Action Time Sharing Policies for Ergodic Control of Markov Chains

Amarjit Budhiraja*, Xin Liu and Adam Shwartz†

August 30, 2011

Abstract: Ergodic control for discrete time controlled Markov chains with a locally compact state space and a compact action space is considered under suitable stability, irreducibility and Feller continuity conditions. A flexible family of controls, called *action time sharing* (ATS) policies, associated with a given continuous stationary Markov control, is introduced. It is shown that the long term average cost for such a control policy, for a broad range of one stage cost functions, is the same as that for the associated stationary Markov policy. In addition, ATS policies are well suited for a range of estimation, information collection and adaptive control goals. To illustrate the possibilities we present two examples: The first demonstrates a construction of an ATS policy that leads to consistent estimators for unknown model parameters while producing the desired long term average cost value. The second example considers a setting where the target stationary Markov control q is not known but there are sampling schemes available that allow for consistent estimation of q. We construct an ATS policy which uses dynamic estimators for q for control decisions and show that the associated cost coincides with that for the unknown Markov control q.

AMS 2000 subject classifications: 90C40, 60K15

Keywords: Markov decision processes, Controlled Markov processes, Adaptive control, Ergodic control, Action time sharing policies, long time average cost

1. Introduction

Markov Decision processes are used extensively as the simplest models that involve both stochastic behavior and control [11]. A common measure of performance is the long-time average (or ergodic) criterion. Given all relevant parameters, a typical goal is to find a simple (e.g. feedback, or deterministic stationary) policy that achieves the optimal value.

The goal of adaptive control is to obtain an optimal policy, when some relevant information

*Research supported in part by the Army Research Office (Grants W911NF-0-1-0080 and W911NF-10-1-0158), National Science Foundation (DMS-1004418 and DMS-1016441), and the US-Israel Binational Science Foundation (Grant 2008466).

†Research supported in part by the US-Israel Binational Science Foundation (Grant 2008466).

concerning the behavior of the system is missing. The relevant information needs to be obtained while controls are chosen at each step. The classical approach is to design an algorithm which collects information, while at the same time choosing controls, in such a way that sufficient information is collected for making good control decisions, in the sense that the chosen controls "approach optimality over time." Existing results include general solutions for the case of countable state space, and specify an estimation and a control scheme (see [4, 10] and references therein). For a more refined criterion of optimality for the adaptive case see [1, 5, 12]. A different approach to this issue, including PAC (Probably Approximately Correct) criteria, can be found in the large literature on Reinforcement learning, e.g. [6]. For results on adaptive control in the non-countable setting we refer the reader to [7–9] and references therein: these deal with the classical setup, namely they seek a combined estimation and control scheme and consider parameterized models.

We are concerned with a more elementary question, namely: What are the basic controlled objects that determine the cost? Since the objective function (see (2.2)) is defined as a Cesaro limit, we can expect that a similar Cesaro definition of the choice of controls would suffice to determine the cost. Indeed, [2] shows the following, for the case of countable state and action spaces. Let q be a stationary Markov control, namely it is a map from the state space $\mathbb{X}$ to the space $\mathcal{P}(\mathbb{A})$ of probability measures on the action space $\mathbb{A}$. Together with an initial distribution μ on $\mathbb{X}$ and a transition probability kernel $\mathcal{Q} : \mathbb{X} \times \mathbb{A} \times \mathcal{B}(\mathbb{X}) \to [0, 1]$, such a Markov control determines a probability measure $\mathbb{P}^q_\mu$ on the infinite product space $\Omega = (\mathbb{X} \times \mathbb{A})^{\otimes\infty}$ by the relation

$$\begin{aligned}&\mathbb{P}^q_\mu\left((X_0, A_0) \in E_0, (X_1, A_1) \in E_1, \cdots, (X_k, A_k) \in E_k\right)\\ &= \int_{E_0}\int_{E_1}\cdots\int_{E_k} q(x_k, da_k)\mathcal{Q}(x_{k-1}, a_{k-1}, dx_k)\cdots q(x_1, da_1)\mathcal{Q}(x_0, a_0, dx_1)q(x_0, da_0)\mu(dx_0),\\ &\qquad\qquad E_0, E_1, \cdots, E_k \in \mathcal{B}(\mathbb{X} \times \mathbb{A}),\ k \in \mathbb{N}_0,\end{aligned}$$

where $(X_k, A_k)_{k\in\mathbb{N}_0}$ is the canonical coordinate sequence on Ω. Defining a general admissible control policy requires additional notation and thus a precise description is postponed to Section 2. Roughly speaking, such a policy is defined in terms of a non-anticipative sequence $\{\pi_t\}_{t\in\mathbb{N}_0}$ of $\mathcal{P}(\mathbb{A})$ valued random variables and, through a formula similar to the above display, describes a probability measure $\mathbb{P}^\pi_\mu$ on Ω. In the setting of countable state and action spaces, an admissible control policy π is called an ATS policy for a stationary Markov control q if the conditional frequencies:

$$f_T(a \mid x) = \frac{\sum_{t=0}^{T-1} 1\{X_t = x, A_t = a\}}{\sum_{t=0}^{T-1} 1\{X_t = x\}} \to q(x)(a) \equiv q(a \mid x), \text{ for all } (x, a) \in \mathbb{X} \times \mathbb{A},\ \mathbb{P}^\pi_\mu, a.e. \tag{1.1}$$

The paper [2] shows that for such a π, for any bounded one stage cost function, the costs (2.2) under $\mathbb{P}^q_\mu$ and under $\mathbb{P}^\pi_\mu$ are the same. Such a result says that the control decisions can deviate from those dictated by the Markov policy q, and still produce the same long term average cost, as long as the conditional frequencies converge to the correct values. This flexibility is useful in many situations, some of which will be described towards the end of this Introduction.

In the current work we are concerned with a setting where the state and action spaces are

not (necessarily) countable. Our main objective is to formulate an appropriate definition for an ATS policy which, similar to the countable case, on the one hand leads to long term costs that are identical to those for the corresponding Markov control, while on the other hand allows for flexible implementation well suited for various estimation and adaptive control goals. Clearly, conditional frequencies of the form in (1.1) are not suitable when $q(x,\cdot)$ and $\mathcal{Q}((x,a),\cdot)$ are not discrete measures. In Section 3 (Definition 3.1) we propose a definition of an ATS policy given in terms of suitable conditional frequencies over a sequence of "converging partitions" of the state space $\mathbb{X}$. We show in Theorem 3.1 that, under suitable stability, irreducibility and Feller continuity conditions (Assumptions 2.1, 2.2 and 2.3) occupation measures for state and action sequences, under an ATS policy given as in Definition 3.1, converge a.s. to the same (deterministic) measure as under the corresponding Markov control. Such a result in particular shows that long term costs for a broad family of one stage cost functions, under the two control policies, coincide.

The rest of the paper is organized as follows. In Section 2 we begin with some preliminary definitions and the main assumptions on the controlled dynamics. Section 3 introduces the definition of an ATS policy through a sequence of "converging partitions" of the state space. The section also presents the main convergence result for occupation measures associated with an ATS policy. In Section 4 we describe how ATS policies can be constructed and used in settings with incomplete model information. Finally in Section 5 we illustrate the advantage of using a flexible family of policies, through an example.

2. Definitions and Assumptions.

The following notation will be used. For two measurable spaces $(\Omega_1,\mathcal{F}_1)$ and $(\Omega_2,\mathcal{F}_2)$, the space of $\mathcal{F}_1/\mathcal{F}_2$ measurable maps from Ω_1 to Ω_2 will be denoted as $\mathcal{M}(\Omega_1,\mathcal{F}_1:\Omega_2,\mathcal{F}_2)$. When $(\Omega_2,\mathcal{F}_2)=(\mathbb{R},\mathcal{B}(\mathbb{R}))$, we will merely write $\mathcal{M}(\Omega_1,\mathcal{F}_1)$ and if $\mathcal{F}_1,\mathcal{F}_2$ are clear from the context, we will write $\mathcal{M}(\Omega_1:\Omega_2)$ and $\mathcal{M}(\Omega_1)$, respectively. The space of all probability measures on a measurable space $(\Omega,\mathcal{F})$ will be denoted by $\mathcal{P}(\Omega,\mathcal{F})$ or $\mathcal{P}(\Omega)$, when clear from the context. Borel sigma fields on a metric space $\mathcal{T}$ will be denoted by $\mathcal{B}(\mathcal{T})$. If $(\Omega,\mathcal{F})=(\mathcal{T},\mathcal{B}(\mathcal{T}))$ for some complete and separable metric (Polish) space $\mathcal{T}$, we will endow $\mathcal{P}(\Omega)\equiv\mathcal{P}(\mathcal{T})$ with the topology of weak convergence. We recall the definition of Bounded-Lipschitz norm on $\mathcal{P}(\mathcal{T})$ for a Polish space $\mathcal{T}$. Let

$$\mathcal{C}_1(\mathcal{T})=\left\{\psi:\mathcal{T}\to\mathbb{R}:\sup_{t,t'\in\mathcal{T},t\neq t'}\left(|\psi(t)|+\frac{|\psi(t)-\psi(t')|}{d(t,t')}\right)\leq 1\right\},$$

where d is the metric given on $\mathcal{T}$. For $\nu_1,\nu_2\in\mathcal{P}(\mathcal{T})$ denote

$$\|\nu_1-\nu_2\|_{\mathrm{BL}}=\sup_{\psi\in\mathcal{C}_1(\mathcal{T})}\left|\int\psi d\nu_1-\int\psi d\nu_2\right|.$$

This norm metrizes the topology of weak convergence making $\mathcal{P}(\mathcal{T})$ a Polish space. Throughout we will consider $\mathcal{P}(\mathcal{T})$ with this metric. The class of real valued continuous and bounded

functions on a metric space $\mathcal{T}$ will be denoted by $C_b(\mathcal{T})$. $C_{\rm buc}(\mathcal{T})$ will denote the subset of $C_b(\mathcal{T})$ consisting of all uniformly continuous functions. A class $\mathcal{S} \subset C_b(\mathcal{T})$ is called separating in $(\mathcal{T}, \mathcal{B}(\mathcal{T}))$ if whenever $\mu, \nu \in \mathcal{P}(\mathcal{T})$ and $\int f d\mu = \int f d\nu$ for all $f \in \mathcal{S}$, then $\mu = \nu$. Since $\mathcal{T}$ is Polish, one can find a countable collection in $C_{\rm buc}(\mathcal{T})$ that is separating and we shall use the notation $\mathcal{S}(\mathcal{T})$ to denote such a class. It is easy to check that if $\mathcal{T}_1, \mathcal{T}_2$ are Polish spaces then $\{f \otimes g : f \in \mathcal{S}(\mathcal{T}_1), g \in \mathcal{S}(\mathcal{T}_2)\}$ is separating in $(\mathcal{T}_1 \times \mathcal{T}_2, \mathcal{B}(\mathcal{T}_1) \otimes \mathcal{B}(\mathcal{T}_2))$. Given a subset C of a metric space $\mathcal{T}$ with a distance d, we define $\mathrm{diam}(C) = \sup\{d(x, y) : x, y \in C\}$.

We will consider a controlled stochastic dynamical system in discrete time (i.e. parametrized by the discrete index set $\mathbb{N}_0 \doteq \{0, 1, 2, \ldots\}$) with state space $\mathbb{X}$ that is a complete and separable locally compact space. A Polish space $\mathbb{A}$ will represent the control (or action) space. For each $x \in \mathbb{X}$ we are given a compact set $\mathbb{U}(x) \subset \mathbb{A}$ representing the set of admissible actions when the system is in state $x \in \mathbb{X}$. We assume that $\mathbb{K} = \{(x, a) : x \in \mathbb{X}, a \in \mathbb{U}(x)\}$ is a measurable subset of $\mathbb{X} \times \mathbb{A}$. The dynamics of the controlled Markov chain is described in terms of a transition kernel

$$\mathcal{Q} : \mathbb{K} \times \mathcal{B}(\mathbb{X}) \to [0, 1]$$

satisfying:
(i) For all $(x, a) \in \mathbb{K}$, $\mathcal{Q}((x, a), \cdot) \equiv \mathcal{Q}(\cdot \mid (x, a))$ is in $\mathcal{P}(\mathbb{X})$ and;
(ii) for every $C \in \mathcal{B}(\mathbb{X})$, $\mathcal{Q}(\cdot, C) \in \mathcal{M}(\mathbb{K})$.

Roughly speaking, denoting the state and control processes by $(X_t)_{t\in\mathbb{N}_0}, (A_t)_{t\in\mathbb{N}_0}$, respectively, $\mathcal{Q}(C \mid (x, a))$ represents the conditional probability of $\{X_1 \in C\}$ given that $\{X_0 = x, A_0 = a\}$. A convenient way to give a precise formulation of the controlled system is through canonical sample spaces (cf. [3]), as follows. Let $\Omega = (\mathbb{X} \times \mathbb{A})^{\otimes\infty}$ and denote by $\mathcal{F}$ the Borel σ field on Ω corresponding to the product topology. Define sequences $\{X_t\}_{t\in\mathbb{N}_0}$, $\{A_t\}_{t\in\mathbb{N}_0}$ of $\mathbb{X}$ and $\mathbb{A}$ valued measurable maps, respectively, on $(\Omega, \mathcal{F})$ as follows:

$$X_t(\omega) = x_t;\ A_t(\omega) = a_t,\ \text{where } \omega = (x_0, a_0, \cdots x_t, a_t, \cdots),\ t \in \mathbb{N}_0.$$

We also introduce the sequence of *History maps*, $\{H_t\}_{t\in\mathbb{N}_0}$, $H_t : \Omega \to \mathbb{H}_t$, where

$$\mathbb{H}_t = (\mathbb{X} \times \mathbb{A})^{\otimes(t-1)} \times \mathbb{X},\ t \in \mathbb{N};\ \mathbb{H}_0 = \mathbb{X}$$

as $H_t(\omega) = (x_0, a_0, \cdots x_{t-1}, a_{t-1}, x_t)$. Let

$$\bar{\mathcal{H}}_t = (\mathcal{B}(\mathbb{X} \times \mathbb{A}))^{\otimes(t-1)} \otimes \mathcal{B}(\mathbb{X}),\ \text{and } \mathcal{H}_t = \sigma(H_t) = H_t^{-1}(\bar{\mathcal{H}}_t).$$

Note that $\mathcal{F} = \bigvee_{t=0}^{\infty} \mathcal{H}_t$.

