Identifiability and Adaptive Control of Markov Chains

Ben Pence
Scott Moura

December 12, 2009

EECS 558 – Stochastic Control
Professor Teneketzis
University of Michigan, Ann Arbor

1. Introduction

This project reviews three papers from the identification and adaptive control literature for Markov chain models. The reviewed papers are by Mandl [1] and Borkar & Varaiya [2,3]. Each paper addresses the problem of controlling a Markov chain in which the transition probability matrices depend on an unknown parameter vector. A key issue in adaptive control, as discussed in the reviewed papers, is the "dual control" problem – that is, when to apply controls to improve performance and when to apply controls to improve parameter estimation. To motivate why this tradeoff is important, *this reports focuses on the concept of "identifiability"*.

The pioneering work for identification and adaptive control of Markov chains was developed by Mandl [1]. This paper proves that, under certain assumptions including the identifiability condition, the sequence of maximum likelihood parameter estimates converges to the true parameter. In addition, the mean difference between the estimates in this sequence and the true parameters is bounded for all time. Then in the limit, the optimal control based on the parameter estimates converges almost surely to the optimal control based on knowing the true parameter *a priori*. Thus the limiting performance of the Markov chain based on the adaptive control policy is optimal. This paper also derives asymptotic uncertainty measures for both the parameter estimates and performance of the Markov chain.

A key assumption in Mandl [1] is a so called identifiability condition. Borkar and Varaiya [2,3] examine the case for which this assumption is relaxed. The first paper by Borkar and Varaiya [2] proves that for finite state controlled Markov chains, the sequence of maximum likelihood parameter estimates converges in finite time to a limit point in the set of possible parameters. Then the resulting performance of the Markov chain with the true parameter matches the performance of a Markov chain with the estimated parameter. The maximum likelihood estimate, however, is not guaranteed to converge to the true parameter value. Neither is the resulting control policy guaranteed to be optimal. This report discusses the reasons why this result occurs, both mathematically and intuitively.

The last paper [3] addresses the inability to identify the true parameter without presuming identifiability. Their solution introduces two methods for randomizing the control policies. Under these randomized control policies the maximum likelihood estimates are guaranteed to converge to the true parameters, and the limiting performance of the Markov chain is shown to be optimal. This result requires a much weaker form of identifiability.

This review is formatted as follows. Section 2 discusses identification and adaptive control for Markov chains given that the identifiability assumption holds. Section 3 reviews the case in which the identifiability condition is relaxed and discusses key results and limitations. Section 4 discusses the control randomization policies and the resulting performance guarantees. Finally, Section 5 provides conclusions and recommendations.

2. Identification and Adaptive Control

This section reviews identification and adaptive control of Markov chains. The majority of the results in this section are due to Mandl [1]. The model considered is summarized as follows:

<u>The Model</u>

- Controlled Markov chain (MC)
- Finite state space $I = \{1, 2, \ldots, r\}$
- Compact and finite action space U
- Compact and finite parameter space A
- Transition probabilities $\{p(i,k;u,\alpha),\ \ i,k \in I,\ \ u \in U,\ \ \alpha \in A\}$
- Stationary Markov policies $U_n = g(X_n, \alpha), \quad \alpha \in A$
- The estimator has perfect recall. That is, the estimator remembers the entire past history of states and controls
- Instantaneous cost $c(i,k;u,\alpha),\ \ i,k \in I$

This model is identical to the systems typically considered in class, with the exception that the transition probabilities depend on the unknown parameter $\alpha \in A$. The objective is to identify the true parameter value $\alpha^0 \in A$ based on clean observations of the states and controls. Moreover, we seek to use the parameter estimates to adaptively adjust the control policy. First we discuss results for identification only, and then consider performance when adaptive control is applied.

2.1 Identification of Markov Chains

The control policy is a function of the unknown parameter α. Thus, before applying the control policy the parameter α must be estimated. The paper by Mandl [1] considers the general case of minimum contrast estimation whereas the papers by Borkar and Varaiya [2,3] consider a subclass of minimum contrast estimators called maximum likelihood estimators. The maximum likelihood estimator is defined as follows: α_n^* is the maximum likelihood estimate of α_0, the true but unknown parameter, at time n if it satisfies

$$\begin{aligned}\alpha_n^* &= \underset{\alpha}{\operatorname{argmax}}\{\mathrm{Prob}(\mathrm{x}_0, \cdots, \mathrm{x}_\mathrm{n} | \mathrm{x}_0, \mathrm{u}_0, \cdots, \mathrm{u}_{\mathrm{n}-1}, \alpha_n)\} \\ &= \underset{\alpha}{\operatorname{argmax}}\left\{\prod_{m=0}^{n-1} p(i,k;u_m,\alpha_m)\right\}\end{aligned}$$

Given the contrast function: $-\log\big(p(i,k;u_n,\alpha_n)\big)$, the minimum contrast estimator described in Mandl [1] is equivalent to the maximum likelihood estimator described above.

