_t \in Q_t,\ t=0,\dots,T}\left|\ Err_{Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(Q_i)\right| > \varepsilon'\right].
$$

Anthony and Bartlett (1999, pg. 241) use Hoeffding’s inequality along with the classical techniques of symmetrization and permutation to provide the upper bound (see also van der Vaart and Wellner, 1996),

$$
\begin{aligned}
&P\left[\text{for some}_{Q_t \in Q_t,\ t=0,\dots,T}\left|\ Err_{Q_{i+1}}(Q_i) - Err_{n,Q_{i+1}}(Q_i)\right| > \varepsilon'\right] \\
&\leq 4N_1\left(\frac{\varepsilon'}{32M'},\ F,\ 2n\right)\exp\left\{-\frac{(\varepsilon')^2 n}{32(M')^2}\right\}.
\end{aligned}
$$

Put $\varepsilon = (16L\ )^{(T+2)/2}\sqrt{\varepsilon'}$ to obtain the results of the theorem.

Suppose the Q functions are approximated by linear combinations of p features; for each $t = 0,\dots,T$, denote the feature vector by $q_t\ (\mathbf{o}_t, \mathbf{a}_t\ )$. The approximation space is then,

$$
\mathcal{Q}_t = \left\{Q_t(\boldsymbol{o}_t,\ \boldsymbol{a}_t) = \theta^T q_t(\boldsymbol{o}_t,\ \boldsymbol{a}_t)\ :\ \theta \in \Theta\right\}
$$

where Θ is a subset of $\mathbf{R}^p$. In this case, the batch Q-learning algorithm may be based on (4); we represent the performance of the functions $\{Q_0,\dots, Q_t\ \}$ on the training set by

$$\widetilde{Err}_{n,Q_{t+1}}(Q_t) = E_n\left[\left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)\right)q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)\right]$$

for $t = 0,...,T$ (recall E_n represents the expectation with respect to the probability obtained by choosing a trajectory uniformly from the training set). In this theorem

$$F' = \bigcup_{i=1}^{p} \bigcup_{t=1}^{T} \left\{\left(r_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}; \theta_{t+1}) - Q_t(\boldsymbol{o}_t, \boldsymbol{a}_t; \theta_t)\right)q_{ti}(\boldsymbol{o}_t, \boldsymbol{a}_t) : \theta_t, \theta_{t+1} \in \Theta\right\}.$$

Define the functions $\{\tilde{Q}_0, \ldots, \tilde{Q}_T\}$, and the policy, $\boldsymbol{\pi}$, as follows. First define $\tilde{Q}_T(\boldsymbol{O}_T, \boldsymbol{A}_T)$ to be the projection of $E[R_T | \mathbf{O}_T, \mathbf{A}_T]$ on the space spanned by q_T. Then set $\boldsymbol{\pi}_T(\boldsymbol{o}_T, \boldsymbol{a}_{T-1}) \in \arg\max_{a_T} \tilde{Q}_T(\boldsymbol{o}_T, \boldsymbol{a}_T)$. Next for $t = T - 1,...,0$, set $\tilde{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t)$as the projection of $E\left[R_t + \tilde{Q}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \boldsymbol{\pi}_{t+1}) \mid \boldsymbol{O}_t, \boldsymbol{A}_t\right]$on the space spanned by q_t (recall $\tilde{Q}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, \boldsymbol{\pi}_{t+1})$ is defined as $\tilde{Q}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t, a_{t+1})$ with a_{t+1} replaced by $\boldsymbol{\pi}_{t+1}(\boldsymbol{O}_{t+1}, \boldsymbol{A}_t)$). And set $\boldsymbol{\pi}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \tilde{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$. These projections are with respect to P, the distribution which generated the trajectories in the training set (the likelihood is in (1)).

Theorem 2

Suppose that there exists a positive constant, say L, for which $p_t(a_t | \mathbf{o}_t, \mathbf{a}_{t-1}) \geq L^{-1}$ for all $(\mathbf{o}_t, \mathbf{a}_{t-1})$, $0 \leq t \leq T$. Suppose that for each t, $x \in \mathbf{R}^p$, $x^T E q_t q_t^T x > \eta \|x\|^2$ where $\eta > 0$ ($\| \cdot \|$ is the Euclidean norm). Also assume that Θ is a closed subset of $\{x \in \mathbf{R}^p : \|x\| \leq M_\Theta\}$ and for all (t, i), the ith component in the vector q_t is pointwise bounded; $|q_{ti}| \leq M_Q$ for M_Q a constant. Then for $\varepsilon > 0$, with probability at least $1 - \delta$, over the random choice of the training set, every choice of functions, $Q_t \in \mathcal{Q}_t$ and functions $\tilde{Q}_t$, $t = 0, \ldots, T$ with associated policies defined by π with $\pi_t(\mathbf{o}_t, \mathbf{a}_{t-1} \in \arg\max_{at} Q_t(\mathbf{o}_t, \mathbf{a}_t)$ and $\boldsymbol{\pi}$ with $\boldsymbol{\pi}_t(\boldsymbol{o}_t, \boldsymbol{a}_{t-1}) \in \arg\max_{a_t} \tilde{Q}_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$ respectively, the following bounds are satisfied

