EECS 558 Term Project: Adaptive Control of Markov Chains

Thomas Brady
Kevin McDonough
Presentation Outline

- Introduction

- Fundamental Work
  - Robbins (1952): Work on dual aspect of control
  - Mandl (1974): Work on system identification

- Results From Reviewed Papers
  - Borkar and Varaiya (1979): Relaxing Mandl’s identifiability condition
  - Doshi and Shreve (1980): Randomized parameter estimates
  - Kumar and Becker (1982): A family of truly optimal adaptive controllers

- Conclusions

- What we have not Reviewed
INTRODUCTION

- EECS558 has covered (in great depth) controlled Markov chains.

- What happens, though, if the transition probabilities of the system are dependent on some parameter that is unknown? This can happen, for instance, when system models are imprecise.

- This is where adaptive control of Markov Chains becomes useful.
**INTRODUCTION—MODEL**

- (M1) Finite state space, $I$: $x_t \in I, t = \{0,1,2 \ldots I\}$

- (M2) Finite action space, $U$, $u_t \in U$

- (M3) Matrix of transition probabilities depends on $u_t$ and a vector of unknown parameters, $\alpha$. $\text{Prob}\{x_{t+1} = j \mid x_t = i\} = p(i, j; u_t, \alpha), \alpha \in A, u \in U, i \in I, j \in I$ and is known.

- (M4) There exists some parameter $\alpha^0$ that corresponds to the true system model and is unknown at time $t=0$.

- (M5) At each time step, $\alpha$ is the estimate of the true parameter, $\alpha^0$. For each $\alpha \in A$, there is a predetermined $\Phi(\alpha)$ (stationary) that determines the control action, $u$. $u_t = \Phi(\alpha) \ \forall \alpha \in A$

- (M6) Estimator has perfect recall

- (M7) Instantaneous cost is known $c(x_t, x_{t+1}, u_t), x_t \in I, u_t \in U$
Herbert Robbins did work on dual action controllers

In his 1952 work on sequential experimentation, he demonstrated that a stochastic controller that takes advantage of state information performs better than one that disregards state information.

Though not surprising, an important step none the less
Mandl did work in the identification of Markov Chains

In his 1974 paper, using a contrast function and a maximum-likelihood estimate, proved that regardless of control chosen, the unknown system parameter will converge to the true parameter.
Assumptions:

(M1) There is $\epsilon > 0$ such that for every $i, j$ either 
$p(i, j; u, \alpha) > \epsilon$ for all $u, \alpha$ OR $p(i, j; u, \alpha) = 0$ for all 
$u, \alpha$.

(M2) If $\alpha, \alpha' \in A$, then there exists an $i \in I$, for which:

$[p(i, 1; u, \alpha), ..., p(i, I; u, \alpha)] \neq [p(i, 1; u, \alpha'), ..., p(i, I; u, \alpha')]$
**Fundamental Work—Mandl**

[4] *Estimation and Control in Markov Chains*

(M1): Ensures that the set of states \( j \) which are reachable from \( i \) is independent of \( u, \alpha \).

→ This is a common assumption made in work in this area.

(M2): Came to be known as “Mandl’s Identifiability Condition”. It guarantees that the transition probabilities are different for all \( \alpha \neq \alpha' \in A \). Hence, \( \alpha \) and \( \alpha' \) are *distinguishable* from one another.

→ This assumption is widely viewed as one that is too restrictive in practice.
**Fundamental Work—Mandl**

**[4] Estimation and Control in Markov Chains**

Why is the identifiability condition too restrictive?

→ Example from [1]

- Consider the familiar but simple, nonfinite state Markov chain:

\[ x_{t+1} = ax_t + bu_t + v_t, \quad t = 0, 1, 2, \ldots \]

- Here the unknown parameter is the vector \((a, b)\). Assume we have \(\alpha = (a, b)\) and \(\alpha' = (a', b')\). We also have that \(ab^{-1} = a'(b')^{-1} = g\).

- Under the control law \(u_t = -gx_t\), we have the following.

\[ p(x_t, x_{t+1}; u_t, \alpha) = p(x_t, x_{t+1}; u_t, \alpha') \]

→ The identifiability condition cannot hold.
Borkar and Varaiya

1. Adaptive Control of Markov Chains, I: Finite Parameter Set

- Showed that Mandl’s identifiability condition can be too restrictive.