The interpretation of the maximum likelihood estimator can be summarized as follows: *the maximum likelihood estimate of α_0 at time n is the value of α_n that maximizes the probability of transitioning from* x_0 *to* x_1*, from* x_1 *to* x_2*, …, and from* $\mathrm{x}_{\mathrm{n}-1}$ *to* x_n *(as depicted in Fig. 1)* *conditioned on the initial state of the Markov chain and all the control actions up to but not including time n.* Thus the maximum likelihood estimator is quite intuitive.

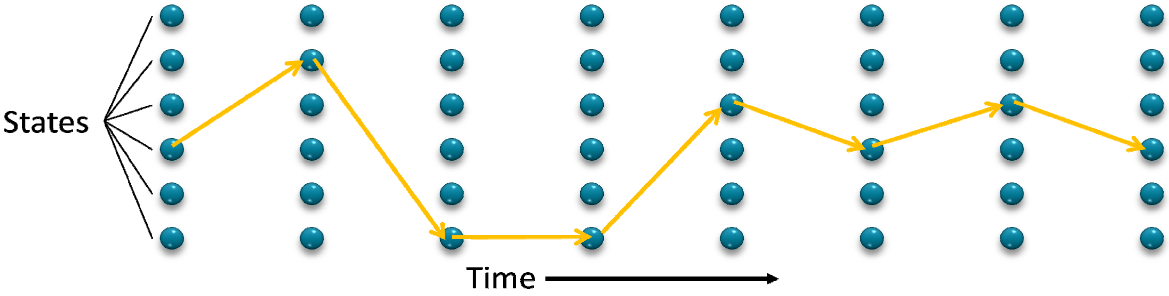

Fig 1. The maximum likelihood estimate is the value of α for which the state trajectory shown is most likely.

A key assumption by Mandl [1] is known as the *identifiability condition*:

Assumption A1:

For any $\alpha \neq \beta$ both in the parameter space A, there exists an $i \in I$ such that the following is true:

$$[p(i,1;u,\alpha), p(i,2;u,\alpha), \cdots, p(i,r;u,\alpha)] \neq [p(i,1;u,\beta), p(i,2;u,\beta), \cdots, p(i,r;u,\beta)] \; for\ all\ u \in U.$$

This key assumption can be interpreted to mean that *the transition matrix with parameter α is different from the transition matrix with parameter β for all admissible control actions.* Intuitively, this assumption appears to conjecture that if the transition matrix with parameter α is identical to the transition matrix with parameter β under any admissible control action, then the true parameter α_0 cannot be identified accurately, hence the label "identifiability condition". This conjecture was addressed by Borkar and Varaiya and will be discussed in Section 4 of this review. They showed that under control randomization, Assumption A1 is too restrictive and the conjecture doesn't hold.

A second important assumption for both the work by Mandl and by Borkar and Varaiya is as follows:

Assumption A2:

The controlled Markov chain contains a single positive recurrent class.

Under A1, A2, and any stationary control $u \in U$, Mandl proved the following important result:

Theorem 1:

The sequence $\{\alpha_n^*\}$ of minimum contrast estimates (and hence maximum likelihood estimates) converges asymptotically to the true parameter α_0.

Theorem 1 establishes a method for accurately estimating the unknown parameter and consequently the transition probabilities of the Markov chain. The proof of Theorem 1 is beyond the scope of this review, and the interested reader is encouraged to consult the paper by Mandl [1].

In addition to establishing parameter convergence, Mandl [1] appeals to the central limit theorem to show the following result: U*nder certain assumptions (see [1]), the distribution of* $(\alpha_n^* - \alpha_0)\sqrt{n}$ *is asymptotically normal with zero mean and known covariance.* This result offers an uncertainty measurement on the parameter estimates α_n^* by specifying an asymptotic distribution of $(\alpha_n^* - \alpha_0)\sqrt{n}$. A formula for calculating the covariance matrix is provided in [1]. This suggests that a property of the minimum contrast (and maximum likelihood) estimator is that as $n \longrightarrow \infty$, the covariance matrix, and hence the uncertainty in the estimates, tends to zero.

Up to this point in the section, the emphasis has been toward the identification of uncontrolled Markov chains, and important results have been reviewed and discussed. The remainder of this section focuses on adaptive control of Markov chains.

2.2 Adaptive Control of Markov Chains

The control policies considered in this review are limited to stationary control policies $g \in \mathcal{G}_{\mathrm{SM}}$ where the control action $u_n = g(x_n, \alpha)$ is a function of the current state $x_n \in I$ and parameter $\alpha \in A$. For each $\alpha \in A$, assume there exists a pre-specified control policy that gives optimal performance with respect to the instantaneous cost described in the model summary. The instantaneous selection of the pre-specified control depends on the current estimate of α.