$$\sum_{t=0}^{T} L^{(t+1)} E\left|\tilde{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)\right| \leq \sqrt{p} M_Q \Big/ \eta \sum_{t=0}^{T} L^{(t+1)} \sum_{j=t}^{T} (L p M_Q^2 / \eta)^{j-t} \| \widetilde{Err}_{n,Q_{j+1}}(Q_j) \|$$
$$+4\varepsilon.$$

for $t = 0,...,T$, where E represents the expectation with respect to the distribution (1) generating the training set and

$$\int \left| V_{\boldsymbol{\pi}}(o) - V_{\pi}(o) \right| dF(o) \leq 2M\sqrt{p} M_Q \Big/ \eta \sum_{t=0}^{T} L^{(t+1)} \sum_{j=t}^{T} (L p M_Q^2 / \eta)^{j-t} \| \widetilde{Err}_{n,Q_{j+1}}(Q_j) \|$$
$$+8M\varepsilon$$
$$+ 2M \sum_{t=0}^{T} L^{(t+1)} E\left|\tilde{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - \tilde{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t)\right|$$
$$+ 2M \sum_{t=0}^{T} L^{(t+1)} E\left|\tilde{Q}_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_{\boldsymbol{\pi},t}(\boldsymbol{O}_t, \boldsymbol{A}_t)\right|$$

for n larger than

$$\left(\frac{C}{\varepsilon}\right)^2 \log\left(\frac{B}{\delta}\right) \qquad (14)$$

where $C = 4\sqrt{2}M' p^{T+1/2} M_Q^{2T+1} \eta^{-(T+1)} L^{T+1}$, M' is a uniform upper bound on the absolute value on all $f \in F'$ and $B = \varepsilon^{-2p} 4^{6p+3} p^{2Tp+p+3} (T+1)^2 e^{2p+2} (M')^{4p} |A|^p M_Q^{(2T+1)2p} \eta^{-2p(T+1)} L^{2p(T+1)}$

Remarks:

1.Define $\hat{Q}_t$ as a zero of $\widetilde{Err}_{n,\hat{Q}_{t+1}}(Q_t)$, $t = T, T-1, \ldots, 0$ (recall that $\hat{Q}_{T+1}$ is identically zero). Suppose that $Q_t^* \in \mathcal{Q}_t$ for each t; in this case $\mathcal{Q}_t = Q_t^*$ for all t (we ignore sets of measure zero in this discussion). Then with probability greater than 1 – δ and $\boldsymbol{\pi} = \pi^*$, $\mathcal{Q}_t = Q_t^*$ we obtain

$$\int V^*(o) - V_{\hat{\pi}}(o) dF(o) \le 8M\varepsilon$$

for all n satisfying (14). Thus estimating each Q_t by solving $\widetilde{Err}_{n,Q_{t+1}}(Q_t) = 0$, $t = T, \ldots, 0$, yields a policy that consistently achieves the optimal value.

2. Again define $\hat{Q}_t$ as a zero of $\widetilde{Err}_{n,\hat{Q}_{t+1}}(Q_t)$, $t = T, T-1, \ldots, 0$.Given two $T+1$ vectors of functions $Q' = \{Q'_0, \ldots, Q'_T\}$ and $Q = \{Q_0, \ldots, Q_T\}$ define $\ell(Q', Q) = \sum_{t=0}^{T} L^{t+1} E|Q'_t(\boldsymbol{O}_t, \boldsymbol{A}_t) - Q_t(\boldsymbol{O}_t, \boldsymbol{A}_t)|$. Then the first result of Theorem 2 implies that $\ell(\mathcal{Q}, \hat{Q})$ converges in probability to zero. From Lemma 2 we have that $\int |V_{\boldsymbol{\pi}}(o) - V_{\pi}(o)| dF(o) \le 2M\ell(Q, \mathcal{Q}) + 2M\ell(Q_{\boldsymbol{\pi}}, \mathcal{Q})$ and thus $\int |V_{\boldsymbol{\pi}}(o) - V_{\hat{\pi}}(o)| dF(o)$is with high probability bounded above by $2M\ell(\mathcal{Q}, \mathcal{Q}) + 2M\ell(Q_{\boldsymbol{\pi}}, \mathcal{Q})$. Consequently the presence of the third and fourth terms in Theorem 2 is not surprising. It is unclear whether the "go-between" $\mathcal{Q}_t$ is necessary.