By a controlled system we will mean a probability measure on $(\Omega, \mathcal{F})$ that is described in terms of an admissible control policy which is defined as follows.

Definition 2.1 (Admissible Control Policy)**.** *A sequence $\pi = \{\pi_t\}_{t\in\mathbb{N}_0}$ of kernels, $\pi_t : \mathbb{H}_t \times \mathcal{B}(\mathbb{A}) \to [0, 1]$ satisfying for all $t \in \mathbb{N}_0$:*

(i) $\pi_t(h_t, \cdot) \equiv \pi_t(\cdot \mid h_t)$ is in $\mathcal{P}(\mathbb{A})$, for all $h_t \in \mathbb{H}_t$;

(ii) $\pi_t(\cdot, D) \in \mathcal{M}(\mathbb{H}_t, \bar{\mathcal{H}}_t)$, *for all* $D \in \mathcal{B}(\mathbb{A})$;
(iii) $\pi_t(h_t, \mathbb{U}(x_t)) = 1$, *for all* $h_t = (x_0, a_0, \cdots x_{t-1}, a_{t-1}, x_t) \in \mathbb{H}_t$,

is called an admissible (control) policy.

The set of all admissible policies is denoted by Π. Given $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \in \Pi$, there is a unique probability measure $\mathbb{P}^\pi_\mu$ on $(\Omega, \mathcal{F})$ satisfying:

- $\mathbb{P}^\pi_\mu(X_0 \in C) = \mu(C)$, $C \in \mathcal{B}(\mathbb{X})$,
- $\mathbb{P}^\pi_\mu(A_t \in D \mid \mathcal{H}_t)(\omega) = \pi_t(D \mid H_t(\omega))$, $\mathbb{P}^\pi_\mu$ a.s.,
- $\mathbb{P}^\pi_\mu((X_t(\omega), A_t(\omega)) \in \mathbb{K}) = 1$ for all $t \in \mathbb{N}_0$.
- $\mathbb{P}^\pi_\mu(X_{t+1} \in C \mid H_t, A_t)(\omega) = \mathcal{Q}(C \mid X_t(\omega), A_t(\omega))$, $\mathbb{P}^\pi_\mu$ a.s.

The measure $\mathbb{P}^\pi_\mu$ represents a controlled system with initial distribution μ and an admissible control policy $\pi \in \Pi$. The corresponding expectation operator will be denoted by $\mathbb{E}^\pi_\mu$. If $\mu = \delta_x$, we will write $\mathbb{P}^\pi_\mu$ and $\mathbb{E}^\pi_\mu$ as $\mathbb{P}^\pi_x$ and $\mathbb{E}^\pi_x$, respectively.

A family of admissible policies that are particularly useful are the so-called *stationary Markov policies*. These correspond to those $\pi \in \Pi$ for which there is a measurable map $q : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ such that $\pi_t(h_t, \cdot) = q(x_t)(\cdot)$ for every $h_t = (x_0, a_0, \cdots x_{t-1}, a_{t-1}, x_t) \in \mathbb{H}_t$. The class of all such policies is denoted by Π_{SM} and frequently we will identify a policy $\pi \in \Pi_{\text{SM}}$ with the associated map q. Note that for every $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \equiv q \in \Pi_{\text{SM}}$, $(X_t)_{t\in\mathbb{N}_0}$ is a Markov chain under $\mathbb{P}^\pi_\mu$ with transition probability kernel

$$\varrho_q(x, C) = \int_{\mathbb{A}} \mathcal{Q}((x, a), C) q(x, da), \ (x, C) \in \mathbb{X} \times \mathcal{B}(\mathbb{X}). \tag{2.1}$$

If $q \in \Pi_{\text{SM}}$ is such that the map $x \mapsto q(x)$ is continuous (from $\mathbb{X}$ to $\mathcal{P}(\mathbb{A})$), we will refer to q as a *continuous stationary Markov policy* and denote the class of all such policies by Π_{SMC}. Occasionally, for $x \in \mathbb{X}$, we will write $q(x)(\cdot)$ as $q(\cdot \mid x)$.

The next step in the formulation of a control problem is the introduction of the cost function that one will like to optimize. Here we are interested in a criterion that is designed for system optimization over a long time horizon. This criterion – usually referred to as the pathwise cost per unit time, or long time average cost – is given in terms of a measurable map $c : \mathbb{K} \to \mathbb{R}_+$, called the *one stage cost function*, as

$$J_S = \limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t), \tag{2.2}$$

where the right side above is a $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ valued random variable on $(\Omega, \mathcal{F})$. Under suitable conditions one can show that there is a $\pi^* \in \Pi$ and $V \in [0, \infty)$ such that, for all $\mu \in \mathcal{P}(X)$, $\mathbb{P}^{\pi^*}_\mu(J_S = V) = 1$ and for all $\pi \in \Pi$, $\mathbb{P}^\pi_\mu(J_S \geq V) = 1$. Such a π^* is then an optimal control policy for the problem. One typically finds that π^* can be taken to be an element of Π_{SM} (i.e.

a stationary Markov policy). For precise conditions under which the above statements hold we refer the reader to Section 6 of [3]. In this work we are not interested in the optimization of a particular one stage cost function but rather in the study of control policies that perform well over a broad family of cost functions. In that regard the following occupation measure plays a key role.

For $N \in \mathbb{N}$, define a $\mathcal{P}(\mathbb{X} \times \mathbb{A})$ valued random variable, Φ_N as

$$\Phi_N(\omega)(F) = \frac{1}{N} \sum_{t=0}^{N-1} 1_F(X_t(\omega), A_t(\omega)),\ F \in \mathcal{B}(\mathbb{X} \times \mathbb{A}),\ \omega \in \Omega.$$

We will make the following assumptions. The first two can be regarded as blanket stability conditions while the third is the weak Feller property. We also provide an example satisfying all these assumptions.

Assumption 2.1. *For each $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \in \Pi$, the sequence of probability measures $\{\Phi_N(\omega), N \in \mathbb{N}\}$ is tight, for $\mathbb{P}^\pi_\mu$ a.e. ω.*

If $\mathbb{X}$ and $\mathbb{A}$ are compact, the above assumption holds trivially. More generally, one can formulate conditions in terms of suitable Lyapunov functions that ensure the above almost sure tightness property. Recall that for every $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \equiv q \in \Pi_{\mathrm{SM}}$, $(X_t)_{t\in\mathbb{N}_0}$ is a Markov chain under $\mathbb{P}^\pi_\mu$ with transition probability kernel defined by (2.1).

Assumption 2.2. *For each $q \in \Pi_{SM}$, the Markov chain with transition kernel ϱ_q has a unique invariant probability measure denoted as λ_q.*

Remark 2.1. *Note that if $q, \tilde{q} \in \Pi_{SM}$ and $q(x) = \tilde{q}(x)$ for λ_q a.e. x, then $\lambda_q = \lambda_{\tilde{q}}$. Indeed, for $C \in \mathcal{B}(\mathbb{X})$*

$$\begin{aligned}\lambda_q(C) &= \int_{\mathbb{X}} \varrho_q(x, C)\lambda_q(dx) = \int_{\mathbb{X}\times\mathbb{A}} \mathcal{Q}((x,a), C)q(x, da)\lambda_q(dx)\\ &= \int_{\mathbb{X}\times\mathbb{A}} \mathcal{Q}((x,a), C)\tilde{q}(x, da)\lambda_q(dx) = \int_{\mathbb{X}} \varrho_{\tilde{q}}(x, C)\lambda_q(dx).\end{aligned}$$

Thus λ_q is an invariant probability measure for the Markov chain with transition kernel $\varrho_{\tilde{q}}$ and consequently, from Assumption 2.2, $\lambda_q = \lambda_{\tilde{q}}$.

Assumption 2.3. *For every $f \in C_b(\mathbb{X})$, the function $(x, a) \mapsto \int_{\mathbb{X}} f(\tilde{x})\mathcal{Q}((x, a), d\tilde{x})$ is in $C_b(\mathbb{X} \times \mathbb{A})$.*

Example 2.1. *Suppose that $\mathbb{X}$ and $\mathbb{A}$ are compact and $\{X_t\}$ is a controlled stochastic dynamical system described as*

$$X_{t+1} = F(X_t, A_t, W_t),\ t \in \mathbb{N}_0,$$

where $F : \mathbb{X} \times \mathbb{A} \times \mathbb{Z} \to \mathbb{X}$ is a continuous function, $\mathbb{Z}$ is some Polish space and $\{W_t\}_{t\in\mathbb{N}_0}$ is a $\mathbb{Z}$ valued i.i.d. sequence with common probability law ϑ. Suppose that there is a $\lambda \in \mathcal{P}(\mathbb{X})$ such that for every $(x, a) \in \mathbb{X} \times \mathbb{A}$, the probability law of $F(x, a, W_0)$, denoted as $\theta_{x,a}$, is absolutely continuous with respect to λ and

$$\text{for some } \kappa \in (0, \infty), \frac{d\theta_{x,a}}{d\lambda}(r) \geq \kappa,\ \lambda \text{ a.e. } r,\ \text{for all } (x, a) \in \mathbb{X} \times \mathbb{A}.$$

Then it is easy to check that all of Assumptions 2.1 - 2.3 are satisfied.

Assumptions 2.1 – 2.3 will hold throughout this work and thus will not be noted explicitly in the statement of results.

3. Action Time Sharing Policies.

For the rest of this work we will consider a $q \in \Pi_{\text{SMC}}$ which leads to close to optimal performance for the controlled system. Indeed, as remarked earlier, under suitable conditions on the one stage cost function, the transition kernel $\mathcal{Q}$ and spaces $(\mathbb{X}, \mathbb{A})$, one can show that an optimal control can be found in the family Π_{SM}. Under further smoothness and non-degeneracy conditions one can obtain a sequence of controls in Π_{SMC} such that the associated costs converge to that for the optimal control; in particular for every $\epsilon > 0$, we can find a ϵ-optimal control that belongs to Π_{SMC}. Although we will not appeal to the (near) optimality properties in our proofs, the control q considered above can be regarded as such an ϵ-optimal control. In applications one often encounters controls which are continuous except across some "boundary" surfaces: these may be, for example, regions where some queue is empty. Such discontinuities may be handled by re-defining the metric so that these surfaces become "isolated." However, in order to focus on the main issues, we shall not pursue this extension here. Our main goal is to construct, for a given $q \in \Pi_{\text{SMC}}$, a family of control policies that allow for much more flexibility in implementation than q and lead to the same cost value (as that for q) for a broad range of one stage cost functions.

Define $\theta_q \in \mathcal{P}(\mathbb{X} \times \mathbb{A})$ as

$$\theta_q(C \times D) = \int_C q(x)(D)\lambda_q(dx),\ C \in \mathcal{B}(\mathbb{X}),\ D \in \mathcal{B}(\mathbb{A}).$$

An immediate consequence of assumptions made in Section 2 is the following lemma. The result can be deduced from a more general result given in Section 4.2 (Lemma 4.2) and thus the proof is omitted.

Lemma 3.1. *For each $\mu \in \mathcal{P}(\mathbb{X})$ the sequence of probability measures $\{\Phi_N(\omega), N \in \mathbb{N}\}$ converges weakly, as $N \to \infty$, to θ_q, for $\mathbb{P}_\mu^q$ a.e. ω.*

Lemma 3.1 in particular says that, if the one stage cost function $c \in C_b(\mathbb{X} \times \mathbb{A})$, then the pathwise cost per unit time associated with q, namely J_S (see (2.2)), in fact exists as a limit and equals $\int_{\mathbb{X}\times\mathbb{A}} c(x,a)\theta_q(dxda)$, $\mathbb{P}_\mu^q$ a.e.

We now introduce a family of control policies that are quite flexible and are also well suited for estimation of unknown parameters and for broader information collection purposes, referred to as *action time sharing* (ATS) control policies. An ATS policy associated with q will be such that the corresponding pathwise cost per unit time is the same as that for q. Such a policy is

defined in terms of a sequence of measurable partitions $\{\Lambda_k\}_{k\geq 1}$ of the state space $\mathbb{X}$:

$$\Lambda_k = \{B_{kl}\}_{l=1}^{\tau(k)},\ \mathbb{X} = \bigcup_{l=1}^{\tau(k)} B_{kl},\ B_{kl}\cap B_{kl'} = \emptyset \text{ if } l \neq l' \tag{3.1}$$

such that $|\Lambda_k| = \sup_{l\in R(k)} \operatorname{diam}(B_{kl}) \to 0$ as $k\to\infty$, where $R(k) = \{1,\cdots,\tau(k)\}$. By convention, when $\tau(k) = \infty$, $R(k) = \mathbb{N}$. We refer to $\{\Lambda_k\}_{k\geq 1}$ as a sequence of *converging partitions*. Associated with such a sequence, consider a sequence of random kernels $\{p_k\}_{k\geq 1}$,

$$p_k : \Omega\times\mathbb{X}\times\mathcal{B}(\mathbb{A}) \to [0,1]$$

defined as follows: For $(\omega,x,D)\in\Omega\times\mathbb{X}\times\mathcal{B}(\mathbb{A})$ and $k\in\mathbb{N}$, fix l so that $x\in B_{kl}$. Then set

$$p_k(\omega,x,D) \equiv p_k^\omega(D\mid x) = \begin{cases} \frac{\sum_{j=0}^{k-1} 1_D(A_j(\omega))1_{B_{kl}}(X_j(\omega))}{\sum_{j=0}^{k-1} 1_{B_{kl}}(X_j(\omega))} & \text{if } \sum_{j=0}^{k-1} 1_{B_{kl}}(X_j(\omega)) \neq 0 \\ 1_{\{a_0(x)\in D\}} & \text{if } \sum_{j=0}^{k-1} 1_{B_{kl}}(X_j(\omega)) = 0 \end{cases} \tag{3.2}$$

where $a_0 : \mathbb{X}\to\mathbb{A}$ is an arbitrary fixed measurable function such that $a_0(x)\in\mathbb{U}(x)$ for all $x\in\mathbb{X}$.