Define $g^0 \in \mathcal{G}_{\mathrm{SM}}$ to be the optimal stationary control given that the true parameter α_0 is known. Also define $\Theta(g^0)$ to be the expected stationary cost under the optimal policy g^0, i.e.

$$\Theta(g^0) \coloneqq E^{g^0}\{c(i,k; g^0(i,\alpha_0), \alpha_0)\}.$$

Given these definitions, the expected cost up to time n under the optimal strategy is

$$E^{g^0} C_n = n\Theta(g^0).$$

Then, Mandl [1] proves the following important result for the adaptive control of Markov chains:
Theorem 2:
Let the control action at time n *be* $u_n = g^*(x_n, \alpha_n^*)$ *where* $\alpha_n^* \in A$ *is the minimum contrast (or maximum likelihood) estimate of* α_0*. Here,* g^* *is the optimal stationary Markov policy under* α_n^**. Then given certain continuity assumptions (see [1]),*

$$\lim_{n\to\infty}\frac{1}{n}C_n = \Theta(g^0)\ a.s.$$

where g^0 is the optimal stationary Markov policy under α_0.

As a consequence of Theorem 2, in the limit, the performance of the Markov chain under the adaptive control converges to the optimal performance. The proof is constructed in Mandl [1].

Another important result by Mandl characterizes the asymptotic distribution of the performance of the adaptively controlled Markov chain: *Under certain assumptions (see [1]), $\frac{\left(C_n-n\Theta(g^0)\right)}{\sqrt{n}}$ is asymptotically normally distributed with zero mean and known covariance.* This result provides an uncertainty measure on the performance of the adaptively controlled Markov chain. Mandl [1] provides formulas for calculating the covariance in the performance.

Example 1:
To demonstrate Mandl's proposed identification and adaptive control scheme, we apply the method on the Markov chain illustrated in Fig. 2:

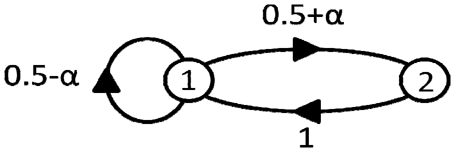

Fig. 2. The controlled Markov chain and transition probabilities

Respectively, the state, action, and parameter spaces are:

- $I = \{1,2\}$
- $U = \{1,2\}$
- $A = \{0.1,0.2,0.3\}$

and the state transition matrix is:

$$P_{ik}(u,\alpha) = \begin{bmatrix} 0.5-\alpha & 0.5+\alpha \\ 1 & 0 \end{bmatrix}$$

The true parameter $\alpha^0 = 0.2$. The control law, as a function of the state and parameter estimate is given by $g(i,0.1) = g(i,0.3) = 2$ and $g = (i,0.2) = 1$. Initially $x_0 = 1$ and $u_0 = 1$. The objective is to identify the true parameter and control the Markov chain as if the unknown parameter were known *a priori*.

Note that the system is identifiable since the transition probability matrix is different for each parameter value, for all possible control actions. As a result, we apply the adaptive control method and provide a sample simulation in Fig. 3. Observe that the parameter estimate in the third subplot converges to the true parameter $\alpha^0 = 0.2$. Moreover, once the parameter estimate converges, the adaptive control law applies the control action $g = (i, 0.2) = 1$ corresponding to if the true parameter were known *a priori*.

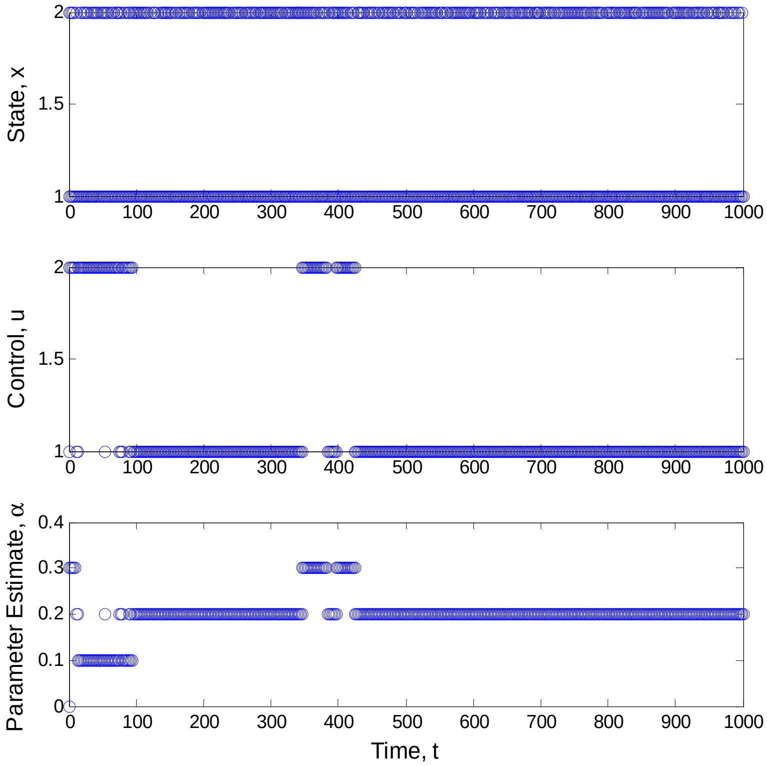

Fig. 3. Sample simulation for an identifiable Markov chain.