3. Recall the space of policies implied by the approximation spaces for the Q-functions is given by $\Pi_Q = \{\pi_\theta, \theta \in \Theta\}$ where $\pi_\theta = \{\pi_{1,\theta}, \ldots, \pi_{T,\theta}\}$ and where each $\pi_{t,\theta}(\mathbf{o}_t, \mathbf{a}_{t-1}) \in \arg\max_{a_t} Q_t(\mathbf{o}_t, \mathbf{a}_t; \theta)$ for some $Q_t \in \mathcal{Q}_t$. Suppose that $\max_{\pi \in \Pi_Q} \int V_\pi(o) dF(o)$ is achieved by some member of Π_Q and $\boldsymbol{\pi} \in \arg\max_{\pi \in \Pi_Q} \int V_\pi(o) dF(o)$. Ideally Q-learning would provide a policy that achieves the highest value as compared to other policies in Π_Q (as is the case with $\boldsymbol{\pi}$). This is not necessarily the case. As discussed in the above remark batch Q-learning yields estimated Q-functions for which $\ell(\mathcal{Q}, \hat{Q})$ converges to zero. The policy $\boldsymbol{\pi}$ may not produce a maximal value; that is $\int V_{\boldsymbol{\pi}}(o) - V_\pi(o) dF(o)$need not be zero (see also the remark following Lemma 2). Recall from Lemma 2 that $2M\ell(\mathcal{Q}, \mathcal{Q}) + 2M\ell(\mathcal{Q}, Q_{\boldsymbol{\pi}})$ is an upper bound on this difference. It is not hard to see that $\ell(\mathcal{Q}, Q_{\boldsymbol{\pi}})$ is zero if and only if $\boldsymbol{\pi}$ is the optimal policy; indeed the optimal Q-function would belong to the approximation space. The Q-learning algorithm does not directly maximize the value function. As remarked in Tsitsiklis and van Roy (1997) the goal of the Q-learning algorithm is to construct an approximation to the optimal Q-function within the constraints imposed by the app! roximation space; this approximation is a projection when the approximation space is linear. Approximating the Q-function yields an optimal policy if the approximating class is sufficiently rich. Ormoneit and Sen (2002) consider a sequence of approximation spaces (kernel based spaces indexed by a bandwidth) and make assumptions on the optimal value function which guarantee that this sequence of approximations spaces is sufficient rich (as the bandwidth decreases with increasing training set size) so as to approximate the optimal value function to any desired degree.

4. Again define $\hat{Q}_t$ as a zero of $\widetilde{Err}_{n,\hat{Q}_{t+1}}(Q_t),\ t = T,\ T-1,\ \dots,\ 0$. Since $\ell(\bar{Q},\ \hat{Q})$ converges in probability to zero, one might think that $\int \left| V_{\pi}(o) - V_{\hat{\pi}}(o) \right| dF(o)$ should be small as well. Referring to Lemma 1, we have that the difference in value functions $\int \left| V_{\pi}(o) - V_{\hat{\pi}}(o) \right| dF(o)$ can be expressed as the sum over t of the expectation of $Q_{\pi,t}(\boldsymbol{O}_t,\ \boldsymbol{A}_{t-1},\ \hat{\pi}_t) - Q_{\pi,t}(\boldsymbol{O}_t,\ \boldsymbol{A}_{t-1},\ \pi_t)$.However $\ell(\bar{Q},\ \hat{Q})$ small does not imply that $\hat{\pi}$ and π will be close nor does it imply that $Q_{\pi,t}(\boldsymbol{O}_t,\ \boldsymbol{A}_{t-1},\ \hat{\pi}_t) - Q_{\pi,t}(\boldsymbol{O}_t,\ \boldsymbol{A}_{t-1},\ \pi_t)$ will be small. To see the former consider an action space with 10 actions, 1,...,10 and $\hat{Q}_t(o_t,\ a_t) = 1$ for a = 1,...,9, $\hat{Q}_t(o_t,\ 10) = 1 + 1/2\varepsilon$ and $\bar{Q}_t(o_t,\ a_t) = 1 - 1/2\varepsilon$ for $a = 2,\ \dots,\ 10,\ \bar{Q}_t(o_t,\ 1) = 1$.So $\bar{Q}_t$ and $\hat{Q}_t$ are uniformly less than ε apart yet the argument of their maxima are 1 and 10.

Proof of Theorem 2—Fix $Q_t = \theta_t^T q_t$, $\theta \in \theta$ for $t = 0,\dots,T$. Define an infinite training sample version of $\widetilde{Err}_n$ as

$$
\begin{aligned}
\widetilde{Err}_{Q_{t+1}}(Q_t) &= E\left[\left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ a_{t+1}) - Q_t\right) q_t\right] \\
&= E\left[\left(\bar{Q}_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ a_{t+1}) - \bar{Q}_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ \pi_{t+1}) - Q_t\right) q_t\right]
\end{aligned}
$$