- Replaced the identifiability condition with a weaker assumption and showed that in the limit, the unknown parameter may not converge to the true parameter, but the closed-loop transition probabilities would converge almost surely to something that is indistinguishable from those of the true parameter.
Assumptions:

(BV1) There is $\epsilon > 0$ such that for every $i, j$ either $p(i, j; u, \alpha) > \epsilon$ for all $u, \alpha$ OR $p(i, j; u, \alpha) = 0$ for all $u, \alpha$.

(BV2) For every $i, j$ there is a sequence of $i_1, \ldots, i_r$ such that for all $u, \alpha$

$p(i_{s-1}, i_s; u, \alpha) > 0$, $s = 1, \ldots, r+1$ where $i_0 = i, i_{r+1} = j$
(BV1): Guarantees the set of states $j$ which can be reached from $i$ is independent of $u$ and $\alpha$.

→ Equivalent to (M1)

(BV2): Guarantees that the Markov Chains generated have a single recurrent class. This condition is necessary for identification.
In mathematical form, Borkar and Varaiya’s key contribution is:

\[ p(i, j; \Phi(\alpha^*, i), \alpha^*) = p(i, j; \Phi(\alpha^*, i), \alpha^0) \]

Here \( \alpha^* \) is the most likely estimate of the true parameter and \( \alpha^0 \) is the true parameter.

Though the evolution of the Markov chain is identical, this does not mean that this evolution is optimal. That is:

\[ p(i, j; \Phi(\alpha^*, i), \alpha^*) \neq p(i, j; \Phi(\alpha^0, i), \alpha^0) \]
DOSHI AND SHREVE

[2] Strong Consistency of a Modified Maximum Likelihood Estimate for Controlled Markov Chains

- Realized the need for adaptive controllers that don’t have the potential to get “stuck” at the wrong parameter estimate.

- Made assumptions that were slightly less restrictive than those made by Mandl.

- Showed that when the estimate of the unknown parameter was chosen at random from all those parameters which nearly maximize the log likelihood function, the estimate converges in the limit to the true parameter.
Why does a parameter estimate get “stuck”?

\[ p(i, j; \Phi(\alpha^*, i), \alpha^*) = p(i, j; \Phi(\alpha^*, i), \alpha^0) \]

The likelihood function

\[ L_t(\alpha) = \prod_{s=0}^{t-1} \frac{p(x_s, x_{s+1}; u_s, \alpha)}{p(x_s, x_{s+1}; u_s, \alpha^0)} = 1 \]

Borkar and Varaiya select the parameter which maximizes the likelihood function at each time, hence choosing the wrong parameter repeatedly. \( \Rightarrow \) “stuck”.
Assumptions:

(DS1) Let $E_1, \ldots, E_\sigma$ be the set of ergodic classes of the Markov chain under the transition matrix $p(i, j; \phi(i, \alpha^0), \alpha^0)$. For $\alpha \in A$ define

$$I(\alpha) = \{i \in I \mid \text{for some } j \in I, \ p(i, j; \Phi(i, \alpha^0), \alpha) \neq p(i, j, \Phi(i, \alpha^0), \alpha^0)\}.$$ We assume that for each $\alpha$ different from $\alpha^0$, $E_k \cap I(\alpha) \neq \emptyset$.

(DS2) The true parameter $\alpha^0$ belongs to a finite set $A$.

Definition: Let $I$ be the state space of a Markov Chain. Let $i, j \in I$. A Markov chain is said to be ergodic if there exists a path of positive probability between any states $i, j \in I$.

$\Rightarrow$ Ergodicity = Irreducibility
DOSHI AND SHREVE
[2] Strong Consistency of a Modified Maximum Likelihood Estimate for Controlled Markov Chains

**Identification**

\[ L_t(\alpha) = \sum_{s=0}^{t-1} \ln p(x_s, x_{s+1}; u_s, \alpha) \]

\[ \Gamma_0(\epsilon) = A \]

\[ \Gamma_t(\epsilon) = \{ \alpha \in A | L_t(\alpha) > \max_{\alpha' \in A} L_t(\alpha') - t\epsilon \} \]

\[ t = 1, 2, ... \]
**DOSHI AND SHREVE**

[2] *Strong Consistency of a Modified Maximum Likelihood Estimate for Controlled Markov Chains*