This section has reviewed adaptive control of Markov chains under the identifiability condition outlined in Assumption A1, mainly from the perspective of Mandl [1]. The subsequent sections review results for the case under which A1 is relaxed.

3. Relaxing the Identifiability Assumption A1

Borkar and Varaiya [2] claim that the "identifiability condition" (Assumption A1) may be too restrictive in some cases and examine cases in which A1 does not hold. They establish the following result for arbitrary stationary Markov policies $g(x_n, \alpha)$ with parameter limit points $\alpha^* \in A$:

Theorem 3:

There exists a random variable $\alpha^* \in A$, and a finite random time T such that $n \geq T$, the following hold:

$$\alpha_n^* = \alpha^*,$$
$$u_n = g(x_n, \alpha^*),$$

$$p(i,k;g(i,\alpha^*),\alpha^*) = p(i,k;g(i,\alpha^*),\alpha_0), \qquad for\ all\ i,k \in I.$$

Borkar and Varaiya in [3] also prove a result that is similar to (but weaker than) Theorem 3 for the case in which the state space I is countably infinite and the parameter space A is a countably infinite compact separable metric space. Theorem 3 guarantees that the maximum likelihood estimate converges in finite time to some parameter α^*, but does not guarantee that it converges to the true parameter. Therefore, the control that is applied may not be the control that one would apply if the true parameter were known *a priori*, i.e.

$$g(x_n,\alpha^*) \neq g(x_n,\alpha_0).$$

Nevertheless, the performance of the Markov chain under the maximum likelihood estimate α_n^* is identical to the performance of the Markov chain under the true parameter α_0 using the (possibly undesirable) control $g(x_n,\alpha^*)$.

Notice that the third equation under Theorem 3 does *not* imply that the performance of the closed loop system is the same as for the case in which the true parameter α_0 is known, i.e.,

$$p(i,k;g(i,\alpha^*),\alpha^*) \neq p(i,k;g(i,\alpha_0),\alpha_0)$$

Borkar and Varaiya [2] illustrate this point by example. This review includes that example because it not only illustrates the statement above, but also provides motivation for the following sections of this review.

Example 2:

Consider the two-state controlled Markov chain with transition probabilities shown in Figure 4.

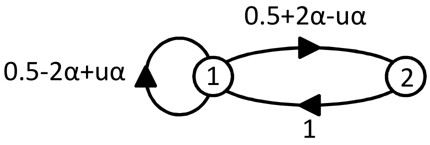

Fig. 4. The controlled Markov chain and transition probabilities

The unknown parameter α can take on one of three possible values 0.1, 0.2, and 0.3. The true parameter is $\alpha_0 = 0.2$. The initial state is $x_0 = 1$, and the stationary control policy is $u = g(i,0.1) = g(i,0.3) = 2$ and $u = g(i,0.2) = 1$ for $i = 1,2$. Notice that under these conditions, Assumption A1 is *not* satisfied. This is because for $u = 2$, the transition matrices for $\alpha = 0.1 \neq 0.3 = \beta$ are identical (since $2\alpha - u\alpha = 0$ when $u = 2$ and the transition matrix no longer depends on α). Suppose $u_0 = 1$. Then at time $n = 1$, the following are possible:

a) $x_1 = 1$, $p(1,1;u_0,0.1) = 0.4$, $p(1,1;u_0,0.2) = 0.3$, $p(1,1;u_0,0.3) = 0.2$. Thus the maximum likelihood estimate of α is 0.1.

b) $x_1 = 2$, $p(1,2;u_0,0.1) = 0.6$, $p(1,2;u_0,0.2) = 0.7$, $p(1,2;u_0,0.3) = 0.8$. Thus the maximum likelihood estimate of α is 0.3.

In either case (a) or (b), the resulting control action is $u_1 = 2$. The estimate α remains unchanged since the transition probabilities $p(i,k;2,\alpha)$ do not depend on α. Thus when $x_n = 1$, $\alpha_n^* = 0.1$, and when $x_n = 2$, $\alpha_n^* = 0.3$. The true parameter $\alpha_0 = 0.2$ is not a limiting point of the sequence $\{\alpha_n^*\}$ and the desired policy $u_n = 1$ won't be applied. Thus the performance of the Markov chain will differ from the desired performance. This illustrates the fact that Theorem 3 does *not* imply that the performance of the closed loop system is the same as for the case in which the true parameter α_0 is known. A sample simulation is provided in Fig. 5, where in this case $x_1 = 2$ and the estimator prematurely converged to $\alpha_n^* = 0.3$ and the control action is constrained to be $u_n = 2$ for $n = 0,1,2$ ...