where Q_t is an abbreviation for Q_t ($\mathbf{O}_t$, $\mathbf{A}_t$). To derive the last equality recall that $\bar{Q}_t(\boldsymbol{O}_t,\ \boldsymbol{A}_t)$ is the projection of $E\left[R_t + \bar{Q}_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ \pi_{t+1}) \mid \boldsymbol{O}_t,\ \boldsymbol{A}_t\right]$ on the space spanned by q_t . Since $\bar{Q}_t$ is a projection we can write $Q_t = \theta_t^T q_t$ for some $\theta_{\pi,t} \in \Theta$. Also we can write $Q_t = \theta^T{}_t\, q_t$ for some $\theta_t \in \theta$. The $\widetilde{Err}$'s provide a pointwise upper bound on the differences, $\left| \bar{Q}_t - Q_t \right|$, as follows. Rearrange the terms in $\widetilde{Err}_{Q_{t+1}}$ using the fact that $Eq_t\, q_t^T$ is invertible to obtain

$$
\begin{aligned}
(\theta_{\pi,t} - \theta_t) &= (Eq_t q_t^T)^{-1}\, \widetilde{Err}_{Q_{t+1}}(Q_t) \\
&- (Eq_t q_t^T)^{-1} E\left[\left(\max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ a_{t+1}) - \bar{Q}_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ \pi_{t+1})\right) q_t\right].
\end{aligned}
$$

Denote the Euclidean norm of a p dimensional vector x by $||x||$. Then

$$
\begin{aligned}
\left| (\theta_{\pi,t} - \theta_t)^T q_t \right| &\le (1/\eta)\| \widetilde{Err}_{Q_{t+1}}(Q_t)\| \ \| q_t\| \ + \\
(1/\eta) & E\left[\left| \max_{a_{t+1}} Q_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ a_{t+1}) - \bar{Q}_{t+1}(\boldsymbol{O}_{t+1},\ \boldsymbol{A}_t,\ \pi_{t+1}) \right| \|q_t\| \right] \|q_t\| \\
&\le (1/\eta)\| \widetilde{Err}_{Q_{t+1}}(Q_t)\| \ \| q_t\| \ + (1/\eta) LE\left[\left| Q_{t+1} - \bar{Q}_{t+1} \right| \| q_t\| \right] \| q_t\| \\
&\le (1/\eta)\sqrt{p} M_Q \| \widetilde{Err}_{Q_{t+1}}(Q_t)\| \ + (1/\eta) LpM_Q^2 E\left[\left| Q_{t+1} - \bar{Q}_{t+1} \right|\right]
\end{aligned}
$$

for $t \le T$. To summarize

$$
E\left| \bar{Q}_t - Q_t \right| \le (1/\eta)\sqrt{p} M_Q \| \widetilde{Err}_{Q_{t+1}}(Q_t)\| \ + (1/\eta) LpM_Q^2 E\left[\left| Q_{t+1} - \bar{Q}_{t+1} \right|\right]
$$

where Q_t, $\bar{Q}_t$ is an abbreviation for Q_t ($\mathbf{O}_t$, $\mathbf{A}_t$), respectively $\bar{Q}_t(\boldsymbol{O}_t,\ \boldsymbol{A}_t)$, for each t.

As in the proof of Theorem 1, these inequalities can be solved for each $\mathrm{E}\left| \bar{Q}_t - Q_t \right|$ to yield

$$E\left|\bar{Q}_t - Q_t\right| \le \left(\sqrt{p}M_Q / \eta\right)\sum_{j=t}^{T}\left(LpM_Q^2 / \eta\right)^{j-t}\| \widetilde{Err}_{Q_{j+1}}(Q_j)\|$$

$$\le \left(\sqrt{p}M_Q / \eta\right)\sum_{j=t}^{T}\left(LpM_Q^2 / \eta\right)^{j-t}\| \widetilde{Err}_{n,Q_{j+1}}(Q_j) - \widetilde{Err}_{Q_{j+1}}(Q_j)\|$$

$$+\left(\sqrt{p}M_Q / \eta\right)\sum_{j=t}^{T}\left(LpM_Q^2 / \eta\right)^{j-t}\| \widetilde{Err}_{n,Q_{j+1}}(Q_j)\| \ .$$

Simplifying terms we obtain

$$\sum_{t=0}^{T} L^{(t+1)}E\left|\bar{Q}_t - Q_t\right| \le \sqrt{p}M_Q \Big/ \eta\sum_{t=0}^{T} L^{(t+1)}\sum_{j=t}^{T}\left(LpM_Q^2 / \eta\right)^{j-t}\| \widetilde{Err}_{n,Q_{j+1}}(Q_j)\|$$
$$+4p^{T+1/2}M_Q^{2T+1}\eta^{-(T+1)}L^{T+1}\max_t\| \widetilde{Err}_{n,Q_{t+1}}(Q_t) - \widetilde{Err}_{Q_{t+1}}(Q_t)\| \ . \qquad (15)$$

