Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Estimation of Non-stationary Markov Chain Transition Models

L. F. Bertuccelli and J. P. How

Aerospace Controls Laboratory
Massachusetts Institute of Technology
{lucab, jhow} @mit.edu

Abstract— **Many decision systems rely on a precisely known Markov Chain model to guarantee optimal performance, and this paper considers the online estimation of unknown, non-stationary Markov Chain transition models with perfect state observation. In using a prior Dirichlet distribution on the uncertain rows, we derive a mean-variance equivalent of the Maximum A Posteriori (MAP) estimator. This recursive mean-variance estimator extends previous methods that recompute the moments at each time step using observed transition counts. It is shown that this mean-variance estimator responds slowly to changes in transition models (especially switching models) and a modification that uses ideas of pseudonoise addition from classical filtering is used to speed up the response of the estimator. This new, discounted mean-variance estimator has the intuitive interpretation of fading previous observations and provides a link to fading techniques used in Hidden Markov Model estimation. Our new estimation techniques is both faster and has reduced error than alternative estimation techniques, such as finite memory estimators.**

I. Introduction

Many decision processes, such as Markov Decision Processes (MDPs) and Jump Markov Linear systems, are modeled as a probabilistic process driven by a Markov Chain. The true parameters of the Markov Chain are frequently unavailable to the modeler, and many researchers have recently been addressing the issue of robust performance in these decision systems [4], [6], [13], [16]. However, a large body of research has also been devoted to the identification of the Markov Chain using available observations. With few exceptions (such as the signal processing community [11], [17]), most of this research has addressed the case of a unique, stationary model.

When the transition matrix Π of a Markov Chain is stationary, classical maximum likelihood (ML) schemes [9], [17] can be used to recursively obtain the best estimate $\hat{\Pi}$ of the transition matrix. Typical Bayesian methods assume a prior Dirichlet distribution on each row of the transition matrix, and exploit the conjugacy property of the Dirichlet distribution with the multinomial distribution to recursively compute $\hat{\Pi}$. This technique amounts to evaluating the empirical frequency of the transitions to obtain a ML or Maximum A Posteriori (MAP) estimate of the transition matrix. In the limit of an infinite observation sequence, this method converges to the true transition matrix, Π. Jilkov and Li [9] discuss the identification of the transition matrices in the context of Markov Jump systems, providing multiple algorithms that can identify Π using noisy measurements that are indirect observations of the transitions. In one of these approaches, a renormalization is used to ensure that the probability estimates sum to unity. Jaulmes et al. [7], [8] study this problem in an active estimation context using Partially Observable Markov Decision Processes (POMDPs). Marbach [14] considers this problem, when the transition probabilities depend on a parameter vector. Borkar and Varaiya [5] treat the adaptation problem in terms of a single parameter as well; namely, the true transition probability model is assumed to be a function a single parameter a belonging to a finite set $\mathcal{A}$. Konda and Tsitsiklis [10] consider the problem of slowly-varying Markov Chains in the context of reinforcement learning. Sato [18] considers this problem and shows asymptotic convergence of the probability estimates also in the context of dual control. Kumar [12] also considered the adaptation problem.

If the Markov Chain, Π_t, is changing over time, classical ML or MAP estimators will generally fail to respond quickly to changes in the model. The intuition behind this is that since these estimators keeps track of all the transitions that have occurred, a large number of new transitions will be required for the change detection, and convergence to the new model. Hence, new estimators are required to compensate for the inherent delay that will occur in classical techniques. Note that if the dynamics of the transition matrix were available to the estimator designer, they could be embedded directly in the estimator. For example, if the transition matrix were known to switch between two systems according to a probabilistic switching schedule, or if the switching time were a random variable with known statistics, these pieces of information could enhance the performance of any estimator. However, in a more general setting, it is unlikely that this information would be available to the estimator designer.

This paper proposes a new technique to speed up the estimator response that does not require information on the dynamics of the uncertain transition model. First, recursions for the mean and variance of the Dirichlet distribution are derived; these are equivalent to a mean-variance interpretation of classical ML or MAP estimation techniques. Importantly, however, we use the similarity of these recursions to Kalman filter-based parameter estimation techniques to notice that the mean-variance estimator does not incorporate any knowledge of the parameter (or transition matrix) dynamics, and therefore results in stationary prediction step. To compensate for this, the responsiveness of the estimator can be improved by adding an artificial pseudonoise to the variance which is implemented by scaling the variance [15]. Scaling the

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

variance leads to a very natural interpretation for updating the Dirichlet parameters, which amounts to progressively fading the impact of older transitions. This result provides an intuition for measurement fading applied to Hidden Markov Models [11]. This insight, and the resulting benefits of faster estimation when applied to decision systems, are the core results of this paper.