Definition 3.1. *Given $\mu\in\mathcal{P}(\mathbb{X})$, a policy $\pi\in\Pi$ is called an action time sharing (ATS) policy for q corresponding to the initial condition μ if for $\mathbb{P}_\mu^\pi$ a.e. ω, there is a sequence of converging partitions $\{\Lambda_k(\omega)\}_{k\geq 1}$, such that for every compact set $K\subset\mathbb{X}$*

$$\sup_{x\in K} \|p_k^\omega(\cdot\mid x) - q(\cdot\mid x)\|_{BL} \to 0,\ \text{as } k\to\infty. \tag{3.3}$$

We denote the collection of all ATS policies for q, corresponding to the initial condition μ, by $\Pi_{ATS}(q,\mu)$.

The following is the main result of this section.

Theorem 3.1. *Let $\mu\in\mathcal{P}(\mathbb{X})$. Fix $\pi\in\Pi_{ATS}(q,\mu)$. Then, as $k\to\infty$, $\Phi_k(\omega)\to\theta_q$ for $\mathbb{P}_\mu^\pi$ a.e. ω.*

Proof. From Assumption 2.1 we can find $\mathcal{N}_1\in\mathcal{F}$ such that $\mathbb{P}_\mu^\pi(\mathcal{N}_1) = 0$ and for all $\omega\in\mathcal{N}_1^c$, $\{\Phi_n(\omega)\}_{n\geq 1}$ is tight. For $f\in\mathcal{S}(\mathbb{X})$, define

$$M_n^f = \sum_{j=0}^{n-1}\left[\int_{\mathbb{X}} f(\tilde{x})\mathcal{Q}((X_j,A_j),d\tilde{x}) - f(X_{j+1})\right].$$

Then, under $\mathbb{P}_\mu^\pi$, $\{M_n^f\}$ is a martingale with bounded increments and so by the strong law of large numbers for such martingales (see e.g.. [13, Theorem VII.5.4]), $\frac{1}{n}M_n^f\to 0$, a.s. $\mathbb{P}_\mu^\pi$. Let $\mathcal{N}_2\in\mathcal{F}$ be such that $\mathbb{P}_\mu^\pi(\mathcal{N}_2) = 0$ and

$$\text{for all } \omega\in\mathcal{N}_2^c,\ \text{and all } f\in\mathcal{S}(\mathbb{X}), \frac{1}{n}M_n^f(\omega)\to 0,\ \text{as } n\to\infty. \tag{3.4}$$

Since $\mathbb{X}$ is locally compact, we can find a sequence $\{K_n\}_{n\geq 1}$ of compact subsets of $\mathbb{X}$ such that

$$K_n^o \subset K_n \subset K_{n+1}^o, \text{ and } \cup_{n\geq 1} K_n = \mathbb{X}.$$

Since $\pi \in \Pi_{\text{ATS}}(q)$, we can find a $\mathcal{N}_3 \in \mathcal{F}$ such that $\mathbb{P}_\mu^\pi(\mathcal{N}_3) = 0$ and, for each $\omega \in \mathcal{N}_3^c$, a sequence $\{\Lambda_k(\omega)\}_{k\geq 1}$ of converging partitions for which, as $k \to \infty$,

$$\sup_{x\in K_n} \left| \int_{\mathbb{A}} g(a) p_k^\omega(da \mid x) - \int_{\mathbb{A}} g(a) q(da \mid x) \right| \to 0, \text{ for every } n \geq 1 \text{ and } g \in \mathcal{S}(\mathbb{A}), \tag{3.5}$$

where p_k is defined through (3.2). Now let $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2 \cup \mathcal{N}_3$ and fix $\omega \in \mathcal{N}^c$. Choose a subsequence $\{n_k\}$ along which $\Phi_{n_k}(\omega)$ converges to some $\Phi(\omega) \in \mathcal{P}(\mathbb{X} \times \mathbb{A})$. Suppressing ω in notation, the measure Φ can be disintegrated as follows: For some $\gamma \in \mathcal{P}(\mathbb{X})$ and a transition probability kernel $\hat{p} : \mathbb{X} \times \mathcal{B}(\mathbb{A}) \to [0, 1]$, $\Phi(dx\, da) = \hat{p}(x, da)\gamma(dx)$, namely

$$\Phi(C \times D) = \int_C \hat{p}(x, D)\gamma(dx), \text{ for all } C \in \mathcal{B}(\mathbb{X}), D \in \mathcal{B}(\mathbb{A}). \tag{3.6}$$

Note that $\hat{p}(\cdot \mid x) \equiv \hat{p}(x, \cdot) \in \Pi_{\text{SM}}$. We claim that

$$\gamma = \lambda_{\hat{p}}. \tag{3.7}$$

To prove the claim it suffices, in view of Assumption 2.2, to show that for all $f \in \mathcal{S}(\mathbb{X})$

$$\int_{\mathbb{X}\times\mathbb{A}} \left(\int_{\mathbb{X}} f(\tilde{x}) \mathcal{Q}((x, a), d\tilde{x}) \right) \hat{p}(da \mid x)\gamma(dx) = \int_{\mathbb{X}} f(x)\gamma(dx). \tag{3.8}$$

Note that the right side of (3.8) equals the limit (as $k \to \infty$) of $\frac{1}{n_k} \sum_{j=0}^{n_k-1} f(X_j(\omega))$, while left side equals (using Assumption 2.3) the limit of

$$\frac{1}{n_k} \sum_{j=0}^{n_k-1} \int_{\mathbb{X}} f(\tilde{x}) \mathcal{Q}((X_j, A_j), d\tilde{x}).$$

Also, from (3.4), the difference of the above two quantities approaches 0 as $k \to \infty$. This proves (3.8) and thus (3.7) follows. To complete the proof of the theorem we will now show that for every $f \in \mathcal{S}(\mathbb{X})$ and $g \in \mathcal{S}(\mathbb{A})$

$$\int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)q(da \mid x)\lambda_{\hat{p}}(dx) = \int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)\hat{p}(da \mid x)\lambda_{\hat{p}}(dx). \tag{3.9}$$

This will prove that $q(\cdot \mid x) = \hat{p}(\cdot \mid x)$, $\lambda_{\hat{p}}$ a.e. x, and consequently, from Remark 2.1, $\lambda_{\hat{p}} = \lambda_q$. Now fix a $(f, g) \in \mathcal{S}(\mathbb{X}) \times \mathcal{S}(\mathbb{A})$. We first show that

$$\lim_{k\to\infty} \left| \int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)p_{n_k}(da \mid x)\Phi_{n_k}^{(1)}(dx) - \int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)q(da \mid x)\lambda_{\hat{p}}(dx) \right| = 0, \tag{3.10}$$

where $\Phi_{n_k}^{(1)}$ is the first marginal of Φ_{n_k}. Let

$$\phi_k(x) = \int_{\mathbb{A}} g(a)p_{n_k}(da \mid x),\ \phi(x) = \int_{\mathbb{A}} g(a)q(da \mid x),\ x \in \mathbb{X}.$$

Since $\omega \in \mathcal{N}_3^c$, we have (see (3.5)) that for every compact K in $\mathbb{X}$

$$\sup_{x\in K} |\phi_k(x) - \phi(x)| \to 0, \text{ as } k \to \infty.$$

Also, from (3.7)

$$\Phi^{(1)}_{n_k} \to \gamma = \lambda_{\hat{p}}.$$

Since $q \in \Pi_{\text{SMC}}$, $\phi \in C_b(\mathbb{X})$ and thus combining the above two displays, we have, as $k \to \infty$,

$$\int_{\mathbb{X}} f(x)\phi_k(x)\Phi^{(1)}_{n_k}(dx) \to \int_{\mathbb{X}} f(x)\phi(x)\lambda_{\hat{p}}(dx).$$

This proves (3.10). We now show

$$\lim_{k\to\infty} \left| \int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)p_{n_k}(da \mid x)\Phi^{(1)}_{n_k}(dx) - \int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)\Phi_{n_k}(da\, dx) \right| = 0. \tag{3.11}$$

Suppressing ω from the notation, suppose that $\Lambda_k(\omega) \equiv \Lambda_k$ is given as in (3.1). Along with the sequence $\{\Lambda_k\}_{k\geq 1}$ we consider a sequence of sets

$$\mathbb{X}_k = \{x_{k1}, \cdots x_{k\tau(k)}\} \subset \mathbb{X},\ k \geq 1 \tag{3.12}$$

such that $x_{kl} \in B_{kl}$ for all $l = 1, \cdots \tau(k)$. We will refer to x_{kl} as the *center* of the set B_{kl}. Define, for $k \geq 1$, $b_k : \mathbb{X} \to \mathbb{X}$ as

$$b_k(x) = \sum_{l=1}^{\tau(k)} x_{kl} 1_{B_{kl}}(x),\ x \in \mathbb{X}.$$

Fix $\epsilon > 0$. Since f is uniformly continuous and $|\Lambda_n| \to 0$ as $n \to \infty$, we can find $n_0 \in \mathbb{N}$ such that

$$\sup_{l\in R(n)} \sup_{x,y\in B_{nl}} |f(x) - f(y)| < \epsilon, \text{ for all } n \geq n_0. \tag{3.13}$$

Fix $k_0 \in \mathbb{N}$ such that $n_k \geq n_0$ whenever $k \geq k_0$. For $k \geq k_0$

$$\int_{\mathbb{A}} g(a)p_{n_k}(da \mid x) = \frac{\sum_{j=0}^{n_k-1} g(A_j)1_{\{b_{n_k}(x)\}}(b_{n_k}(X_j))}{\sum_{j=0}^{n_k-1} 1_{\{b_{n_k}(x)\}}(b_{n_k}(X_j))}.$$

This shows that

$$\int_{\mathbb{X}\times\mathbb{A}} f(x)g(a)p_{n_k}(da \mid x)\Phi^1_{n_k}(dx) = \frac{1}{n_k}\sum_{i=0}^{n_k-1} f(X_i)\frac{\sum_{j=0}^{n_k-1} g(A_j)1_{\{b_{n_k}(X_i)\}}(b_{n_k}(X_j))}{\sum_{j=0}^{n_k-1} 1_{\{b_{n_k}(X_i)\}}(b_{n_k}(X_j))}.$$

Note that $b_{n_k}(X_j) = b_{n_k}(X_i)$ if and only if X_j and X_i are in the same B_{n_kl} and in that case, whenever $k \geq k_0$, $|f(X_i) - f(X_j)| \leq \epsilon$. Using this observation the right side of the above display can be written as

$$\frac{1}{n_k}\sum_{i=0}^{n_k-1} \frac{\sum_{j=0}^{n_k-1} f(X_j)g(A_j)1_{\{b_{n_k}(X_i)\}}(b_{n_k}(X_j))}{\sum_{j=0}^{n_k-1} 1_{\{b_{n_k}(X_i)\}}(b_{n_k}(X_j))} + \varpi(k),$$

where $|\varpi(k)| \le \epsilon \sup_{a\in\mathbb{A}} |g(a)|$ for $k \ge k_0$. The first term in the display can be written as

$$\begin{aligned}
& \frac{1}{n_k} \sum_{i=0}^{n_k-1} \sum_{l=1}^{\tau(n_k)} 1_{B_{n_k l}}(X_i) \frac{\sum_{j=0}^{n_k-1} f(X_j) g(A_j) 1_{\{x_{n_k l}\}}(b_{n_k}(X_j))}{\sum_{j=0}^{n_k-1} 1_{\{x_{n_k l}\}}(b_{n_k}(X_j))} \\
= & \frac{1}{n_k} \sum_{l=1}^{\tau(n_k)} \left(\sum_{i=0}^{n_k-1} 1_{B_{n_k l}}(X_i) \right) \frac{\sum_{j=0}^{n_k-1} f(X_j) g(A_j) 1_{\{x_{n_k l}\}}(b_{n_k}(X_j))}{\sum_{j=0}^{n_k-1} 1_{B_{n_k l}}(X_j)} \\
= & \frac{1}{n_k} \sum_{l=1}^{\tau(n_k)} \sum_{j=0}^{n_k-1} f(X_j) g(A_j) 1_{\{x_{n_k l}\}}(b_{n_k}(X_j)) \\
= & \frac{1}{n_k} \sum_{j=0}^{n_k-1} f(X_j) g(A_j) \\
= & \int_{\mathbb{X}\times\mathbb{A}} f(x) g(a) \Phi_{n_k}(da\, dx).
\end{aligned}$$

Thus for $k \ge k_0$

$$\left| \int_{\mathbb{X}\times\mathbb{A}} f(x) g(a) p_{n_k}(da \mid x) \Phi^1_{n_k}(dx) - \int_{\mathbb{X}\times\mathbb{A}} f(x) g(a) \Phi_{n_k}(da\, dx) \right| \le |\varpi(k)| \le \epsilon \sup_{a\in\mathbb{A}} |g(a)|.$$

Since $\epsilon > 0$ is arbitrary, this proves (3.11). Combining (3.10) and (3.11) we have

$$\lim_{k\to\infty} \left| \int_{\mathbb{X}\times\mathbb{A}} f(x) g(a) \Phi_{n_k}(da\, dx) - \int_{\mathbb{X}\times\mathbb{A}} f(x) g(a) q(da \mid x) \lambda_{\hat{p}}(dx) \right| = 0.$$

The above display, along with (3.6) and (3.7) yields (3.9) and, as noted above (3.9), shows that $q(\cdot \mid x) = \hat{p}(\cdot \mid x)$, $\lambda_{\hat{p}}$ a.e. x, and $\lambda_{\hat{p}} = \lambda_q$. Thus $\Phi = \theta_q$ and the result follows. ■

As a immediate corollary of the above theorem and Lemma 3.1 we have to following result on the convergence of costs. The result says that for a broad family of one stage cost functions, the pathwise cost per unit time for q is same as that for any $\pi \in \Pi_{\text{ATS}}(q)$.