- Proved the following:
  - For some almost surely finite random time $N(\omega)$ and $t > N(\omega)$,
    $$\Gamma_t(\epsilon) = \{\alpha_0\}$$
  - This implies that within an almost surely finite time, the possible set of parameters reduces to the true parameter.
    - The control law converges to a control law which is optimal for the true parameter
    - The average cost achieved is equivalent to the optimal long-term average cost expected if the true parameter were known *a priori.*
KUMAR AND BECKER


- Presented a family of adaptive controllers that exhibited the following characteristics:
  - Convergence, in a Cesaro sense, of the estimates of the unknown parameter to the true parameter
  - Convergence, in a Cesaro sense, to a control law to one that is optimal for the true parameter
  - Long-term cost equivalent to the optimal long-term cost achievable if the true parameter were known

- Used the well-known MLE slight biased towards those parameter estimates that yielded lower long-term expected cost
Kumar and Becker emphasize that in their results, Cesaro sense convergence can *not* be replaced by convergence. We shall discuss this later.
**Assumptions**

(KB1) $X,U,A$ are finite sets

(KB2) For every $(i,j) \in X \times X$ either $p(i,j;u,\alpha) > 0$ $\forall (u,\alpha) \in U \times A$ or $p(i,j;u,\alpha) = 0$ $\forall (u,\alpha) \in U \times A$.

(KB3) For every $(i,j) \in X \times X$, there exists an integer $m_{ij}$ and a sequence of states $i = k_0, k_1, \ldots, k_{m_{ij}} = j$ such that $p(k_s,k_{s+1};u,\alpha) > 0$ $\forall (u,\alpha) \in U \times A$ and $S = 0,1,\ldots, m_{ij} - 1$. 
The “Maximum Likelihood Estimate”

- Let \( f(J^*(\alpha)) \) be a strictly monotone increasing function such that \( f(J^*(\alpha)) > 0 \ \forall \alpha \in A \)

- Let \( o \) be any positive valued function such that \( \lim_{t \to \infty} o(t) = \infty \) while \( \lim_{t \to \infty} o(t)/t = 0 \).

\[
\overline{D}_t(\alpha) := [f(J^*(\alpha))]^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}; u_s, \alpha)
\]

- The term \([f(J^*(\alpha))]^{-o(t)}\) gives the MLE a slight bias towards those parameter estimates which yield a lower long-term expected cost.
**Results** (Theorem 10 [5])

The sequence of parameter estimates \( \{\alpha_t(\omega)\} \) converges almost surely, in a Cesaro sense to a random variable \( \alpha^* \). Under the optimal feedback control law for \( \alpha^* \), the closed-loop transition probabilities under \( \alpha^* \) and \( \alpha^0 \) are equivalent. That is:

\[
p(i,j; \Phi(i, \alpha^*), \alpha^*) = p(i,j; \Phi(i, \alpha^*), \alpha^0)
\]

The adaptive control law converges almost surely in a Cesaro sense to \( \Phi(\cdot, \alpha^*) \) which is an optimal feedback control law for \( \alpha^0 \).

\[
J(\Phi(\cdot, \alpha^*), \alpha^0) = J^*(\alpha^0)
\]

The adaptive controller attains optimal performance almost surely.

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} c(x_s, x_{s+1}, \Phi(x_s, \alpha_s)) = J^*(\alpha^0)
\]
CONCLUSIONS

- The ability to distinguish between the parameters $\alpha \in A$ is a vital aspect in the adaptive control of Markov chains.

- The problem of identifying an unknown parameter while controlling a Markov chain is another manifestation of the dual control problem.

- Slight changes to the assumptions in these types of problems can lead to drastically different results!
CONCLUSIONS

When $p(i, j; \Phi(\alpha^*, i), \alpha^*) = p(i, j; \Phi(\alpha^*, i), \alpha^0)$, while the evolution of the Markov chain under the estimate of the true parameter and the true parameter is identical, this does not mean that the evolution is optimal! This also does not mean that the performance of the Markov chain under the estimated parameter will be optimal.
SITUATIONS WE HAVE NOT COVERED

- Partially observed stochastic systems
- Nonfinite state or action space models
- Situations where the unknown parameter $\alpha$ does not belong to a known set
Questions?
REFERENCES