II. Markov Chain and the Dirichlet Distribution

A transition matrix Π of an N-state Markov Chain is defined as $\Pi \in \mathcal{R}^{N\times N}$ given by

$$\Pi = \begin{bmatrix} \pi(1,1) & \pi(1,2) & \dots & \pi(1,N) \\ \pi(2,1) & \pi(2,2) & \dots & \pi(2,N) \\ \dots & & & \\ \pi(N,1) & \pi(N,2) & \dots & \pi(N,N) \end{bmatrix}$$

where $\pi(i,j)$ entry is the probability that the a transition to state j at time $k+1$, given the state was i at the previous time step

$$\pi(i,j) = \Pr[x_{k+1} = j \mid x_k = i] \tag{1}$$

Note that $\sum_j \pi(i,j) = 1$. When the transition matrix Π is uncertain, we can take a fairly common Bayesian viewpoint [7] and assume a prior Dirichlet distribution on each row of the transition matrix, and recursively update this distribution with observations.[1]

The Dirichlet distribution f_D at time k for a row of the N-dimensional transition model is given by $\mathbf{p}_k = [p_1, p_2, \dots, p_N]^T$ and hyperparameters (with $\alpha_i > 1$) $\alpha(k) = [\alpha_1, \alpha_2, \dots, \alpha_N]^T$, is defined as

$$\begin{aligned} f_D(\mathbf{p}_k|\alpha(\mathbf{k})) &= K \prod_{i=1}^{N} p_i^{\alpha_i - 1}, \quad \sum_i p_i = 1 \qquad (2) \\ &= K\ p_1^{\alpha_1-1} p_2^{\alpha_2-1} \dots (1 - \sum_{i=1}^{N-1} p_i)^{\alpha_N - 1} \end{aligned}$$

where K is a normalizing factor that ensures the probability distribution integrates to unity. Each p_i is the i^{th} column of the m^{th} row, that is: $p_i = \pi(m,i)$ and and $0 \le p_i \le 1$ and $\sum_i p_i = 1$. The primary reasons for using the Dirichlet distribution is that the mean $\bar{p}_i$ satisfies the requirements of a probability mass function $0 \le \bar{p}_i \le 1$ and $\sum_i \bar{p}_i = 1$ by construction. In fact, by sampling the Dirichlet distribution, each sample p_i^s will satisfy $\sum_i p_i^s = 1, \quad \forall s$ and $0 \le p_i^s \le 1, \quad \forall s$. Furthermore, the hyperparameters α_i that can be interpreted as "counts", or times that a particular state transition was observed, thus easily updating the distribution based on new observations.

The mean and the variance of the Dirichlet distribution can then be calculated directly as

$$\bar{p}_i = \alpha_i/\alpha_0 \tag{3}$$

$$\Sigma_{ii} = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{4}$$

[1]Since each row of the transition matrix satisfies the properties of a probability mass function, the following description of the Dirichlet distribution is interpreted to apply to *each row* of the transition matrix.

These are the mean and the variance of each column of the transition model, and need to be evaluated for all rows (recalling $p_i = \pi(m,i)$).

III. Derivation of Mean-Variance Estimator

It is well known that the Dirichlet distribution is conjugate to the multinomial distribution; therefore, performing a Bayesian measurement update step on the Dirichlet amounts to a simple addition of currently observed transitions to the previously observed counts $\alpha(k)$. The posterior distribution $f_D(\mathbf{p}_{k+1}|\alpha(k+1))$ is given in terms of the prior $f_D(\mathbf{p}_k|\alpha(k))$ as

$$\begin{aligned} f_D(\mathbf{p}_{k+1}|\alpha(k+1)) &\propto f_D(\mathbf{p}_k|\alpha(k)) f_M(\beta(k)(|\mathbf{p}_k) \\ &= \prod_{i=1}^{N} p_i^{\alpha_i - 1} p_i^{\beta_i} = \prod_{i=1}^{N} p_i^{\alpha_i + \beta_i - 1} \end{aligned}$$

where $f_M(\beta(k)|\mathbf{p}_k)$ is a multinomial distribution with hyperparameters $\beta(k) = [\beta_1, \dots, \beta_N]$. Each β_i is the total number of transitions observed from state i to a new state i': mathematically $\beta_{i'} = \sum_i \delta_{i,i'}$ and

$$\delta_{i,i'} = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{else} \end{cases} \tag{5}$$

indicates how many times transitions were observed from state i to state i'. For the next derivations, we assume that only a single transition can occur per time step, $\beta_i = \delta_{i,i'}$.