Corollary 3.1. *Let $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \in \Pi_{ATS}(q, \mu)$. Then for any $c \in C_b(\mathbb{X}\times\mathbb{A})$, J_S defined by (2.2) in fact exists as a limit and equals $\int c(x,a)\theta_q(dxda)$, both, a.e. $\mathbb{P}^\pi_\mu$ and $\mathbb{P}^q_\mu$.*

4. Construction of ATS Policies.

In this section we will give a basic construction for a $\pi \in \Pi_{\text{ATS}}(q, \mu)$ for an arbitrary $q \in \Pi_{\text{SMC}}$ and $\mu \in \mathcal{P}(\mathbb{X})$. We will then describe how this construction can be modified in a simple manner to define control policies that are well suited for estimation and information collection purposes while producing the same value for the pathwise cost per unit time. To keep the presentation simple we assume that $\mathbb{U}(x) = \mathbb{A}$ and that $\mathbb{A}$ is a compact metric space. We will further make the following recurrence assumption.

Assumption 4.1. *For every $\pi \in \Pi$, $\mu \in \mathcal{P}(\mathbb{X})$, and $C \in \mathcal{B}(\mathbb{X})$ with a nonempty interior,*

$$\mathbb{P}^{\pi}_{\mu}(X_t \in C, \text{ for some } t \in \mathbb{N}) = 1.$$

The above assumption will hold throughout this section. Note that this assumption is satisfied under the setting of Example 2.1 if the probability measure λ in the example satisfies $\lambda(C) > 0$ for every $C \in \mathcal{B}(\mathbb{X})$ with a nonempty interior.

We begin with the following lemma. Let

$$\Theta = \{\vartheta \in \mathcal{P}(\mathbb{A}) : \vartheta \text{ is supported on finitely many points}\}. \tag{4.1}$$

For $\vartheta \in \Theta$, denote by $S(\vartheta)$ the support of ϑ.

Lemma 4.1. *There is a $\Psi \equiv (\Psi_1, \cdots) : \Theta \to \mathbb{A}^{\infty}$ such that for every $\vartheta \in \Theta$: (i) $\Psi_i(\vartheta) \in S(\vartheta)$, $i \geq 1$; (ii) The probability measure $m_n(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\Psi_i(\vartheta)}$ satisfies*

$$||m_n(\vartheta) - \vartheta||_{BL} \leq \frac{4\ \#(S(\vartheta))}{n}$$

where $\#(S(\vartheta))$ is the cardinality of $S(\vartheta)$.

Proof. Fix $\vartheta \in \Theta$. Then ϑ can be written as

$$\vartheta = \sum_{j=1}^{l} p_j \delta_{a_j},$$

where $l = \#(S(\vartheta)) \in \mathbb{N}$, $a_j \in \mathbb{A}$, $p_j \in (0, 1]$, and $\sum_{j=1}^{l} p_j = 1$.

Define, for $m \in \mathbb{N}$ and $j = 1, \ldots, l$,

$$k_j(m) = \lfloor m l p_j \rfloor, \text{ and } \alpha(m) = \sum_{j=1}^{l} k_j(m).$$

Set $\alpha(0) = 0$. It is easily seen that

$$(m-1)l \leq \alpha(m) \leq ml, \tag{4.2}$$

and so $\alpha(m) \to \infty$ as $m \to \infty$.

We now define a sequence $\{\psi_j\}_{j=1}^{\infty}$ with values in $\mathbb{A}$, such that, for each $m \geq 1$ and $r = 1, \ldots, l$,

$$\#\{j \in \{1, \ldots, \alpha(m)\} : \psi_j = a_r\} = k_r(m). \tag{4.3}$$

One can define $\{\psi_j\}_{j=1}^{\infty}$ inductively as follows.

Consider $m = 1$. Define

$$\psi_j = a_r \text{ whenever } \sum_{i=1}^{r-1} k_i(1) < j \leq \sum_{i=1}^{r} k_i(1), r = 1, \cdots l.$$

This defines $\{\psi_j\}_{j=1}^{\alpha(1)}$. Suppose now $\{\psi_j\}_{j=1}^{\alpha(N)}$ has been defined such that (4.3) holds with $m = N$. Assume without loss of generality that $\alpha(N+1) > \alpha(N)$. We now define $\{\psi_j\}_{j=\alpha(N)+1}^{\alpha(N+1)}$. Note that $k_r(N+1) \geq k_r(N)$. Let $b_r(N+1) = k_r(N+1) - k_r(N)$, and set

$$\psi_j = a_r \text{ whenever } \alpha(N) + \sum_{i=1}^{r-1} b_i(N+1) < j \leq \alpha(N) + \sum_{i=1}^{r} b_i(N+1), r = 1, \cdots l.$$

This completes the definition of $\{\psi_j\}_{j=1}^{\alpha(N+1)}$.

Define $\Psi_j(\vartheta) = \psi_j, j \in \mathbb{N}$. Fix $n \in \mathbb{N}$ such that $\alpha(N) \leq n \leq \alpha(N+1)$ for some $N \in \mathbb{N}$. If $N = 0$,

$$\|m_n(\vartheta) - \vartheta\|_{\mathrm{BL}} = \left\| \frac{1}{n} \sum_{j=1}^{n} \delta_{\psi_j} - \sum_{i=1}^{l} p_i \delta_{a_i} \right\|_{\mathrm{BL}} = \sup_{f \in \mathcal{C}_1(\mathbb{A})} \left| \frac{1}{n} \sum_{j=1}^{n} f(\psi_j) - \sum_{i=1}^{l} p_i f(a_i) \right| \leq 2 \leq \frac{2l}{n},$$

where the last inequality follows from (4.2). Consider now the case $N \geq 1$. Then

$$\begin{aligned} \|m_n(\vartheta) - \vartheta\|_{\mathrm{BL}} &= \left\| \frac{1}{n} \sum_{j=1}^{n} \delta_{\psi_j} - \sum_{i=1}^{l} p_i \delta_{a_i} \right\|_{\mathrm{BL}} \\ &\leq \sup_{f \in \mathcal{C}_1(\mathbb{A})} \left(\left| \frac{1}{n} \sum_{i=1}^{l} k_i(N) f(a_i) - \sum_{i=1}^{l} p_i f(a_i) \right| + \left| \frac{1}{n} \sum_{j=\alpha(N)+1}^{n} f(\psi_j) \right| \right) \\ &\leq \sum_{i=1}^{l} \left| \frac{k_i(N) - np_i}{n} \right| + \left| \frac{\alpha(N+1) - \alpha(N)}{n} \right|. \end{aligned} \tag{4.4}$$

By (4.2),

$$\alpha(N+1) - \alpha(N) \leq (N+1)l - (N-1)l = 2l.$$

Also, for $j = 1, \cdots l$,

$$k_j(N) - np_j \leq Nlp_j - np_j \leq Nlp_j - \alpha(N)p_j \leq Nlp_j - (N-1)lp_j \leq lp_j,$$

and

$$k_j(N) - np_j \geq Nlp_j - 1 - \alpha(N+1)p_j \geq Nlp_j - 1 - (N+1)lp_j = -1 - lp_j.$$

Using the above estimate in (4.4) we now have

$$\|m_n(\vartheta) - \vartheta\|_{\mathrm{BL}} \leq \frac{4l}{n}.$$

The lemma follows. ■

4.1. A Basic Construction.

Fix $q \in \Pi_{\rm SMC}$ and $\mu \in \mathcal{P}(\mathbb{X})$. We now give a pathwise construction of a $\pi \in \Pi_{\rm ATS}(q,\mu)$. Let $\{\tilde{\Lambda}_k\}_{k\geq 1}$ be a sequence of measurable partitions of $\mathbb{X}$:

$$\tilde{\Lambda}_k = \{\tilde{B}_{kl}\}_{l=1}^{\tilde{\tau}(k)},\ \mathbb{X} = \bigcup_{l=1}^{\tilde{\tau}(k)} \tilde{B}_{kl},\ \tilde{B}_{kl} \cap \tilde{B}_{kl'} = \emptyset \text{ if } l \neq l' \tag{4.5}$$

such that $|\tilde{\Lambda}_k| = \sup_{l\in\tilde{R}(k)} \mathrm{diam}(\tilde{B}_{kl}) \to 0$ as $k \to \infty$, where $\tilde{R}(k) = \{1, \cdots, \tilde{\tau}(k)\}$. Each $\tilde{B}_{kl}$ is required to have a nonempty interior. Also, we assume that the sequence $\tilde{\Lambda}_k$ is nested, namely, for every $k \geq 1$ and $l \in \tilde{R}(k+1)$, there is a $l' \in \tilde{R}(k)$ such that $\tilde{B}_{(k+1)l} \subset \tilde{B}_{kl'}$. We also assume that for any compact $K \subset \mathbb{X}$ and $k \geq 1$,

$$\#\{l : \tilde{B}_{kl} \bigcap K \neq \emptyset\} < \infty.$$

Associated with the sequence $\{\tilde{\Lambda}_k\}$, we define sets $\{\tilde{\mathbb{X}}_k\}$ and maps $\{\tilde{b}_k\}$ analogous to as below (3.11). Namely, for $k \geq 1$

$$\tilde{\mathbb{X}}_k = \{\tilde{x}_{k1}, \cdots \tilde{x}_{k\tilde{\tau}(k)}\} \subset \mathbb{X}, \tag{4.6}$$

is such that $\tilde{x}_{kl} \in \tilde{B}_{kl}$ for all $l \in \tilde{R}(k)$ and $\tilde{b}_k : \mathbb{X} \to \mathbb{X}$ is given as

$$\tilde{b}_k(x) = \sum_{l\in\tilde{R}(k)} \tilde{x}_{kl} 1_{\tilde{B}_{kl}}(x),\ x \in \mathbb{X}.$$

As before, $\tilde{x}_{kl}$ is called the *center* of the set $\tilde{B}_{kl}$. Since $x \mapsto q(\cdot \mid x)$ is a continuous map from $\mathbb{X}$ to $\mathcal{P}(\mathbb{A})$, we have that for every compact $K \subset \mathbb{X}$

$$\sup_{x\in K} ||q(\cdot \mid x) - q(\cdot \mid \tilde{b}_k(x))||_{\rm BL} \to 0,\ \text{as } k \to \infty. \tag{4.7}$$

Next let $\{\Lambda'_k\}_{k\geq 1}$ be a sequence of measurable partitions of $\mathbb{A}$:

$$\Lambda'_k = \{F_{km}\}_{m=1}^{\ell(k)},\ \mathbb{A} = \bigcup_{m=1}^{\ell(k)} F_{km},\ F_{km} \cap F_{km'} = \emptyset \text{ if } m \neq m' \tag{4.8}$$

such that $\ell(k) < \infty$ for all k and $|\Lambda'_k| \to 0$ as $k \to \infty$. Define a sequence of finite sets $\mathbb{A}_k = \{a_{k1}, \cdots, a_{k\ell(k)}\}$ such that $a_{km} \in F_{km}$ for all $m = 1, \cdots \ell(k)$. Let, for $k \geq 1$, $b'_k : \mathbb{A} \to \mathbb{A}_k$ be defined as

$$b'_k(a) = \sum_{m=1}^{\ell(k)} a_{km} 1_{\tilde{F}_{km}}(a),\ a \in \mathbb{A}.$$

We will now construct a sequence of $\mathbb{X}\times\mathbb{A}$ valued random variables $Z \equiv (\bar{X}_t, \bar{A}_t)_{t\in\mathbb{N}_0}$ on a suitable probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that $\bar{X}_0$ has probability law μ and the probability law of Z corresponds to a controlled system associated with a policy $\pi \in \Pi_{\rm ATS}(q,\mu)$. More

precisely, denoting the measure induced by Z on $(\Omega, \mathcal{F})$, by $\mathbb{P}^*$ (i.e. $\mathbb{P}^* = \bar{\mathbb{P}} \circ Z^{-1}$), we will obtain an admissible control policy $\pi = \{\pi_t\}_{t\in\mathbb{N}_0}$ by disintegrating, for $t \in \mathbb{N}_0$, the measure $\bar{\mathbb{P}}_t = \mathbb{P}^* \circ (H_t, A_t)^{-1} \in \mathcal{P}(\mathbb{H}_t \times \mathbb{A})$, as

$$\bar{\mathbb{P}}_t(dh, da) = \pi_t(h, da) \ \left(\mathbb{P}^* \circ H_t^{-1}\right)(dh). \tag{4.9}$$

Note that with π defined through the above equation, we have that the controlled system $\mathbb{P}^\pi_\mu = \mathbb{P}^*$. The construction of $(\bar{X}_t, \bar{A}_t)_{t\in\mathbb{N}_0}$ will be carried out in a recursive fashion such that

$$\mathbb{P}(\bar{X}_{t+1} \in C \mid (\bar{X}_j, \bar{A}_j), j \le t) = \mathcal{Q}((\bar{X}_t, \bar{A}_t), C), \ C \in \mathcal{B}(\mathbb{X}), \ t \in \mathbb{N}_0.$$

The recursive construction of the sequence $(\bar{A}_t)$ is described in what follows.

Let $\{K_n\}_{n\ge1}$ be the sequence of compact sets in $\mathbb{X}$ introduced in Section 3. Let, for $r \ge 1$, by relabeling sets if needed,

$$\tilde{\Lambda}^0_r = \{\tilde{B}_{r1}, \cdots \tilde{B}_{rj(r)}\} \subset \tilde{\Lambda}_r$$

be the finite collection of sets such that $\tilde{B}_{rm} \in \tilde{\Lambda}^0_r$ if and only if $\tilde{B}_{rm} \bigcap K_r$ is non-empty. For $m = 1, \cdots, j(r)$, define $q^{r,m} \in \mathcal{P}(\mathbb{A})$ as $q^{r,m} = q(\cdot \mid x_{rm})$. Define, for $r \ge 1$, $\eta_r : \mathcal{P}(\mathbb{A}) \to \mathcal{P}(\mathbb{A}_r)$ as

$$\eta_r(\vartheta) = \sum_{j=1}^{\ell(r)} \delta_{a_{rj}} \vartheta(F_{rj}), \ \vartheta \in \mathcal{P}(\mathbb{A}).$$

Note that

$$\sup_{\vartheta\in\mathcal{P}(\mathbb{A})} ||\eta_r(\vartheta) - \vartheta||_{\mathrm{BL}} \le |\Lambda'_r| \to 0, \text{ as } r \to \infty. \tag{4.10}$$

Set $\tilde{q}^{r,m} = \eta_r(q^{r,m})$, $r \ge 1$. Note that $\tilde{q}^{r,m} \in \Theta$ (cf. 4.1) for all $r \in \mathbb{N}$, $m \le j(r)$. Denote, for $i \ge 1$, the i^{th} component of $\Psi(\tilde{q}^{r,m})$ by $e^r[m, i]$, i.e.

$$\Psi(\tilde{q}^{r,m}) = (e^r[m, 1], e^r[m, 2], \cdots).$$

Note that by definition of Ψ, $e^r[m, i] \in \mathbb{A}_r$ for all $i, r \in \mathbb{N}$, $m \le j(r)$. Furthermore, from Lemma 4.1, for every $N \ge 1$,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \delta_{e^r[m,i]} - \tilde{q}^{r,m} \right\|_{\mathrm{BL}} \le \frac{4\ell(r)}{N}.$$

The sequences $\Psi(\tilde{q}^{r,m})$, $m \le j(r)$, $r \in \mathbb{N}$, will form the basic building blocks for the sequence $(\bar{A}_t)_{t\in\mathbb{N}_0}$. Let $\{\varepsilon_r\}_{r\ge1}$ be a sequence of positive reals such that $\varepsilon_r \downarrow 0$ as $r \to \infty$.