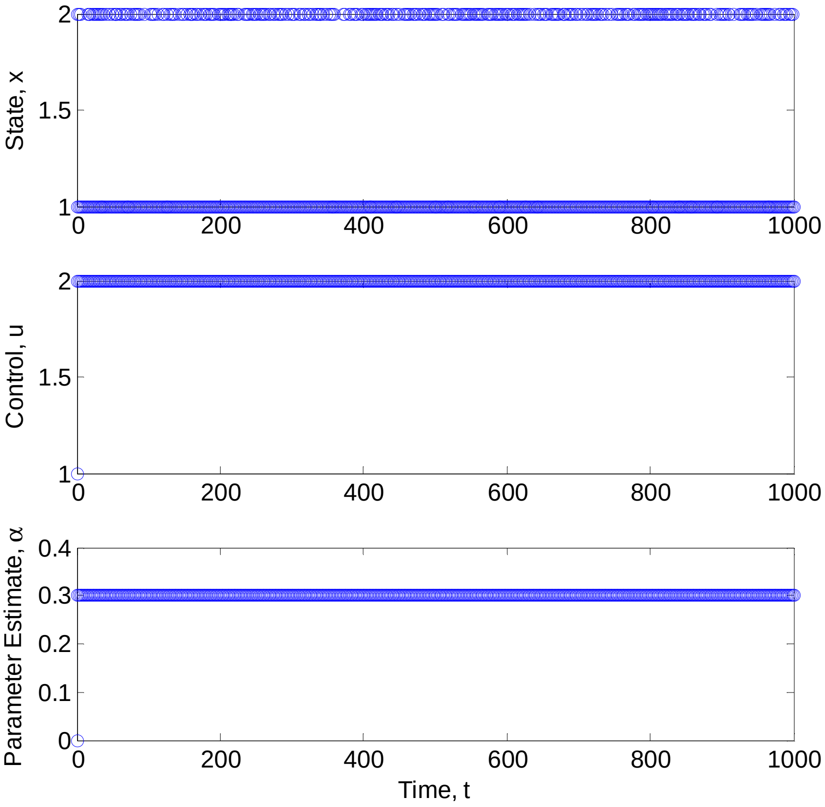

Fig. 5. Sample simulation for a Markov chain that is not identifiable.

The difficulty in Example 2 is due to the fact that, when $u = 2$, the transition matrix does not depend on α. For any other control action, the transition probability matrix depends on α and thus leads to different maximum likelihood estimates. This idea motivates the concept of randomized control which is presented in Section 4.

4. Randomized Policies for Identification and Adaptive Control

The second paper [3] by Borkar and Varaiya examines identification and adaptive control for Markov chains with countable state space I. They assumed for each i, k that $p(i, k; u, \alpha)$ is continuous in u, α.

Recall in Example 1 that when $u = 2$, the transition matrix does not depend on α, making it impossible to identify the true parameter since it cannot be a limit point of the estimation sequence. For any other control action, the transition probability matrix depends on α and thus leads to different maximum likelihood estimates. This would have increased the probability of identifying the true parameter. Based on this intuition, consider the following idea: randomly perturb the control action at each time step so that the system investigates controls different from $u = 2$. This concept allows the transition probability matrix to depend on α. Then it may be possible to accurately estimate the true parameter, and thus apply the desired control. This section discusses the development of control randomization and shows results for both identification and adaptive control. Identification is presented first.

4.1 Randomized Policies for Identification

Borkar and Varaiya described two schemes for randomizing the control law, and proved two important results. To obtain these results, they imposed assumptions that are weaker than the identifiability condition in A1.

Assumption A2:
For any $\alpha \neq \beta$ both in the parameter space A, there exists an $i \in I$ such that for every open set $O \subset U$ there exists at least one $u \in O$ for which the following is true:

$$[p(i, 1; u, \alpha), p(i, 2; u, \alpha), \cdots] \neq [p(i, 1; u, \beta), p(i, 2; u, \beta), \cdots]$$

Assumption A2 is weaker than A1 because even if A1 fails for some control action u, there may be a control action u' in an open ball centered at u such that $[p(i, 1; u, \alpha), p(i, 2; u, \alpha), \cdots] \neq [p(i, 1; u, \beta), p(i, 2; u, \beta), \cdots]$, and hence A2 holds.