Consider each component of each of the T +1, p dimensional vectors, $\widetilde{Err}_{n,Q_{i+1}}(Q_i)) - \widetilde{Err}_{Q_{i+1}}(Q_i)$ for an ε′> 0:

$$P\left[\bigcup_{i=0}^{T}\bigcup_{j=1}^{p}\left\{\text{for some}_{\theta_i,\theta_{i+1}\in\Theta, q_i\in Q_i, q_{i+1}\in Q_{i+1}}\left|\widetilde{Err}_{n,Q_{i+1}}(Q_i)_j - \widetilde{Err}_{Q_{i+1}}(Q_i))_j\right| > \varepsilon'\right\}\right].$$

This probability is in turn bounded above by

$$\sum_{i=0}^{T}\sum_{j=1}^{p} P\left[\text{for some}_{\theta_i,\theta_{i+1}\in\Theta, q_i\in Q_i, q_{i+1}\in Q_{i+1}}\left|\widetilde{Err}_{n,Q_{i+1}}(Q_i)_j - \widetilde{Err}_{Q_{i+1}}(Q_i))_j\right| > \varepsilon'\right].$$

In Lemmas 17.2, 17.3, 17.5, Anthony and Bartlett (1999) provide an upper bound on the probability

$$P\left[\text{for some } f \in F \text{ has } \left|E_n(\ell_f) - E(\ell_f)\right| \ge \varepsilon'\right]$$

where $\ell_f(x, y) = (y - f(x))^2$. These same lemmas (based on the classical arguments of symmetrization, permutation and reduction to a finite set) can be used for $f \in$ F′ since the functions in F′ are uniformly bounded. Hence for each j = 1,..., p and t = 0,...,T

$$P\left[\text{for some } \theta_t,\ \theta_{t+1} \in \Theta,\ q_t \in Q_t,\ q_{t+1} \in Q_{t+1} \text{ has } \left|\widetilde{Err}_{n,Q_{t+1}}(Q_t)_j - \widetilde{Err}_{Q_{t+1}}(Q_t)_j\right| > \varepsilon'\right]$$
$$\le 4N_1\left(\frac{\varepsilon'}{16M'}, F', 2n\right)\exp\left\{-\frac{(\varepsilon')^2 n}{32(M')^2}\right\}.$$

Set ε $= p^{T+1/2}M^{2T+1}{}_{\mathrm{Q}}\eta^{-(T+1)}L^{T+1}$ε′. Thus for n satisfying

$$4p(T+1)N_1\left(\frac{\varepsilon}{16M' p^{T+1/2}M_Q^{2T+1}\eta^{-(T+1)}L^{T+1}}, F', 2n\right)$$
$$\exp\left\{-\frac{\varepsilon^2 n}{32(M')^2\left(p^{T+1/2}M_Q^{2T+1}\eta^{-(T+1)}L^{T+1}\right)^2}\right\} \le \delta, \qquad (16)$$

the first result of the theorem holds.

To simplify the constraint on *n*, we derive a covering number for F′ from covering numbers for the Q$_t$'s. Apply Lemma A2 part 1, to obtain

$$N_1(\varepsilon,\ V_{t+1,n}) \le N_1\left(\frac{\varepsilon}{|A|},\ Q_{t+1},\ |A|n\right)$$

for V$_{t+1}$ = {max$_{at+1}$Q$_{t+1}$(**o**$_{t+1}$, **a**$_{t+1}$) : Q$_{t+1}$ ∈ Q$_{t+1}$}. Next apply Lemma A2, parts 2 and 3, to obtain

$$N_1(\varepsilon,\ F',\ n) \le \sum_{t=0}^{T-1} N_1\left(\frac{\varepsilon}{2|A|M'},\ Q_{t+1},\ |A|n\right) N_1\left(\frac{\varepsilon}{2M'},\ Q_t,\ n\right)$$
$$+N_1\left(\frac{\varepsilon}{M'},\ Q_T,\ n\right).$$

Theorems 11.6 and 18.4 of Anthony and Bartlett imply that $N_1(\varepsilon,\ Q_t,\ n) \le e(p+1)\left(\frac{2e}{\varepsilon}\right)^p$ for each *t*. Combining this upper bound with (16) and simplifying the algebra yields (14).

Next Lemma 2 implies:

$$\int |V_{\pi}(o) - V_{\pi}(o)|\, dF(o) \le M \sum_{t=0}^{T} 2L^{(t+1)} E|Q_t - \tilde{Q}_t|$$
$$+M \sum_{t=0}^{T} 2L^{(t+1)} E|\tilde{Q}_t - \tilde{Q}_t| + M \sum_{t=0}^{T} 2L^{(t+1)} E|\tilde{Q}_t - Q_{\pi,t}|.$$

This combined with the first result of the theorem implies the second result.

6. Discussion

Planning problems involving a single training set of trajectories are not unusual and can be expected to increase due to the widespread use of policies in the social and behavioral/medical sciences (see, for example, Rush et al., 2003; Altfeld and Walker, 2001; Brooner, and Kidorf, 2002); at this time these policies are formulated using expert opinion, clinical experience and/or theoretical models. However there is growing interest in formulating these policies using empirical studies (training sets). These training sets are collected under fixed exploration policies and thus while they allow exploration they do not allow exploitation, that is, online choice of the actions. If subjects are recruited into the study at a much slower rate than the calendar duration of the horizon, then it is possible to permit some exploitation; some of this occurs in the field of cancer research (Thall, Sung and Estey, 2002).