Construction of Z. We are now ready to specify the sequence $(\bar{X}_t, \bar{A}_t)$ on a suitable probability space. The definition of the probability space will be implicit in the construction and a detailed description of the space will be omitted. Let $\bar{X}_0$ be a $\mathbb{X}$ valued random variable with probability law μ.

We now define, recursively in r, sequences $\{\xi^r_k, s^r_k, \zeta^r_k, (i^r[m, k])_{m=1,\cdots j(r)}\}_{k\ge0}$, $r \ge 1$, as follows.

Case $r = 1$**:** Define $\xi_0^r = \bar{X}_0$ and let

$$\begin{aligned} &i^r[m,0] = 1_{\tilde{B}_{r,m}}(\xi_0^r),\ m = 1, \cdots j(r), \\ &m_0^r = \sum_{m=1}^{j(r)} m 1_{\tilde{B}_{r,m}}(\xi_0^r),\ s_0^r = i^r[m_0^r, 0],\ \text{and}\ \zeta_0^r = e^r[m_0^r, s_0^r]. \end{aligned} \tag{4.11}$$

Note that $s_0^r = 1$. Having defined $\{\xi_k^r, s_k^r, \zeta_k^r, (i^r[m,k])_{m=1,\cdots j(r)}\}$ for $k \le k_0$, define $\xi_{k_0+1}^r$ through the relation

$$\bar{\mathbb{P}}(\xi_{k_0+1}^r \in C \mid \mathcal{G}_{k_0}^r) = \mathcal{Q}((\xi_{k_0}^r, \zeta_{k_0}^r), C),\ C \in \mathcal{B}(\mathbb{X}), \tag{4.12}$$

where $\mathcal{G}_{k_0}^r = \sigma\{(\xi_j^r, \zeta_j^r) : j \le k_0\}$, and set

$$i^r[m, k_0+1] = i^r[m,k_0] + 1_{\tilde{B}_{r,m}}(\xi_{k_0+1}^r),\ m = 1, \cdots j(r),\ m_{k_0+1}^r = \sum_{m=1}^{j(r)} m 1_{\tilde{B}_{r,m}}(\xi_{k_0+1}^r), \tag{4.13}$$

and

$$s_{k_0+1}^r = i^r[m_{k_0+1}^r, k_0+1],\ \zeta_{k_0+1}^r = e^r[m_{k_0+1}^r, s_{k_0+1}^r]. \tag{4.14}$$

This completes the definition for $\{\xi_k^r, s_k^r, \zeta_k^r, (i^r[m,k])_{m=1,\cdots j(r)}\}$ for $r = 1$ and $k \in \mathbb{N}_0$.

Set $\varrho_0 = 0$ and define, for $r = 1$,

$$\alpha_r = \varepsilon_r^{-1} \left(2\varrho_{r-1} + 4\left(\ell(r) + \ell(r+1)\right)\right), \tag{4.15}$$

$$\sigma_r = \inf\{k : i^r[m,k] \ge \alpha_r \text{ for all } m = 1, \cdots j(r)\}, \tag{4.16}$$

$$\varrho_r = \varrho_{r-1} + \sigma_r \tag{4.17}$$

Case $r > 1$**:** Let

$$(\xi_0^r, \zeta_0^r) = (\xi_{\sigma_{r-1}}^{r-1} \zeta_{\sigma_{r-1}}^{r-1}),\ i^r[m,0] = 0, m = 1, \cdots j(r).$$

Definition of $\{\xi_k^r, \zeta_k^r, s_k^r, (i^r[m,k])_{m=1,\cdots j(r)}\}_{k\ge 1}$ and $(\alpha_r, \sigma_r, \varrho_r)$, for $r > 1$, is given recursively, exactly as above through (4.12) – (4.17).

Finally, the sequence $(\bar{X}_k, \bar{A}_k)$ is now constructed on the probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ that supports the random variables $\{\xi_k^r, \zeta_k^r, s_k^r, (i^r[m,k])_{m=1,\cdots j(r)}\}_{k\ge 0}$, $(\alpha_r, \sigma_r, \varrho_r)$, $r \in \mathbb{N}$, by piecing together the sequence $(\xi_k^r, \zeta_k^r; k, r \in \mathbb{N}_0)$ as follows,

$$(\bar{X}_k, \bar{A}_k) = (\xi_{k-\varrho_r}^{r+1}, \zeta_{k-\varrho_r}^{r+1}),\ \text{whenever}\ \varrho_r \le k < \varrho_{r+1},\ r \in \mathbb{N}_0.$$

Recall from (4.9) the definition of π and $\mathbb{P}_\mu^\pi$ corresponding to the sequence $(\bar{X}_k, \bar{A}_k)_{k\in\mathbb{N}_0}$. We now show that π constructed in the above fashion is an ATS policy for q with initial condition μ.

Theorem 4.1. *The policy $\pi \in \Pi$ constructed above is in $\Pi_{ATS}(q, \mu)$.*

Proof. From Assumption 4.1 it follows that, with $\bar{\Omega}_0 = \{\omega \in \bar{\Omega} : \varrho_r(\omega) < \infty \text{ for all } r \geq 1\}$, $\bar{\mathbb{P}}(\bar{\Omega}_0) = 1$. Define for $\omega \in \bar{\Omega}_0$, $(x, D) \in \mathbb{X} \times \mathcal{B}(\mathbb{A})$, $\bar{p}_k(\omega, x, D) \equiv \bar{p}_k^{\omega}(D \mid x)$ by the right side of (3.2), replacing (A_j, X_j) there by $(\bar{A}_j, \bar{X}_j)$ and $\{\Lambda_k(\omega)\}$ (suppressing ω from notation throughout) defined as follows: For $k \geq 1$,

$$\Lambda_k = \tilde{\Lambda}_\beta \text{ if } \varrho_\beta < k \leq \varrho_{\beta+1}, \beta = 0, 1, \cdots$$

where $\tilde{\Lambda}_0$ is taken to be $\tilde{\Lambda}_1$. In order to prove the result, it suffices to show that for all $\omega \in \bar{\Omega}_0$ and compact $K \subset \mathbb{X}$

$$\sup_{x \in K} \|\bar{p}_k^{\omega}(\cdot \mid x) - q(\cdot \mid x)\|_{\text{BL}} \to 0, \text{ as } k \to \infty. \tag{4.18}$$

Fix now a compact set $K \subset \mathbb{X}$ and $\epsilon \in (0, 1)$. Using (4.7) and (4.10), choose r_0 large enough so that for all $r \geq r_0$, $K \subset K_r$,

$$\sup_{x \in K} \|q(\cdot \mid x) - q(\cdot \mid \tilde{b}_r(x))\|_{\text{BL}} \leq \epsilon \tag{4.19}$$

and

$$\sup_{\vartheta \in \mathcal{P}(\mathbb{A})} \|\vartheta - \eta_r(\vartheta)\|_{\text{BL}} \leq \epsilon. \tag{4.20}$$

We introduce some additional notation. For $t \geq 1$ and $l = 1, \cdots j(t)$, let

$$n_{tl}(m_1, m_2) = \#\{\bar{X}_j \in \tilde{B}_{tl} : m_1 \leq j < m_2\},\ 0 \leq m_1 \leq m_2 < \infty$$

and for such m_1, m_2, let $\mu_{tl}[m_1, m_2] \in \mathcal{P}(\mathbb{A})$ be defined as follows: For $D \in \mathcal{B}(\mathbb{A})$,

$$\mu_{tl}[m_1, m_2](D) = \begin{cases} n_{tl}(m_1, m_2)^{-1} \sum_{j=m_1}^{m_2-1} 1_D(\bar{A}_j) 1_{\tilde{B}_{tl}}(\bar{X}_j), & \text{if } n_{tl}(m_1, m_2) > 0, \\ \delta_{a_0}(D), & \text{otherwise,} \end{cases}$$

where a_0 is some fixed element of $\mathbb{A}$.

Fix $\beta_0 \geq r_0 + 1$ and consider $k > \varrho_{\beta_0}$. Let $\beta \in \mathbb{N}$, $\beta \geq \beta_0$ be such that $\varrho_\beta < k \leq \varrho_{\beta+1}$. We will now estimate the quantity on the left side of (4.18) for such a k. Fix $x \in K$ and let $i \in \{1, \cdots j(\beta)\}$ be such that $x \in \tilde{B}_{\beta i}$. Since $B_{ki} = \tilde{B}_{\beta i}$ for $\varrho_\beta < k \leq \varrho_{\beta+1}$, we can write

$$\bar{p}_k(\cdot \mid x) = n^{-1}(n_1\nu_1 + n_2\nu_2 + n_3\nu_3), \tag{4.21}$$

where

$$\nu_1 = \mu_{\beta i}[0, \varrho_{\beta-1}],\ \nu_2 = \mu_{\beta i}[\varrho_{\beta-1}, \varrho_\beta],\ \nu_3 = \mu_{\beta i}[\varrho_\beta, k]$$

and

$$n_1 = n_{\beta i}(0, \varrho_{\beta-1}),\ n_2 = n_{\beta i}(\varrho_{\beta-1}, \varrho_\beta),\ n_3 = n_{\beta i}(\varrho_\beta, k),\ n = n_1 + n_2 + n_3.$$

Recall that the sequence $\{\tilde{\Lambda}_k\}$ is nested. Denote the sets in $\tilde{\Lambda}_{\beta+1}$ that are contained in $\tilde{B}_{\beta i}$ as $G_1, G_2, \cdots G_\gamma$ and denote the corresponding centers by $g_1, \cdots, g_\gamma$. Let, for $t = 1, \cdots \gamma$, $m_t = \#\{X_j \in G_t : \varrho_\beta \leq j < k\}$. Then

$$\sum_{t=1}^{\gamma} m_t = n_3 \text{ and } \nu_3 = n_3^{-1} \sum_{t=1}^{\gamma} m_t \nu_{3t}, \tag{4.22}$$

where

$$\nu_{3t}(D) = \begin{cases} m_t^{-1} \sum_{j=\varrho_\beta}^{k-1} 1_D(\bar{A}_j) 1_{G_t}(\bar{X}_j), & \text{if } m_t > 0, \\ \\ \delta_{a_0}(D), & \text{otherwise.} \end{cases}$$

Then, whenever $m_t \neq 0$,

$$\|\nu_{3t} - \eta_{\beta+1}(q(\cdot \mid g_t))\|_{\mathrm{BL}} \leq \frac{4\ell(\beta+1)}{m_t}. \tag{4.23}$$

Also, since $\beta > r_0$, from (4.20), whenever $n_3 > 0$,

$$\|\tilde{\nu}_3 - n_3^{-1} \sum_{t=1}^{\gamma} m_t q(\cdot \mid g_t)\|_{\mathrm{BL}} \leq \epsilon$$

where $\tilde{\nu}_3 = n_3^{-1} \sum_{t=1}^{\gamma} m_t \eta_{\beta+1}(q(\cdot \mid g_t))$, and from (4.19)

$$\|n_3^{-1} \sum_{t=1}^{\gamma} m_t q(\cdot \mid g_t) - q(\cdot \mid x)\|_{\mathrm{BL}} \leq 2\epsilon.$$

Thus, whenever $n_3 > 0$, $\|\tilde{\nu}_3 - q(\cdot \mid x)\|_{\mathrm{BL}} \leq 3\epsilon$. Letting $\tilde{\nu}_1 = \tilde{\nu}_2 = \eta_\beta(q(\cdot \mid \tilde{b}_\beta(x)))$, we have by this estimate and (4.19), (4.20) that

$$\|q(\cdot \mid x) - n^{-1}(n_1\tilde{\nu}_1 + n_2\tilde{\nu}_2 + n_3\tilde{\nu}_3)\|_{\mathrm{BL}} \leq 3\epsilon.$$

Also, from Lemma 4.1,

$$\|\nu_2 - \tilde{\nu}_2\|_{\mathrm{BL}} \leq \frac{4\ell(\beta)}{n_2}$$

and if $n_3 \neq 0$, from (4.23), (4.22) and Lemma 4.1, we have

$$\|\nu_3 - \tilde{\nu}_3\|_{\mathrm{BL}} \leq \frac{4\ell(\beta+1)}{n_3}.$$

Combining the above three displays with (4.21) and the trivial estimate $\|\nu_1 - \tilde{\nu}_1\|_{\mathrm{BL}} \leq 2$, we have

$$\begin{aligned} \|\bar{p}_k(\cdot \mid x) - q(\cdot \mid x)\|_{\mathrm{BL}} &\leq 3\epsilon + n^{-1}\left(2n_1 + 4\left(\ell(\beta) + \ell(\beta+1)\right)\right) \\ &\leq 3\epsilon + \varepsilon_\beta. \end{aligned}$$

where the last inequality follows on observing that $n_1 \leq \varrho_{\beta-1}$, $n \geq n_2 \geq \alpha_\beta$ and using (4.15). Since $x \in K$ and $\epsilon > 0$ are arbitrary and $\beta \to \infty$ as $k \to \infty$, the result follows. ■

4.2. ATS Policies for Simultaneous Estimation and Optimization.

Consider a setting where one has a (near) optimal $q \in \Pi_{\mathrm{SMC}}$ for pathwise cost per unit time associated with some one stage cost function $c \in C_b(\mathbb{X} \times \mathbb{A})$. However, in addition to cost

optimization one has a secondary objective of estimating some unknown parameter in the model. Consistent estimation may require using actions that are not optimal. For example, analogous to the example discussed in the introduction, it could be that under the policy q, estimation is impossible because transitions do not depend at all on the parameter that need to be estimated and thus one needs to deviate from the optimal q in order to gain information on the parameter. ATS policies provide a framework that allows one to introduce such deviations without "paying a price" in terms of the optimization problem. In this section we describe the construction of ATS policies for one such estimation problem.