Given that Assumption A2 holds, consider the following control randomization scheme:
Control Randomization Method 1 ($\{\varepsilon_i\}$ – *randomization of* u):
Specify a probability measure μ_i on U that assigns positive values to each open set. Select a small $\varepsilon_i > 0$, and for each $u \in U$, let $B(i, u)$ be the open ball of radius ε_i centered at u. Let α_n be the value of the maximum likelihood estimate at time n, and the state $x_n = i$ at time n. Then, using an independent experiment, select a control u_n from $B(i, g(i, \alpha_n))$ corresponding to the restriction of μ_i to the set $B(i, g(i, \alpha_n))$. The construction is described graphically in Fig. 6.

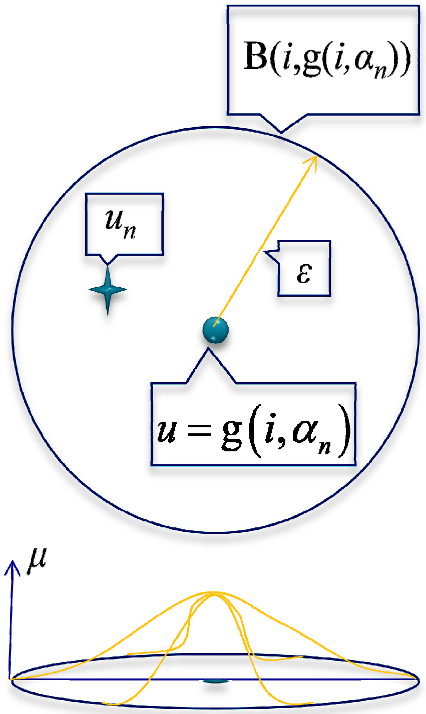

Fig. 6. Construction of $\{\varepsilon_i\}$ – randomization of u.

With this control randomization scheme, Borkar and Varaiya proved the following result.

Theorem 4:

Let the control be assigned according to the $\{\varepsilon_i\}$ – randomization of u scheme. Then the maximum likelihood estimate α_n^* converges to the true value α_0 almost surely.

In Theorem 4, under a weaker assumption than the "identifiability condition" of Assumption A1, Borkar and Varaiya demonstrate that by randomly perturbing the control action, the maximum likelihood estimate converges to the true parameter value.

The first control randomization method directly perturbs the control action by randomization. Borkar and Varaiya [3] also present the following control randomization scheme that perturbs the control indirectly by perturbing the maximum likelihood estimate. First, however, they consider the following assumption in place of Assumptions A2 and A1.

Assumption A3:

For any $\alpha \neq \beta$ both in the parameter space A, $i \in I$, and neighborhood O of A, there is an open set $\tilde{O} \subset O$ such that for every $\tilde{\alpha} \in \tilde{O}$

$$[p(i,1;g(i,\tilde{\alpha}),\alpha),p(i,2;g(i,\tilde{\alpha}),\alpha),\cdots] \neq [p(i,1;g(i,\tilde{\alpha}),\beta),p(i,2;g(i,\tilde{\alpha}),\beta),\cdots].$$

As with A2, Assumption A3 is weaker than A1 because even when A1 fails for the control action $u = g(i, \alpha)$, there may be a set $\tilde{O} \subset O$ such that for every $\tilde{\alpha} \in \tilde{O}$, $[p(i, 1; g(i, \tilde{\alpha}), \alpha), p(i, 2; g(i, \tilde{\alpha}), \alpha), \cdots] \neq [p(i, 1; g(i, \tilde{\alpha}), \beta), p(i, 2; g(i, \tilde{\alpha}), \beta), \cdots]$. Now consider the following control randomization scheme:

Control Randomization Method 2 (*$\gamma-$ randomization of α, or parameter estimate randomization*):
Specify a probability measure υ on A that assigns positive values to each open set. Select a small $\gamma > 0$ and let $B(\alpha)$ be the open ball of radius γ centered at α. Let α_n^* be the value of the maximum likelihood estimate at time n, and the state $x_n = i$ at time n. Then, using an independent experiment, select a control $u_n = g(i, \tilde{\alpha})$ where $\tilde{\alpha}$ is selected from $B(\alpha_n^*)$ corresponding to the restriction of υ to the set $B(\alpha_n^*)$.

This second control randomization method is different from the first because it randomly perturbs the parameter estimate instead of directly perturbing the control. Since the control action is a function of the parameter estimate, the control action is also perturbed due to this scheme.

Under Control Randomization Method 2, Borkar and Varaiya proved the following result which is similar to Theorem 4:
Theorem 5:
Under $\gamma-$ randomization of α, α_0 is the only frequent limit point of the sequence of maximum likelihood estimates $\{\alpha_n^*\}$ almost surely.

As with the other theorems in this review, the proofs are not included here, and the interested reader is referred to the original work [3]. The following section describes the performance of Markov chains using the randomized control schemes of this section.