Upon receipt of the observations $\beta(k)$, the parameters $\alpha(k)$ are thus updated in the following manner

$$\alpha_i(k+1) = \begin{cases} \alpha_i(k) + \delta_{i,i'} & \text{Transition } i \text{ to } i' \\ \alpha_i(k) & \text{Else} \end{cases}$$

The mean and the variance can then be calculated by using Eqs. 3 and 4.

Instead of calculating the mean and variance from the transitions at each time step, we can directly find recursions for the mean $\bar{p}_i(k)$ and variance $\Sigma_{ii}(k)$ of the Dirichlet distribution by deriving the Mean-Variance Estimator with the following proposition [3].

Proposition 1: The posterior mean $\bar{p}_i(k+1)$ and variance $\Sigma_{ii}(k+1)$ of the Dirichlet distribution can be found in terms of the prior mean $\bar{p}_i(k)$ and variance $\Sigma_{ii}(k)$ by using the following recursion for the Mean-Variance Estimator:

$$\begin{aligned} \bar{p}_i(k+1) &= \bar{p}_i(k) + \Sigma_{ii}(k) \tfrac{\delta_{i,i'} - \bar{p}_i(k)}{\bar{p}_i(k)(1-\bar{p}_i(k))} \\ \Sigma_{ii}^{-1}(k+1) &= \gamma_{k+1}\ \Sigma_{ii}^{-1}(k) + \tfrac{1}{\bar{p}_i(k+1)(1-\bar{p}_i(k+1))} \end{aligned}$$

where $\gamma_{k+1} = \frac{\bar{p}_i(k)(1-\bar{p}_i(k))}{\bar{p}_i(k+1)(1-\bar{p}_i(k+1))}$.

Remark 1: The recursion for the mean is actually the maximum a posteriori (MAP) estimator of the mean of the Dirichlet distribution, expressed in terms of prior mean and variance. If the updated counts are $\alpha'(k+1)$, then the posterior distribution is given by

$$f_D(\mathbf{p}_{k+1}|\alpha'(\mathbf{k+1})) = K \prod_{i=1}^{N} p_i^{\alpha_i'}, \quad \sum_i p_i = 1$$

TABLE I

MEAN VARIANCE RECURSION SHOWN IN PREDICTION AND UPDATE STEP

	Mean-variance
Prediction	$\bar{p}_i(k+1\|k) = \bar{p}_i(k\|k)$ $\Sigma_{ii}(k+1\|k) = \Sigma_{ii}(k\|k)$
Measurement update	$\bar{p}_i(k+1\|k+1) = \bar{p}_i(k+1\|k) + \Sigma_{ii}(k+1\|k)\frac{\delta_{i,i'} - \bar{p}_i(k+1\|k)}{\bar{p}_i(k+1\|k)(1-\bar{p}_i(k+1\|k))}$ $\Sigma_{ii}^{-1}(k+1\|k+1) = \gamma_{k+1}\Sigma_{ii}^{-1}(k+1\|k) + \frac{1}{\bar{p}_i(k+1\|k)(1-\bar{p}_i(k+1\|k))}$

and the MAP estimate $\hat{p}_i$ is $\hat{p}_i = \arg\max f_D(\mathbf{p}|\alpha^{'}(\mathbf{k}+\mathbf{1}))$.

Remark 2: This mean-variance estimator explicitly guarantees that the estimates sum to unity, $\sum_i \bar{p}_i(k|k) = 1, \quad \forall k$, since they are calculated directly from the MAP estimate. Other mean-variance approaches [9] only enforce the unit sum constraint at the end of each estimator cycle, through some form of normalization. However, in the mean-variance form for the Dirichlet, no approximations are needed to ensure that the estimates remain within the unit simplex.

Remark 3: The convergence of the mean-variance estimator is guaranteed since the MAP estimator is guaranteed to converge [7]. After a large number of observations, the MAP estimate of the probability $\bar{p}_i = \alpha_i/\alpha_0$ will be equal to the true probability p_i, and the variance asymptotically approaches 0.

This is immediately clear from the mean-variance formulation as well. From proposition 1, the estimate $\bar{p}_i(k)$ will converge if $\lim_{k\to\infty} \bar{p}_i(k+1) - \bar{p}_i(k) = 0$, which implies that for any arbitrary measurement $\delta_{i,i'}$, that this will be true if the variance asymptotically approaches 0, $\lim_{k\to\infty} \Sigma_{ii}(k) = 0$.