This paper considers the use of Q-learning with dynamic programming and function approximation for this planning purpose. However the mismatch between Q-learning and the goal of learning a policy that maximizes the value function has serious consequences and emphasizes the need to use all available science in choosing the approximation space. Often the available behaviorial or psychosocial theories provide qualitative information concerning the importance of different observations. In addition these theories are often represented graphically via directed acyclic graphs. However information at the level of the form of the conditional distributions connecting the nodes in the graphs is mostly unavailable. Also due to the complexity of the problems there are often *unknown* missing common causes of different nodes in the graphs. See http://neuromancer.eecs.umich.edu/dtr for more information and references. Methods that can use this qualitative information to minimize t! he mismatch are needed.

Acknowledgements

We would like to acknowledge the help of the reviewers and of Min Qian in improving this paper. Support for this project was provided by the National Institutes of Health (NIDA grants K02 DA15674 and P50 DA10075 to the Methodology Center).

References

Altfeld M, Walker BD. Less is more? STI in acute and chronic HIV-1 infection. Nature Medicine 2001;7:881–884.

M. Anthony and P. L. Bartlett. Neural Network Learning: Theoretical Foundations. Cambridge, UK: Cambridge University Press, 1999.

L. Baird. Advantage updating. Technical Report. WL-TR-93-1146, Wright-Patterson Air Force Base, 1993.

Baxter J, Bartlett PL. Infinite-horizon policy-gradient estimation. J Artificial Intelligence Research 2001;15:319–350.

R. E. Bellman Dynamic Programming. Princeton: Princeton University Press, 1957.

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

Brooner RK, Kidorf M. Using behavioral reinforcement to improve methadone treatment participation. Science and Practice Perspectives 2002;1:38–48.

Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, Quitkin FM, Wisniewski S, Lavori PW, Rosenbaum JF, Kupfer DJ. Background and rationale for the sequenced treatment alternative to relieve depression (STAR*D) study. Psychiatric Clinics of North America 2003;26(3): 457–494. [PubMed: 12778843]

C. N. Fiechter Efficient Reinforcement Learning. In Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory (COLT 1994), pages 88–97, New Brunswick, NJ, 1994.

C. N. Fiechter. Expected Mistake Bound Model for On-Line Reinforcement Learning. In Proceedings of the Fourteenth International Conference on Machine Learning, Douglas H. Fisher (Ed.), Nashville, Tennessee, pages 116–124, 1997.

S. M. Kakade. On the Sample Complexity of Reinforcement Learning. Ph.D. thesis, University College, London, 2003.

Kearns M, Mansour Y, Ng AY. A Sparse sampling algorithm for near-optimal planning in large Markov decision processes. Machine Learning 1999;49(2–3):193–208.

M. Kearns, Y. Mansour and A. Y. Ng. Approximate planning in large POMDPs via reusable trajectories. In Advances in Neural Information Processing Systems, 12, MIT Press, 2000.

Ormoneit D, Sen S. Kernel-Based Reinforcement Learning. Machine Learning 2002;49(2–3):161–178.

L. Peshkin and C. R. Shelton. Learning from scarce experience. In Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002) Claude Sammut, Achim G. Hoffmann (Eds.) pages 498–505, Sydney, Australia, 2002.

Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ, Trivedi MH, Suppes T, Miller AL, Biggs MM, Shores-Wilson K, Witte BP, Shon SP, Rago WV, Altshuler KZ. TMAP Research Group. Texas medication algorithm project, phase 3 (TMAP-3): Rationale and study design. Journal of Clinical Psychiatry 2003;64(4):357–69. [PubMed: 12716235]

Schneider LS, Tariot PN, Lyketsos CG, Dagerman KS, Davis KL, Davis S, Hsiao JK, Jeste DV, Katz IR, Olin JT, Pollock BG, Rabins PV, Rosenheck RA, Small GW, Lebowitz B, Lieberman JA. National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE). American Journal of Geriatric Psychiatry 2001;9(4):346–360. [PubMed: 11739062]

Schapire RE, Bartlett P, Freund Y, Lee WS. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 1998;26(5):1651–1686.

D. I. Simester, P. Sun, and J. N. Tsitsiklis. Dynamic catalog mailing policies. unpublished manuscript, Available electronically at http://web.mit.edu/jnt/www/Papers/P-03-sun-catalog-rev2.pdf, 2003.

R. S. Sutton and A. G. Barto. Reinforcement Learning: An Introduction. The MIT Press, Cambridge, Mass, 1998.

Thall PF, Millikan RE, Sung HG. Evaluating multiple treatment courses in clinical trials. Statistics and Medicine 2000;19:1011–1028.

Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. Journal of the American Statistical Association 2002;97:29–39.

Tsitsiklis JN, Van Roy B. Feature-based methods for large scale dynamic programming. Machine Learning 1996;22:59–94.

Tsitsiklis JN, Van Roy B. An analysis of temporal-difference learning with function approximation. IEEE Transactions on Automatic Control 1997;42(5):674–690.

A. W. van der Vaart and J. A. Wellner. Weak Convergence and Empirical Processes. Springer, New York, 1996.

C. J. C. H. Watkins. Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, 1989.

Appendix A

Recall that the distributions, P and P_π differ only with regards to the policy (see (1) and (2)). Thus the following result is unsurprising. Let $f(\mathbf{O}_{T+1}, \mathbf{A}_T)$ be a (measurable) nonnegative function; then $E_\pi f$ can be expressed in terms of an expectation with respect to the distribution P if we assume that $p_t(a_t|\mathbf{o}_t, \mathbf{a}_{t-1}) > 0$ for each $(\mathbf{o}_t, \mathbf{a}_t)$ pair and each t. The presence of the p_js in denominator below represent the price we pay because we only have access to training trajectories with distribution P; we do not have access to trajectories from distribution P_π.

Lemma A1 Assume that $P_\pi[p_0(A_0|S_0) > 0] = 1$ and $P_\pi[p_t(A_t|\mathbf{O}_t, \mathbf{A}_{t-1}) > 0] = 1$ for $t = 1,...,T$. For any (measurable) nonnegative function of $g(\mathbf{O}_t, \mathbf{A}_t)$, the P-probability that

$$E_\pi\left[g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, S_0\right] = E\left[\left(\prod_{\ell=0}^{t} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell, \boldsymbol{A}_{\ell-1})}\right) g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, S_0\right]$$

is one for $t = 0,...,T$.

Proof: We need only prove that

$$E\left[h(S_0) E_\pi\left[g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, S_0\right]\right] = E\left[h(S_0) E\left[\left(\prod_{\ell=0}^{t} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell, \boldsymbol{A}_{\ell-1})}\right) g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\middle|\, S_0\right]\right]$$

for any (measurable) nonnegative function, h. Consider the two likelihoods ((1) and (2)) for a trajectory up to time t. Denote the dominating measure for the two likelihoods for the trajectory up to time t as λ_t. By assumption,

$$\int h(s_0) g(\boldsymbol{o}_t, \boldsymbol{a}_t) \left(\prod_{\ell=0}^{T} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(a_\ell \mid \boldsymbol{o}_\ell, \boldsymbol{a}_{\ell-1})}\right) f_0(s_0) p_0(a_0 \mid s_0)$$

$$\prod_{j=1}^{t} f_j(s_j \mid \boldsymbol{o}_{j-1}, \boldsymbol{a}_{j-1}) p_j(a_j \mid \boldsymbol{o}_j, \boldsymbol{a}_{j-1}) d\lambda_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$$

$$= \int h(s_0) g(\boldsymbol{o}_t, \boldsymbol{a}_t) f_0(s_0) 1_{a_0=\pi_0(s_0)} \prod_{j=1}^{t} f_j(s_j \mid \boldsymbol{o}_{j-1}, \boldsymbol{a}_{j-1}) 1_{a_j=\pi_j(\boldsymbol{o}_j, \boldsymbol{a}_{j-1})} d\lambda_t(\boldsymbol{o}_t, \boldsymbol{a}_t).$$

By definition the left hand side is $E\left[h(S_0) g(\boldsymbol{O}_t, \boldsymbol{A}_t)\left(\prod_{\ell=0}^{j} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell, \boldsymbol{A}_{\ell-1})}\right)\right]$ and the right hand side is $E_\pi[h(S_0)g(\mathbf{O}_t, \mathbf{A}_t)]$. Expressing both sides as the expectation of a conditional expectation, we obtain,

$$E_\pi[h(S_0)E_\pi[g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\Big|\, S_0]] = E\left[h(S_0)E\left[g(\boldsymbol{O}_t, \boldsymbol{A}_t)\left(\prod_{\ell=0}^{t} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{O}_\ell, \boldsymbol{A}_{\ell-1})}\right) \mid S_0\right]\right].$$

Note that the distribution of S_0 is the same regardless of how the actions are chosen, that is the distribution of S_0 is the same under both P and P_π. Thus

$$E[h(S_0)E_\pi[g(\boldsymbol{O}_t, \boldsymbol{A}_t) \,\Big|\, S_0]] = E\left[h(S_0)E\left[g(\boldsymbol{O}_t, \boldsymbol{A}_t)\left(\prod_{\ell=0}^{t} \frac{1_{A_\ell=\pi_\ell}}{p_\ell(A_\ell \mid \boldsymbol{A}_\ell, \boldsymbol{A}_{\ell-1})}\right) \mid S_0\right]\right].$$

Lemma A2 For p, q, r, s, N positive integers and M_F, M_G, M_Θ positive reals, define the following classes of real valued functions,

$$\begin{aligned}
H &\subseteq \left\{h(x, a) : x \in \mathbb{R}^p,\ a \in \{1, \ldots, N\}\right\} \\
F &\subseteq \left\{f(x) : x \in \mathbb{R}^q, \sup_x |f(x)| \le M_F\right\} \\
G &\subseteq \left\{g(x, y) : x \in \mathbb{R}^q,\ y \in \mathbb{R}^r \sup_{x,y} |g(x, y)| \le M_G\right\}
\end{aligned}$$

and

$$\Theta \subseteq \left\{\theta \in \mathbb{R}^s : \max_{i=1,\ldots,s} |\theta_i| \le M_\Theta\right\}.$$

The following hold.