4.2 Randomized Policies for Adaptive Control

The preceding section characterized the performance of the maximum likelihood estimators based on two control randomization methods and provided important results for each method. This section examines the performance of the Markov chain under these same randomized control schemes. Borkar and Varaiya [3] show the following result.

Theorem 6:
Let $g(\alpha_0)$ be the unique optimal stationary control policy under α_0. Then for any $\delta > 0$, there exists an $\varepsilon > 0$ (alternatively $\gamma > 0$) such that if $\{u_n\}$ is a ε - randomization (alternatively γ - randomization) of g, then

$$\Theta(g(\alpha_0)) \leq \lim_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} c(x_m, x_{m+1}; u_m) \leq \Theta(g(\alpha_0)) + \delta$$

Theorem 6 shows that in the limit, the performance of the Markov chain with either randomized control policy is optimal (since δ can be chosen arbitrarily small).

Example 3:
Consider the controlled Markov chain described in Example 2. In this example we apply Control Randomization Method 1: $\{\varepsilon_i\}$ – randomization of u and observe that the parameter estimate converges to the true parameter and the adaptive control law applies the corresponding control action. The approach is identical to Example 2, except that we must define a probability measure over the action space which assigns positive numbers between 0 and 1 to a ball centered around the unperturbed control action. In this case the ball is taken to be the entire action space $U = \{1,2\}$. Using this ball, note that the transition probabilities satisfy the weak form of identifiabilty defined in Assumption A2. The probabilities are defined as follows:

$$\begin{array}{ll} \mathrm{Prob}(u = 2|\alpha \in \{0.1,0.3\}) = 0.75 & \mathrm{Prob}(u = 1|\alpha \in \{0.2\}) = 0.75 \\ \mathrm{Prob}(u = 1|\alpha \in \{0.1,0.3\}) = 0.25 & \mathrm{Prob}(u = 2|\alpha \in \{0.2\}) = 0.25 \end{array}$$

Hence, the randomized control actions take on the nominal control action 75% of the time. The remaining 25% of the time, the controller applies the other possible control action for “probing” purposes. This is, of course, the essence of “dual control” in Markov chains. A sample simulation is provided in Fig. 7. Note that the control explores values different than $u = 2$. Most importantly, the parameter estimate converges to the true parameter value of $\alpha_0 = 0.2$. However, once convergence occurs, the controller still applies randomized controls which generally reduces the overall performance of the system. This limitation, along with several others, are discussed in the subsequent section.

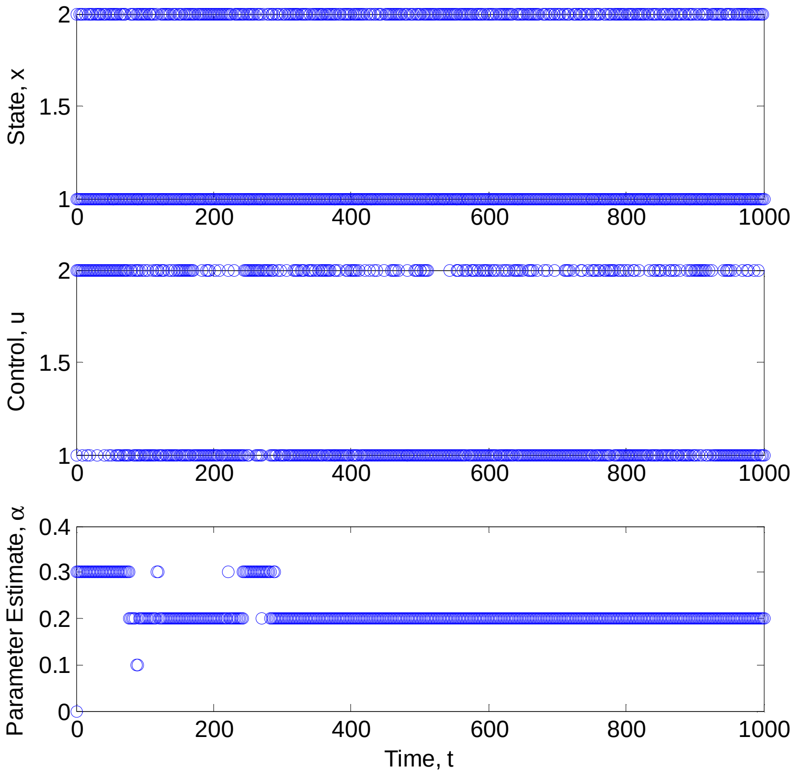

Fig. 7. Sample simulation for a Markov chain that is not identifiable, but controlled using $\{\varepsilon_i\}$-randomization of u.