Let $q \in \Pi_{\text{SMC}}$ be as in Section 4.1 and $c \in C_b(\mathbb{X}\times\mathbb{A})$. Suppose we are given another $q_0 \in \Pi_{\text{SMC}}$ and one would like to obtain consistent estimators for

$$\mathcal{J}_f = \int_{\mathbb{X}\times\mathbb{A}} f(x,a)\theta_{q_0}(dxda),\ f \in C_b(\mathbb{X}\times\mathbb{A}),$$

while achieving the pathwise cost per unit time $\int_{\mathbb{X}\times\mathbb{A}} c(x,a)\theta_q(dxda)$. We will show below that by an appropriate modification of the ATS policy constructed in Section 4.1 one can achieve both goals. We begin by introducing a strengthening of Assumption 2.1.

Let $\{\jmath_k\}_{k\in\mathbb{N}}$ be a sequence of $\{\mathcal{H}_t\}_{t\in\mathbb{N}_0}$- stopping times given on $(\Omega,\mathcal{F})$ such that

$$\jmath_k < \jmath_k + m_k \le \jmath_{k+1},\ \text{for all } \omega \in \Omega$$

for some $m_k \in \mathbb{N}$, $k \ge 1$. Write $\varpi = (\jmath_k, m_k)_{k\in\mathbb{N}}$ and let $\mathbb{T}$ be the family of all such sequences. For $\varpi = (\jmath_k, m_k)_{k\in\mathbb{N}} \in \mathbb{T}$, and $N \ge 1$, let $\Phi_N[\varpi]$ be a measurable map from $\Omega \to \mathcal{P}(\mathbb{X}\times\mathbb{A})$ defined as

$$\Phi_N^\omega[\varpi](F) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{m_k}\sum_{j=\jmath_k}^{\jmath_k+m_k-1} 1_F(X_j,A_j),\ F \in \mathcal{B}(\mathbb{X}\times\mathbb{A}),\ \omega\in\Omega.$$

We will make the following assumption.

Assumption 4.2. *For all $\varpi \in \mathbb{T}$, $\mu \in \mathcal{P}(\mathbb{X})$ and $\pi \in \Pi$, $\{\Phi_N^\omega[\varpi] : N \in \mathbb{N}\}$ is tight for $\mathcal{P}_\mu^\pi$ a.e. ω.*

We note that the assumption is trivially satisfied if $\mathbb{X}$ and $\mathbb{A}$ are compact spaces. More generally, blanket stability conditions in terms of a suitable Lyapunov function can be formulated under which Assumption 4.2 holds.

An immediate consequence of the above assumption and other assumptions from Section 2 is the following.

Lemma 4.2. *Let $\varpi = (\jmath_k, m_k)_{k\in\mathbb{N}} \in \mathbb{T}$ be such that $m_k \to \infty$ as $k \to \infty$. Let $\mu \in \mathcal{P}(\mathbb{X})$, $\pi \in \Pi$ and $q_0 \in \Pi_{SMC}$ be such that for all $k \ge 1$ and $j \in \{0,1,\cdots m_k - 1\}$*

$$\mathbb{P}_\mu^\pi\left((A_{\jmath_k+j}, X_{\jmath_k+j+1}) \in D\times C \mid \mathcal{H}_{\jmath_k+j}\right) = \int_D \mathcal{Q}\left((X_{\jmath_k+j},a),C\right) q_0(X_{\jmath_k+j},da),$$

for all $D\times C \in \mathcal{B}(\mathbb{A}\times\mathbb{X})$, a.e. $\mathbb{P}_\mu^\pi$. Then, as $N\to\infty$,

$$\|\Phi_N^\omega[\varpi] - \theta_{q_0}\|_{BL} \to 0,\quad \text{a.e. } \omega\ [\mathbb{P}_\mu^\pi].$$

Proof. For $f \in \mathcal{S}(\mathbb{X})$, let $\psi_f(x) \doteq \int_{\mathbb{X}} f(y)\varrho_{q_0}(x, dy), x \in \mathbb{X}$. Then, suppressing ω in notation, we have

$$\begin{aligned}
\Psi_f^N &\doteq \left| \int_{\mathbb{X}\times\mathbb{A}} f(x)\Phi_N[\varpi](dxda) - \int_{\mathbb{X}\times\mathbb{A}} \psi_f(x)\Phi_N[\varpi](dxda) \right| \\
&= \left| \frac{1}{N}\sum_{k=1}^{N} \frac{1}{m_k} \sum_{j=\jmath_k}^{\jmath_k+m_k-1} (f(X_j) - \psi_f(X_j)) \right| \\
&= \left| \frac{1}{N}\sum_{k=1}^{N} \frac{1}{m_k} \sum_{j=\jmath_k+1}^{\jmath_k+m_k-1} (f(X_j) - \psi_f(X_{j-1})) + \frac{1}{N}\sum_{k=1}^{N} \frac{f(X_{\jmath_k})}{m_k} - \frac{1}{N}\sum_{k=1}^{N} \frac{\psi_f(X_{\jmath_k+m_k-1})}{m_k} \right| \\
&\le \left| \frac{1}{N}\sum_{k=1}^{N} \frac{1}{m_k} \sum_{j=0}^{m_k-2} (f(X_{\jmath_k+j+1}) - \psi_f(X_{\jmath_k+j})) \right| + \frac{1}{N}\sum_{k=1}^{N} \frac{2|f|_\infty}{m_k}.
\end{aligned} \tag{4.24}$$

Note that for all $k \ge 1$ and $j \in \{0, 1, \ldots, m_k - 2\}$,

$$\begin{aligned}
\psi_f(X_{\jmath_k+j}) &= \int_{\mathbb{X}} f(y)\varrho_{q_0}(X_{\jmath_k+j}, dy) \\
&= \int_{\mathbb{X}\times\mathbb{A}} f(y)\mathcal{Q}((X_{\jmath_k+j}, a), dy)q_0(X_{\jmath_k+j}, da) \\
&= \mathbb{E}_\mu^\pi[f(X_{\jmath_k+j+1})|\mathcal{H}_{\jmath_k+j}] \text{ a.e. } \mathbb{P}_\mu^\pi.
\end{aligned}$$

By the strong law of large numbers for martingales (cf. [13, Theorem VII.5.4]) and the assumption that $m_k \to \infty$ as $k \to \infty$, $\Psi_f^N \to 0$ as $N \to \infty$ a.e. $\mathbb{P}_\mu^\pi$.

Next, for $g \in \mathcal{S}(\mathbb{X})$ and $h \in \mathcal{S}(\mathbb{A})$, let $\phi_{(g,h)}(x) \doteq g(x)\int_{\mathbb{A}} h(a)q_0(x, da)$. Then

$$\begin{aligned}
\Phi_{(g,h)}^N &\doteq \left| \int_{\mathbb{X}\times\mathbb{A}} g(x)h(a)\Phi_N[\varpi](dxda) - \int_{\mathbb{X}\times\mathbb{A}} \phi_{(g,h)}(x)\Phi_N[\varpi](dxda) \right| \\
&= \left| \frac{1}{N}\sum_{k=1}^{N} \frac{1}{m_k} \sum_{j=0}^{m_k-1} \big(g(X_{\jmath_k+j})h(A_{\jmath_k+j}) - \phi_{(g,h)}(X_{\jmath_k+j})\big) \right|
\end{aligned} \tag{4.25}$$

For all $k \ge 1$ and $j \in \{0, 1, \ldots, m_k - 1\}$,

$$\begin{aligned}
\phi_{(g,h)}(X_{\jmath_k+j}) &= g(X_{\jmath_k+j}) \int_{\mathbb{A}} h(a)q_0(X_{\jmath_k+j}, da) \\
&= \mathbb{E}_\mu^\pi[g(X_{\jmath_k+j})h(A_{\jmath_k+j})|\mathcal{H}_{\jmath_k+j}] \text{ a.e. } \mathbb{P}_\mu^\pi.
\end{aligned}$$

Again from the strong law of large numbers for martingales and the fact that $m_k \to \infty$ as $k \to \infty$, we have $\Phi_{(g,h)}^N \to 0$ as $N \to \infty$ a.e. $\mathbb{P}_\mu^\pi$.

From Assumption 4.2, and the above two conclusions, we can find $\Omega_0 \in \mathcal{F}$ with $\mathbb{P}_\mu^\pi(\Omega_0) = 1$ such that for any $\omega \in \Omega_0$, $\{\Phi_N^\omega[\varpi] : N \in \mathbb{N}\}$ is tight and $\Psi_f^N(\omega) \to 0, \Phi_{(g,h)}^N(\omega) \to 0$ as $N \to \infty$, for all $f \in \mathcal{S}(\mathbb{X})$ and all $(g, h) \in \mathcal{S}(\mathbb{X}) \times \mathcal{S}(\mathbb{A})$. Fix such $\omega \in \Omega_0$ and let $\{N_k : k \in \mathbb{N}\}$ be some

subsequence along which $\Phi^{\omega}_{N_k}[\varpi]$ converges weakly to some $\Phi \in \mathcal{P}(\mathbb{X}\times\mathbb{A})$. We now show that $\Phi = \theta_{q_0}$. The continuity of q_0 implies that $\psi_f \in \mathcal{C}_b(\mathbb{X})$ for $f \in \mathcal{C}_b(\mathbb{X})$, and so

$$\int_{\mathbb{X}\times\mathbb{A}} f(x)\Phi^{\omega}_{N_k}[\varpi](dxda) - \int_{\mathbb{X}\times\mathbb{A}} \psi_f(x)\Phi^{\omega}_{N_k}[\varpi](dxda) \to \int_{\mathbb{X}} f(x)\Phi^{(1)}(dx) - \int_{\mathbb{X}} \psi_f(x)\Phi^{(1)}(dx)$$

where $\Phi^{(1)}$ is, as before, the first marginal of Φ. Therefore, for any $f \in \mathcal{S}(\mathbb{X})$,

$$\int_{\mathbb{X}} f(x)\Phi^{(1)}(dx) = \int_{\mathbb{X}} \psi_f(x)\Phi^{(1)}(dx) = \int_{\mathbb{X}}\int_{\mathbb{X}} f(y)\varrho_{q_0}(x,dy)\Phi^{(1)}(dx).$$

By Assumption 2.2, $\Phi^{(1)} = \lambda_{q_0}$. Similarly, for $g \in \mathcal{S}(\mathbb{X})$ and $h \in \mathcal{S}(\mathbb{A})$,

$$\begin{aligned}&\int_{\mathbb{X}\times\mathbb{A}} g(x)h(a)\Phi^{\omega}_{N}[\varpi](dxda) - \int_{\mathbb{X}\times\mathbb{A}} \phi_{(g,h)}(x)\Phi^{\omega}_{N}[\varpi](dxda)\\ &\to \int_{\mathbb{X}\times\mathbb{A}} g(x)h(a)\Phi(dxda) - \int_{\mathbb{X}} \phi_{(g,h)}(x)\Phi^{(1)}(dx).\end{aligned}$$

Hence

$$\int_{\mathbb{X}\times\mathbb{A}} g(x)h(a)\Phi(dxda) = \int_{\mathbb{X}} \phi_{(g,h)}(x)\Phi^{(1)}(dx) = \int_{\mathbb{X}} \phi_{(g,h)}(x)\lambda_{q_0}(dx) = \int_{\mathbb{X}\times\mathbb{A}} g(x)h(a)\theta_{q_0}(dxda).$$

Recalling that $\{g\otimes h : g \in \mathcal{S}(\mathbb{X}), h \in \mathcal{S}(\mathbb{A})\}$ is separating in $(\mathbb{X}\times\mathbb{A}, \mathcal{B}(\mathbb{X})\otimes\mathcal{B}(\mathbb{A}))$, we have $\Phi = \theta_{q_0}$. Consequently, $\Phi^{\omega}_{N}[\varpi]$ converges weakly to θ_{q_0} as $N\to\infty$ a.e. ω $[\mathbb{P}^{\pi}_{\mu}]$. ■

Similar to Section 4.1, we will now construct a sequence of $\mathbb{X}\times\mathbb{A}$ valued random variables $Z \equiv (\bar{X}_t, \bar{A}_t)_{t\in\mathbb{N}_0}$ on a suitable probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that: (i) $\bar{X}_0$ has probability law μ, (ii) the probability law of Z corresponds to a controlled system associated with a policy $\pi \in \Pi_{\text{ATS}}(q,\mu)$, and (iii) consistent estimation of $\mathcal{J}_f$ can be achieved using the sequence Z. The sequence will be obtained by piecing together suitable sequences $(\xi^r_k, \zeta^r_k; k, r \in \mathbb{N}_0)$, $(\bar{\xi}^r_k, \bar{\zeta}^r_k; k, r \in \mathbb{N}_0)$ of $\mathbb{X}\times\mathbb{A}$ valued random variables. To construct these sequences we proceed recursively in r. Let $m_r = -\log(\varepsilon_r)$ for $r \in \mathbb{N}$, where $\{\varepsilon_r\}_{r\in\mathbb{N}}$ is as in Section 4.1, and set $m_0 = 0$.