1. If V = {$\max_a h(x, a) : h \in$ H}then $N_1(\varepsilon, V\, n) \le N_1(\varepsilon/N, H\, Nn)$.

2. For $|a| \le 1$, $|b| \le 1$, if V = {$a\,f(x) + bg(x, y) : f \in$ F, $g \in$ G} then $N_1(\varepsilon, V, n) \le N_1(\varepsilon/2, F, n)\, N_1(\varepsilon/2, G, n)$.

3. If V = F ∪ G then $N_1(\varepsilon, V, n) \le N_1(\varepsilon, F, n) + N_1(\varepsilon, G, n)$.

4. If V = {$\theta_1 f_1(x) + \ldots + \theta_s f_s(x) : f_i \in$ F, $(\theta_1, \ldots, \theta_s) \in \Theta$} then $N_1(\varepsilon, V, n) \le e(s+1)\left(\frac{4esM_\Theta M_F}{\varepsilon}\right)^s N_\infty\left(\frac{\varepsilon}{4sM_\Theta}, F, n\right)^s.$

Proof. We prove 1. and 4.; the proofs of 2. and 3. are straightforward and are omitted. Consider 1. Given $(x_1, \ldots, x_n)$, the ε-covering number for the class of points in $\mathbb{R}^{Nn}$, $\{(h(x_i, a) : i = 1, \ldots, n, a = 1, \ldots N); h \in H\}$ is bounded above by $N_1(\varepsilon, H, Nn)$. Note that for $(z_{ia}, i = 1, \ldots, n, a = 1, \ldots, N)$,

$$\begin{aligned}
1\Big/n \sum_{i=1}^{n} \left| \max_{a=1,\ldots,N} h(x_i, a) - \max_{a=1,\ldots,N} z_{ia} \right| &\le 1\Big/n \sum_{i=1}^{n} \max_{a=1,\ldots,N} |h(x_i, a) - z_{ia}| \\
&\le 1\Big/n \sum_{i=1}^{n} \sum_{a=1}^{N} |h(x_i, a) - z_{ia}|.
\end{aligned}$$

Thus the ε-covering number for the class of points in $\mathbb{R}^n$, $\mathbb{R}^n, \left\{(\max_{a=1}^{N} h(x_i, a) : i = 1, \ldots, n);\ h \in H\right\}$ is bounded above by $N_1(\varepsilon, H, Nn)$. Using the definition of covering numbers for classes of functions we obtain $N_1(\varepsilon, V, n) \le N_1\left(\frac{\varepsilon}{N}, H, Nn\right).$

Next consider 4. Put $x = (x_1, \ldots, x_n)$ (each $x_i \in \mathbb{R}^q$) and $f(x_i) = (f_1(x_i), \ldots, f_s(x_i))^T$. Then there exists $\{z_1, \ldots, z_N\}$, ($N = N_\infty(\varepsilon/(4sM_\Theta), F, n)$; $z_j \infty \mathbb{R}^n$) that form the centers of an $\varepsilon/(4sM_\Theta)$-cover for F. To each z_j we can associate an $f \in$ F, say f^*_j so that $\{f^*_1, \ldots, f^*_N\}$ form the centers of an $\varepsilon/(2sM_\Theta)$-cover for F. Then given $\{f_1, \ldots, f_s\} \in$ F there exists $j^* \in \{1, \ldots, N\}$ for $j = 1, \ldots, s$, so that $\max_{1 \le j \le s} \max_{1 \le i \le n} |f_j(x_i) - f^*_{j^*}(x_i)| \le \varepsilon/(2sM_\Theta)$. Then

$$(1 \big/ n) \sum_{i=1}^{n} \Big| \sum_{j=1}^{s} \theta_j f_j(x_i) - \theta_j f^*_{j^*}(x_i) \Big| \le \varepsilon / 2.$$

Define $F' = \left\{ \sum_{j=1}^{s} \theta_j f^*_{j^*} : \theta_j \in \Theta \right\}$. Theorems 11.6 and 18.4 of Anthony and Bartlett (1996) imply that $N_1(\varepsilon/2, F', n) \le e(s+1) \left(\frac{4esM_\Theta M_F}{\varepsilon} \right)^s$ These two combine to yield the result.