5. Conclusions and Recommendations

This report reviews three papers on identification and adaptive control of Markov chains. A general framework for this problem was presented in Section 2, assuming the identifiability condition. A key theme of this report, however, focuses on the ability to relax the "identifiability condition" (Assumption A1) and still guarantee convergence to the true parameter. Section 3 relaxed this assumption completely and examined the results. Neither convergence to the true parameter nor satisfactory performance of the Markov chain is guaranteed. Indeed, Section 3 showed by example that the true unknown parameter may not even be a limit point of the estimation sequence if the identifiability condition does not hold. Thus, Section 3 concludes that the identifiability condition cannot be completely relaxed without jeopardizing estimation accuracy and/or control performance.

Section 4 imposed less restrictive forms of the "identifiability condition" and proposed randomized control policies which guarantee estimate convergence and optimal performance (if optimal control policies are applied). Thus, Section 4 demonstrates that while the identifiability condition cannot be completely dismissed, it can be made less restrictive under certain modifications to the control policies. These modifications are the essence of the dual control problem, in which control actions are chosen to tradeoff performance with learning.

Recommendations

The above comparative analysis of the three papers on identification and adaptive control of Markov chains motivates several areas for investigation, in our view:

1. *Alternative estimation methods*: A maximum likelihood estimation approach, in its most basic form, requires the identifiability condition to be satisfied to achieve convergence to the true parameter value. Hence, one might consider alternative estimation approaches which guarantee convergence (and may have other nice properties, e.g. convergence speed, intuitive separation between learning and control). Some examples previously studied in the literature are listed below:
    a. Biased maximum likelihood estimation [4]
    b. View identification and adaptive control as a multi-armed bandit problem and then apply stochastic optimization techniques [5]
    c. Bayesian estimation
2. *Algorithms for computing MLE*: Calculating the maximum likelihood estimate is generally tractable for small finite parameter, state, and control spaces. However, large finite or infinite parameter spaces require iterative approximation techniques that are generally non-trivial. A theoretical issue related to this topic is the following: If one uses iterates of likelihood estimates that are not exactly the maximizers, then does the identification algorithm still converge to the true parameter? An application based issue related to this topic is: How would one implement an iterative scheme to compute the maximum likelihood estimate (e.g. gradient-based algorithms).
3. *Control policy synthesis*: Pre-computing an optimal control policy for every $\alpha \in A$ may be intractable, particularly for large finite or infinite parameter spaces. Hence one could investigate precomputing controls for a finite subset of *A*, and then interpolating control policies for other parameter estimates. If the control policies for the finite subset are computed to be optimal in some sense, what loss of optimality is incurred for applying interpolated control laws?
4. *Modifications to control randomization*:
    a. A critical issue encountered in Example 3 is that the control is still randomized once the parameter estimated has converged to the true value. This may lead to an unnecessary loss of performance. This fact motivates a stopping criterion for random perturbations. For example, as the parameter estimate converges to the true value, the radius of the open ball ε can be constructed to simultaneously converge to 0. That is, $\varepsilon \to 0$ as $\alpha_n \to \alpha^0$.
    b. There exist clear relationships between control randomization and extremum seeking with stochastic perturbations (see paper's by Prof. Semyon Meerkov and Prof. Miroslav Krstic [6]). It would be instructive to explore these relationships and possibly share techniques between each problem area.

5. *Applications*: The theory on adaptive control of Markov chains could be readily applied to supervisory control of hybrid vehicles. Specifically, one could extend previous work on stochastic optimization of power management for PHEVs by attempting to identify the type of drive cycle the PHEV is driving on. These drive cycles are modeled as Markov chains, which can be parameterized according to the type of cycle (e.g. highway, city, suburban, etc.). Using maximum likelihood estimation, the adaptive controller can apply the optimal supervisory controller corresponding to the estimated drive cycle type, which would be computed offline.

References

[1] P. Mandl (1974). “Estimation and control in Markov chains,” *Advances in Applied Probability*, vol 6, 40-60.

[2] V. Borkar and P. Varaiya (1979), “Adaptive control of Markov chains, I: finite parameter set,” *IEEE Transactions on Automatic Control*, vol AC-24, 953-958.

[3] V. Borkar and P. Varaiya (1982), “Identification and adaptive control of Markov chains,” *SIAM Journal on Control and Optimization*, vol 20, 470-489.

[4] P.R. Kumar and A. Becker (1982), “A new family of optimal adaptive controllers for Markov chains,” *IEEE Transactions on Automatic Control*, vol AC-27, 137-146.

[5] R. Argawal, D. Teneketzis, and V. Anatharam (1989), “Asymptotically Efficient Adaptive Allocation Schemes for Controlled Markov Chains: Finite Parameter Space”, *IEEE Transactions on Automatic Control*, vol AC-34, 1249-1259.

[6] C. Manzie and M. Krstic (2009), “Extremum seeking with stochastic perturbations,” *IEEE Transactions on Automatic Control*, vol 54, p. 580-585.