Case $r = 1$**:** Define $\{\xi^r_k, s^r_k, \zeta^r_k, (i^r[m,k])_{m=1,\cdots j(r)}\}$, $\alpha_r, \sigma_r, \varrho_r$ for $r = 1$ and $k \in \mathbb{N}_0$ exactly as in Section 4.1. For $k = 0, 1, \cdots m_r$, define $\mathbb{X}\times\mathbb{A}$ valued random variables, $(\bar{\xi}^r_k, \bar{\zeta}^r_k)$ recursively in k, by setting $(\bar{\xi}^r_0, \bar{\zeta}^r_0) = (\xi^r_{\varrho_r}, \zeta^r_{\varrho_r})$ and through the following two equations

$$\bar{\mathbb{P}}\left(\bar{\xi}^r_k \in C \mid \hat{\mathcal{G}}^r_{k-1}\right) = \mathcal{Q}\left((\bar{\xi}^r_{k-1}, \bar{\zeta}^r_{k-1}), C\right), \tag{4.26}$$

$$\bar{\mathbb{P}}\left(\bar{\zeta}^r_k \in D \mid \hat{\mathcal{G}}^r_{k-1}, \bar{\xi}^r_k\right) = q_0(\bar{\xi}^r_k, D), \tag{4.27}$$

where for $k = 0, 1, \cdots m_r - 1$, $\hat{\mathcal{G}}^r_k = \mathcal{G}^r_{\varrho_r} \vee \sigma\{(\bar{\xi}^r_j, \bar{\zeta}^r_j), j = 0, 1, \cdots k\}$ and $\mathcal{G}^r_k$ is as in Section 4.1.

Case $r > 1$**:** Definition of $\{\xi^r_k, \zeta^r_k, s^r_k, (i^r[m,k])_{m=1,\cdots j(r)}\}_{k\geq 0}$ and σ_r, α_r for $r > 1$, is given exactly as in Section 4.1 through (4.11) – (4.16), in a recursive fashion, but with ϱ_r defined as

$$\varrho_r = \sigma_r + \varrho_{r-1} + m_{r-1} \tag{4.28}$$

and by setting

$$(\xi_0^r, \zeta_0^r) = (\bar{\xi}_{m_{r-1}}^{r-1}, \bar{\zeta}_{m_{r-1}}^{r-1}),\ i^r[m,0] = 0.$$

The sequence $(\bar{\xi}_k^r, \bar{\zeta}_k^r)$, for $k = 0, 1, \cdots m_r$, is defined exactly as for the case $r = 1$ through equations (4.26)-(4.27) (and by setting $(\bar{\xi}_0^r, \bar{\zeta}_0^r) = (\xi_{\varrho_r}^r, \zeta_{\varrho_r}^r)$).

Finally, the sequence $(\bar{X}_k, \bar{A}_k)$ is now constructed as follows. Recall that $\varrho_0 = 0$.

$$(\bar{X}_k, \bar{A}_k) = \begin{cases} (\bar{\xi}_{k-\varrho_r}^r, \bar{\zeta}_{k-\varrho_r}^r), & \text{whenever } \varrho_r \le k < \varrho_r + m_r,\ r \in \mathbb{N}. \\ \\ (\xi_{k-\varrho_r-m_r}^{r+1}, \zeta_{k-\varrho_r-m_r}^{r+1}), & \text{whenever } \varrho_r + m_r \le k < \varrho_{r+1},\ r \in \mathbb{N}_0. \end{cases} \tag{4.29}$$

The above sequence yields a $\pi \in \Pi$ and $\mathbb{P}_\mu^\pi \in \mathcal{P}(\Omega)$ as before. Consistent estimators for $\mathcal{J}_f$, $f \in C_b(\mathbb{X} \times \mathbb{A})$ can now be obtained as follows. Define on $(\bar{\Omega}, \bar{\mathcal{F}})$, a sequence of $\mathcal{P}(\mathbb{X} \times \mathbb{A})$ valued random variables, $\tilde{\Phi}_N$, $N \in \mathbb{N}$, as follows.

$$\tilde{\Phi}_N^\omega(F) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{m_k} \sum_{j=\varrho_k}^{\varrho_k + m_k - 1} 1_F(\bar{X}_j, \bar{A}_j),\ F \in \mathcal{B}(\mathbb{X} \times \mathbb{A}),\ \omega \in \bar{\Omega}.$$

The following is the main result of this section. The second part of the theorem says that for every $f \in C_b(\mathbb{X} \times \mathbb{A})$, $\int_{\mathbb{X} \times \mathbb{A}} f(x,a) \tilde{\Phi}_N(dxda)$ is an (a.e.) consistent estimator for $\mathcal{J}_f$. The proof is very similar to that of Theorem 4.1 and so only a sketch will be provided.

Theorem 4.2. *The policy constructed above is in $\Pi_{ATS}(q, \mu)$. Furthermore, as $N \to \infty$*

$$\|\tilde{\Phi}_N^\omega - \theta_{q_0}\|_{BL} \to 0, \quad a.e.\ \bar{\mathbb{P}}.$$

Proof. As in Theorem 4.1, we define $\bar{\Omega}_0$, $\bar{p}_k^\omega(D|x)$, Λ_k, $n_{ti}(m_1, m_2)$, and $\mu_{ti}(m_1, m_2)$. To show $\pi \in \Pi_{\text{ATS}}(q, \mu)$, it suffices to show that for all $\omega \in \bar{\Omega}_0$ and compact $K \subset \mathbb{X}$, (4.18) holds. Fix such a ω and K. As in Theorem 4.1, we can find $r_0 \in \mathbb{N}$, such that (4.19) and (4.20) hold for all $r \ge r_0$ and $K \subset K_r$. Fix $\beta_0 \ge r_0 + 1$ and let $\beta \in \mathbb{N}, \beta \ge \beta_0$ such that $\varrho_\beta < k \le \varrho_{\beta+1}$. Also fix $x \in K$ and let $i \in \{1, \ldots, j(\beta)\}$ be such that $x \in \tilde{B}_{\beta i}$. Similar to (4.21), we can write

$$\bar{p}_k(\cdot|x) = \begin{cases} l^{-1}(l_1\tau_1 + l_2\tau_2 + l_3\tau_3 + l_4\tau_4), & \text{whenever } \varrho_\beta \le k < \varrho_\beta + m_\beta. \\ \\ \check{l}^{-1}(l_1\tau_1 + l_2\tau_2 + l_3\tau_3 + l_5\tau_5 + l_6\tau_6), & \text{whenever } \varrho_\beta + m_\beta \le k < \varrho_{\beta+1}. \end{cases} \tag{4.30}$$

Here

$$\tau_1 = \mu_{\beta,i}[0, \varrho_{\beta-1}],\ \tau_2 = \mu_{\beta,i}[\varrho_{\beta-1}, \varrho_{\beta-1} + m_{\beta-1}],\ \tau_3 = \mu_{\beta,i}[\varrho_{\beta-1} + m_{\beta-1}, \varrho_\beta],$$
$$\tau_4 = \mu_{\beta,i}[\varrho_\beta, k],\ \tau_5 = \mu_{\beta,i}[\varrho_\beta, \varrho_\beta + m_\beta],\ \tau_6 = \mu_{\beta,i}[\varrho_\beta + m_\beta, k]$$

and

$$l_1 = n_{\beta i}(0, \varrho_{\beta-1}),\ l_2 = n_{\beta i}(\varrho_{\beta-1}, \varrho_{\beta-1} + m_{\beta-1}),\ l_3 = n_{\beta i}(\varrho_{\beta-1} + m_{\beta-1}, \varrho_\beta),$$
$$l_4 = n_{\beta i}(\varrho_\beta, k),\ l_5 = n_{\beta i}(\varrho_\beta, \varrho_\beta + m_\beta),\ l_6 = n_{\beta i}(\varrho_\beta + m_\beta, k),$$
$$l = l_1 + l_2 + l_3 + l_4,\ \check{l} = l_1 + l_2 + l_3 + l_5 + l_6.$$

Analogous to $\tilde{\nu}_3$ in Theorem 4.1, we can define a $\tilde{\tau}_6 \in \mathcal{P}(\mathbb{A})$ such that, if $n_6 > 0$,

$$\|\tilde{\tau}_6 - q(\cdot|x)\|_{\text{BL}} \leq 3\epsilon, \ \|\tau_6 - \tilde{\tau}_6\|_{\text{BL}} \leq \frac{4\ell(\beta+1)}{l_6}.$$

Now let $\tilde{\tau}_1 = \tilde{\tau}_2 = \tilde{\tau}_3 = \tilde{\tau}_4 = \tilde{\tau}_5 = \eta_\beta(q(\cdot|\tilde{b}_\beta(x)))$. Then, by our choice of r_0,

$$\begin{aligned} &\|q(\cdot|x) - l^{-1}(l_1\tilde{\tau}_1 + l_2\tilde{\tau}_2 + l_3\tilde{\tau}_3 + l_4\tilde{\tau}_4)\|_{\text{BL}} \leq 2\epsilon \text{ when } \varrho_\beta \leq k < \varrho_\beta + m_\beta. \\ &\|q(\cdot|x) - \check{l}^{-1}(l_1\tilde{\tau}_1 + l_2\tilde{\tau}_2 + l_3\tilde{\tau}_3 + l_5\tilde{\tau}_5 + l_6\tilde{\tau}_6)\|_{\text{BL}} \leq 3\epsilon \text{ when } \varrho_\beta + m_\beta \leq k < \varrho_{\beta+1}. \end{aligned} \tag{4.31}$$

Also note that

$$\|\tau_3 - \tilde{\tau}_3\|_{BL} \leq \frac{4\ell(\beta)}{l_3}.$$

When $\varrho_\beta \leq k < \varrho_\beta + m_\beta$, we have

$$\begin{aligned} \|\bar{p}_k(\cdot|x) - q(\cdot|x)\|_{BL} &\leq 2\epsilon + l^{-1}(2l_1 + 2l_2 + 4\ell(\beta) + 2l_4) \\ &\leq 2\epsilon + \alpha_\beta^{-1}(2\varrho_{\beta-1} + 2m_{\beta-1} + 4\ell(\beta) + 2m_\beta) \\ &\leq 2\epsilon + \varepsilon_\beta + \frac{4m_\beta}{\alpha_\beta}. \end{aligned} \tag{4.32}$$

When $\varrho_\beta + m_\beta \leq k < \varrho_{\beta+1}$,

$$\begin{aligned} \|\bar{p}_k(\cdot|x) - q(\cdot|x)\|_{BL} &\leq 3\epsilon + \check{l}^{-1}(2l_1 + 2l_2 + 4\ell(\beta) + 2l_5 + 4\ell(\beta+1)) \\ &\leq 3\epsilon + \alpha_\beta^{-1}(2\varrho_{\beta-1} + 2m_{\beta-1} + 4\ell(\beta) + 2m_\beta + 4\ell(\beta+1)) \\ &\leq 3\epsilon + \varepsilon_\beta + \frac{4m_\beta}{\alpha_\beta}. \end{aligned} \tag{4.33}$$

Recalling that $m_\beta = -\log(\varepsilon_\beta)$ and that $\alpha_\beta \geq \varepsilon_\beta^{-1}$, $\frac{4m_\beta}{\alpha_\beta} \to 0$ as $\beta \to \infty$. Since $\beta \to \infty$ as $k \to \infty$, we have that $\|\bar{p}_k(\cdot|x) - q(\cdot|x)\|_{BL} \to 0$ as $k \to \infty$ and therefore $\pi \in \Pi_{\text{ATS}}(q, \mu)$. Finally, the second part of the theorem is an immediate consequence of Lemma 4.2. ■

4.3. Adaptive Control.

In this section we consider a setting where the (near) optimal $q \in \Pi_{\text{SMC}}$ is not known but there are available sampling schemes that allow for consistent estimation of q. The goal is then to estimate q dynamically and use the estimators of q to construct a control policy for which the associated pathwise cost per unit time coincides with that for q.

In order to give a precise formulation, suppose that $q \in \Pi_{\text{SMC}}$ is given as

$$q(\cdot \mid x) = q(\cdot \mid \kappa_0, x), \tag{4.34}$$

where κ_0 is an unknown parameter taking values in some compact metric space Γ. We assume that the map $(\kappa, x) \mapsto q(\cdot \mid \kappa, x)$, from $\Gamma \times \mathbb{X} \to \mathcal{P}(\mathbb{A})$, is a continuous function. Also suppose that there is a $q_0 \in \Pi_{\text{SMC}}$ and a continuous function $G : \mathcal{P}(\mathbb{X} \times \mathbb{A}) \to \Gamma$ such that

$$G(\theta_{q_0}) = \kappa_0.$$

This relationship, in view of Lemma 3.1, says that as $N \to \infty$, $G(\Phi_N)$ is an (a.e.) consistent estimator for κ_0, under $\mathbb{P}_\mu^{q_0}$ for all $\mu \in \mathcal{P}(\mathbb{X})$. However the corresponding pathwise cost is $\int_{\mathbb{X}\times\mathbb{A}} c(x,a)\theta_{q_0}(dxda)$ ($\mathbb{P}_\mu^{q_0}$ a.e.) and thus although the policy q_0 achieves the goal of parameter estimation, it does not meet the criterion of cost (near) optimization. In order to meet both objectives we will now construct a policy π which uses dynamic estimators for κ_0 (and consequently for q) for control decisions and is such that it is an ATS policy for q corresponding to the initial condition μ.

Let $\{\tilde{\Lambda}_k, \tilde{\mathbb{X}}_k, \tilde{b}_k, \Lambda'_k, \mathbb{A}_k, b'_k, \tilde{\Lambda}^0_k, \eta_k\}_{k\in\mathbb{N}}$ be as in Section 4.1. Let m_r be as in Section 4.2. As in Section 4.2 we begin by introducing sequences $(\xi^r_k, \zeta^r_k; k, r \in \mathbb{N}_0)$, $(\bar{\xi}^r_k, \bar{\zeta}^r_k; r \in \mathbb{N}_0, k = 1, \cdots m_r)$ of $\mathbb{X} \times \mathbb{A}$ valued random variables, recursively in r. We will use notation and constructions from Sections 4.1 and 4.2.

Case $r = 1$**:** Set $\hat{q}_r = q_0$. For $m = 1, \cdots j(r)$, define

$$\hat{q}^{r,m}_r(\cdot) = \hat{q}_r(\cdot \mid x_{rm}),\ \tilde{q}^{r,m}_r = \eta_r(\hat{q}^{r,m}_r),\ m = 1, \cdots j(r). \tag{4.35}$$

Abusing notation from Section 4.1, denote the i-th component of $\Psi(\tilde{q}^{r,m}_r)$ by $e^r[m, i]$, i.e.

$$\Psi(\tilde{q}^{r,m}_r) = (e^r[m, 1], e^r[m, 2], \cdots). \tag{4.36}$$

With this new definition of $e^r[m, i]$, the definition of $\{\xi^r_k, s^r_k, \zeta^r_k, (i^r[m, k])_{m=1,\cdots j(r)}\}$, for $r = 1$ and $k \in \mathbb{N}_0$ is given exactly as in Section 4.1, through equations (4.11) – (4.14). Also define $\alpha_r, \sigma_r, \varrho_r$ through equations (4.15) – (4.17) (with $\varrho_0 = 0$). Next, for $t = 0, 1, \cdots, m_r$, define $\mathbb{X} \times \mathbb{A}$ valued random variables $(\bar{\xi}^r_t, \bar{\zeta}^r_t)$, $t = 0, 1, \cdots m_r$, recursively in t, by (4.26) – (4.27) (and by setting $(\bar{\xi}^r_0, \bar{\zeta}^r_0) = (\xi^r_{\varrho_r}, \zeta^r_{\varrho_r})$). Define a $\mathcal{P}(\mathbb{X} \times \mathbb{A})$ valued random variable $\tilde{\Phi}_r$ by the relation

$$\tilde{\Phi}_r(F) = \frac{1}{m_r} \sum_{t=1}^{m_r} 1_F(\bar{\xi}^r_t, \bar{\zeta}^r_t),\ F \in \mathcal{B}(\mathbb{X} \times \mathbb{A}).$$

and let $\kappa_r = G(\tilde{\Phi}_r)$.

Case $r > 1$**:** Set $\hat{q}_r(\cdot \mid x) = q(\cdot \mid \kappa_{r-1}, x)$, $x \in \mathbb{X}$. Define for $m = 1, \cdots j(r)$, $\hat{q}^{r,m}_r$ and $\tilde{q}^{r,m}_r$, , through (4.35); and $e^r[m, i]$, $i \in \mathbb{N}$, through (4.36). With this definition of $e^r[m, i]$, the definition of $\{\xi^r_k, s^r_k, \zeta^r_k, (i^r[m, k])_{m=1,\cdots j(r)}\}$, for $k \in \mathbb{N}_0$ and $\alpha_r, \sigma_r, \varrho_r$ is given as in Sections 4.1 and 4.2, through equations (4.11) – (4.16) and (4.28). The sequence $(\bar{\xi}^r_k, \bar{\zeta}^r_k)$, for $k = 0, 1, \cdots m_r$, is defined exactly as for the case $r = 1$ through equations (4.26)-(4.27) (and by setting $(\bar{\xi}^r_0, \bar{\zeta}^r_0) = (\xi^r_{\varrho_r}, \zeta^r_{\varrho_r})$). To complete the recursion we define

$$\tilde{\Phi}_r(F) = \frac{1}{M_{r-1} + m_r} \left(M_{r-1}\tilde{\Phi}_{r-1}(F) + \sum_{t=1}^{m_r} 1_F(\bar{\xi}^r_t, \bar{\zeta}^r_t) \right),\ F \in \mathcal{B}(\mathbb{X} \times \mathbb{A}),$$

where $M_{r-1} = \sum_{t=1}^{r-1} m_t$, and let $\kappa_r = G(\tilde{\Phi}_r)$.

The definition of the sequence $(\bar{X}_k, \bar{A}_k)$ is now given through (4.29). This sequence yields a $\pi \in \Pi$ and $\mathbb{P}_\mu^\pi \in \mathcal{P}(\Omega)$ as before.

The following is the main result of the section. Assumption 4.2 will be taken to hold. The proof is similar to that of Theorems 4.1 and 4.2 and so only a sketch will be provided.

Theorem 4.3. *The policy constructed above is in $\Pi_{ATS}(q, \mu)$. Furthermore, for every compact K in $\mathbb{X}$, as $r \to \infty$*

$$\sup_{x \in K} \|\hat{q}_r(\cdot \mid x) - q(\cdot \mid x)\|_{BL} \to 0,$$

a.e. $\bar{\mathbb{P}}$.

Proof. We use the same notation and definitions as in the proof of Theorem 4.2. First, we show that, for every compact set $K \subset \mathbb{X}$,

$$\sup_{x \in K} \|\hat{q}_r(\cdot \mid x) - q(\cdot \mid x)\|_{\mathrm{BL}} \to 0 \text{ a.e. } \bar{\mathbb{P}}. \tag{4.37}$$

By Theorem 3.1, $\tilde{\Phi}_r$ converges weakly to θ_{q_0} a.e. $\bar{\mathbb{P}}$. Since G is continuous, $\kappa_r = G(\tilde{\Phi}_r) \to G(\theta_{q_0}) = \kappa_0$ as $r \to \infty$, a.e. $\bar{\mathbb{P}}$. Note that

$$\|\hat{q}_r(\cdot \mid x) - q(\cdot \mid x)\|_{\mathrm{BL}} = \|q(\cdot \mid \kappa_r, x) - q(\cdot \mid \kappa, \tilde{b}_r(x))\|_{\mathrm{BL}}.$$

Equation (4.37) is now an immediate consequence of the continuity of the map $(\kappa, x) \mapsto q(\cdot \mid \kappa, x)$.

For $\epsilon > 0$, choose r_0 such that all $r > r_0, K \subset K_r$, (4.20) holds,

$$\sup_{(x,\kappa) \in K \times \Gamma} \|q(\cdot \mid \kappa, x) - q(\cdot \mid \kappa, \tilde{b}_r(x))\|_{\mathrm{BL}} \le \epsilon, \tag{4.38}$$

and

$$\sup_{x \in K} \|\hat{q}_r(\cdot \mid x) - q(\cdot \mid x)\|_{\mathrm{BL}} \le \epsilon. \tag{4.39}$$

Fix $\beta_0 > r_0 + 1$ and let $\beta \in \mathbb{N}$, $\beta \ge \beta_0$ be such that $\varrho_\beta < k \le \varrho_{\beta+1}$.

Let $l, \check{l}, l_i, \tau_i, i = 1, 2, \ldots, 6$, and $\bar{p}_k(\cdot|x)$ be the same as in the proof of Theorem 4.2. In particular, we have that (4.30) holds. Let $\tilde{\tau}_1 = \tilde{\tau}_2 = \tilde{\tau}_3 = \tilde{\tau}_4 = \tilde{\tau}_5 = \eta_\beta(\hat{q}_\beta(\cdot|\tilde{b}_\beta(x)))$. Construct $\tilde{\tau}_6$ in the same way as in Theorem 4.2 with $q(\cdot|\tilde{x})$ replaced by $\hat{q}_{\beta+1}(\cdot|\tilde{x})$ for $\tilde{x} \in \mathbb{X}$. Using (4.38) and (4.39) it is now easily checked that (4.31) holds with 2ϵ and 3ϵ, replaced by 3ϵ and 4ϵ respectively. Also note that, if $l_6 > 0$,

$$\|\tau_3 - \tilde{\tau}_3\| \le \frac{4\ell(\beta)}{l_3}, \quad \|\tau_6 - \tilde{\tau}_6\| \le \frac{4\ell(\beta+1)}{l_6}.$$

Rest of the proof now follows as for Theorem 4.2. ■

5. Discussion.

In this section we comment on the usefulness of the results obtained in this work. To see the flexibility that ATS policies offer let's first consider the countable setting. Consider the elementary model where $\mathbb{X}$ is a singleton and $\mathbb{A}$ is a finite set. A Markov control in this setting is just a single probability measure on $\mathbb{A}$ and the long term cost for a typical one stage cost function $c : \mathbb{A} \to [0, \infty)$ under q, by the strong law of large numbers is $c_q = \int_{\mathbb{A}} c(a)q(da)$. Also, the corresponding asymptotic mean square error:

$$\lim_{N\to\infty} \mathbb{E}^q \left(\frac{1}{N} \sum_{k=0}^{N-1} c(A_k) - \int_{\mathbb{A}} c(a)q(da) \right)^2 = \frac{\sigma^2}{N},$$

where $\sigma^2 = \int_{\mathbb{A}} (c(a) - c_q)^2 q(da)$. It is easy to see that one can construct an ATS policy π for the Markov control q (cf. Lemma 4.1) under which, for some $\alpha(c) \in (0, \infty)$

$$\left| \frac{1}{N} \sum_{k=0}^{N-1} c(A_k) - \int_{\mathbb{A}} c(a)q(da) \right| \le \frac{\alpha(c)}{N}, \ \mathbb{P}^\pi \text{ a.e.}$$

and thus the asymptotic mean square error under π is $\frac{\alpha^2(c)}{N^2}$.

The above simple example illustrates how ATS policies can be used to develop variance reduction schemes for ergodic control problems. Additionally, ATS policies provide much flexibility for sampling (namely using controls without regards to the ensuing cost), for example for the purpose of collecting information. This could be information which is related to the main optimization objective, but could also be other information which is of interest. Consider, for example, the following elementary setting. Suppose that $\mathbb{X} = \{-1, 0, 1\}$ and $\mathbb{A} = \{a, b\}$. Suppose that the one stage cost function is given as

$$c(\pm 1, a) = c(\pm 1, b) = 0, \ c(0, a) = 1, \ c(0, b) = 2$$

and the transition probability kernel is defined as

$$\mathcal{Q}((0, a), \cdot) = \frac{1}{2}\delta_{\{1\}}(\cdot) + \frac{1}{2}\delta_{\{-1\}}(\cdot); \qquad \mathcal{Q}((0, b), \cdot) = \beta\delta_{\{1\}}(\cdot) + (1 - \beta)\delta_{\{-1\}}(\cdot);$$
$$\mathcal{Q}((x, a), \cdot) = \frac{1}{2}\delta_{\{-x\}}(\cdot) + \frac{1}{2}\delta_{\{0\}}(\cdot); \qquad \mathcal{Q}((x, b), \cdot) = (1 - \gamma)\delta_{\{-x\}}(\cdot) + \gamma\delta_{\{0\}}(\cdot), \ x = \pm 1,$$

where $0 < \beta, \gamma < 1$. If our goal is the minimize to average cost, then we prefer to stay at states ± 1, and so we should use a at all states if $\gamma > \frac{1}{2}$, but use b at states ± 1 if $\gamma < \frac{1}{2}$. Thus, from the optimization point of view, we need to find γ but β is irrelevant.

Since the probability that $X_t = 1$ is bounded below for $t > 1$, consider the following estimation procedure for γ. Action b will be used at time t if $X_t = \pm 1$ and in addition $t = 10^n$ for some integer n. Let

$$\hat{\gamma}_t = \frac{\sum_{m=1}^n 1\{X_{10^m} = \pm 1, X_{10^m+1} = 0\}}{\sum_{m=1}^n 1\{X_{10^m} = \pm 1\}} \quad \text{for } 10^n < t \le 10^{n+1} \ . \tag{5.1}$$

This estimator converges to γ a.s. under any policy that is consistent with the above requirement for time instants $t = 10^n$, $n \in \mathbb{N}$. In particular, we can now choose the following policy: At $t \neq 10^n$ use b iff $X_t = \pm 1$ and $\hat{\gamma}_t < \frac{1}{2}$. It then follows that there is some (random) time so that, at all later times, the optimal policy is used. It is easy to check that the above recipe defines an implementable ATS policy for the (unknown) optimal stationary policy and is thus optimal as well. Furthermore, we can modify the above policy slightly to define a new ATS policy that also delivers estimates for β, with no effect on the cost. This is done similarly to above: use action b at state 0 at time t if $X_t = 0$ and in addition $t = 10^n$ for some integer n. An estimator as in (5.1) will be consistent. Since the number of time points where b is used increases logarithmically it is easy to see that the limits in (1.1) are not affected, and consequently the limiting cost does not change.

The above example illustrates the use of ATS policies for estimation and adaptive control for a rather elementary setting. However similar ideas are applicable for general state and action space models as well. In Section 4.2 we showed how ATS policies introduced in Section 3 of this work can be used for estimation of unknown model parameters and in Section 4.3 we described how they can be used for adaptive control problems as well.

References

[1] R. Agrawal, D. Teneketzis and V. Anantharam, "Asymptotically efficient adaptive allocation rules for controlled Markov chains: finite parameter space," *IEEE Trans. Auto. Control* **34** pp. 1249–1259, 1989.

[2] A. Altman and A. Shwartz, "Markov decision problems and state-action frequencies," *SIAM J. Control Opt.* **29** pp. 786–809, 1991.

[3] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. K. Ghosh and S. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*," **31**, pp. 282-344, 1993.

[4] V. S. Borkar, "On the Milito-Cruz adaptive control scheme for Markov chains," *J. Opt. Theory Appl.* **77** pp. 387–398 (1993).

[5] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for Markov decision processes," *Math. Oper. Res.* **22** pp. 222–255, 1997.

[6] N. Cesa-Bianchi and G. Lugosi *Prediction, learning and games*, Cambridge University Press Cambridge, UK 2006.

[7] T. E. Duncan, B. Pasik-Duncan and L. Stettner, "Adaptive control of discrete time Markov processes by the large deviations method," Appl. Math. **27** pp. 265–285, 2000.

[8] K. Dyagilev, S. Mannor and N. Shimkin, "Efficient Rinforcement Learning in Parameterized Models: Discrete Parameter Case," *8th European Workshop on Reinf. Learning, LNAI* **5323** pp. 41–54, 2008.

[9] E. I. Gordienko and J. A. Minjarez-Sosa, "Adaptive control for discrete-time Markov processes with unbounded costs: Average criterion," *Math. Methods Oper, Res.* **48** pp. 37–55, 1998.

[10] J. A. Minjarez-Sosa, "Empirical estimation in average Markov control processes," *Appl. Math. Letters* **21** pp. 459-464 (2008).

[11] S. M. Ross, *Inroduction to probability models*, 9th edition, Academic Press, Orlando, USA, 2006.
[12] N. Shimkin and A. Shwartz, "Asymptotically efficient adaptive strategies in repeated games II: asymptotic optimality," *Math. Oper. Res.* **21** pp. 487–512, 1996.
[13] A. N. Shiryaev, *Probability*, Graduate Texts in Mathematics, 95, Second Edition, Springer-Verlag, New York, 1996.

A. Budhiraja and X. Liu
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599, USA
email: budhiraj@email.unc.edu, xinliu@unc.edu

A. Shwartz
The Julius M. and Bernice Naiman Chair in Engineering
Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel
email: adam@ee.technion.ac.il