The steady-state covariance can be found explicitly in the mean-variance estimator by rearranging the expression in Proposition 1, and taking the limit.

$$\lim_{k\to\infty} \Sigma_{ii} = \lim_{k\to\infty} (1-\gamma_{k+1})\ \bar{p}_i(k+1)(1-\bar{p}_i(k+1)) = 0$$

Note that we have used the fact that, since the estimate converges, then by definition of γ_k, $\lim_{k\to\infty} \gamma_{k+1} = 1$.

Remark 4: The mean-variance estimator can also be expressed more explicitly in a prediction step and a measurement update step, much like in conventional filtering. The prior distribution is given by $f_D(\mathbf{p}_{k|k}|\alpha(k))$ where the prior estimate is now written as $\bar{p}_i(k|k)$. The propagated distribution is $f_D(\mathbf{p}_{k+1|k}|\alpha(k))$ and the propagated estimate is denoted as $\bar{p}_i(k+1|k)$. The posterior distribution is $f_D(\mathbf{p}_{k+1|k+1}|\alpha(k+1))$, where $\alpha(k+1)$ are the updated counts, and the updated estimate is written as $\bar{p}_i(k+1|k+1)$. These steps are shown in Table I. In the (trivial) prediction step, the mean and the variance do not change, while the measurement update step is the proposition we just derived.

IV. DERIVATION OF THE DISCOUNTED MEAN VARIANCE ESTIMATOR

The general limitation of applying this estimation technique to a non-stationary problem is that the variance of the estimator decreases to 0 rapidly after $N_m \ll \infty$ measurements, which in turn implies that new observations $\delta_{i,i'}$ will not be heavily weighted in the measurement update. This can be seen in the measurement update step of Table I: as the variance Σ_{ii} approaches zero, then new measurements have very little weighting.

This covariance can be thought of as the measurement *gain* of classical Kalman filtering recursions. A way to modify this gain is by embedding transition matrix dynamics. If transition matrix dynamics were available, these could be embedded in the estimator by using the Chapman-Kolmogorov equation $\int P(\pi_{k+1}|\pi_k)P(\pi_k|\alpha(k)d\pi_k$ in the prediction step. However, in general, the dynamics of the parameter may not be well known or easily modeled.

In parameter estimation, well known techniques are used to modify this prediction step for a time-varying unknown parameters, such as through the addition of artificial pseudonoise [19], or scaling the variance by a (possibly time-varying) factor greater than unity [15]. Both pseudonoise addition or covariance scaling rely on the fundamental idea of increasing the covariance of the estimate in the prediction step.

In Ref. [15], Miller artificially scales the predicted covariance matrix $\Sigma_{k+1|k}$ by a time-varying scale factor ω_k ($\omega_k > 1$) and shows that the Kalman filter recursions remain virtually unchanged, except that that predicted variance $\Sigma_{k+1|k}$ is modified to $\Sigma^{'}_{k+1|k} = \omega_k\Sigma_{k+1|k}$. Since $\omega_k > 1$, this has the effect of increasing the covariance, thereby reducing the estimator's confidence and changing the Kalman gain to be more responsive to new measurements.

We thus use this similar intuition to our mean-variance estimator for the case of the Dirichlet distribution; define $\lambda_k = 1/\omega_k$ (where now $\lambda_k < 1$), and modify the prediction steps in a similar way to Miller, and obtain the direct analog for our mean-variance estimator. Our new update step for the variance is given by

$$\Sigma_{ii}^{-1}(k+1|k) = \lambda_k\Sigma_{ii}^{-1}(k|k) \tag{6}$$

The variance is now scaled by a factor $1/\lambda_k > 1$ at each iteration. The complete recursion for the Discounted Mean-Variance Estimator is as follows (the prediction and measurement update step have been combined)

$$\bar{p}_i(k+1|k+1) = \bar{p}_i(k|k) + 1/\lambda_k\Sigma_{ii}(k|k)\frac{\delta_{i,i'} - \bar{p}_i(k|k)}{\bar{p}_i(k|k)(1-\bar{p}_i(k|k))}$$
$$\Sigma_{ii}^{-1}(k+1|k+1) = \lambda_k\gamma_{k+1}\Sigma_{ii}^{-1}(k|k) + \frac{1}{\bar{p}_i(k|k)(1-\bar{p}_i(k|k))}$$

Note that since the posterior mean $\bar{p}_i(k+1|k+1)$ is directly dependent on $\Sigma_{ii}(k|k)$, scaling the variance by $1/\lambda_k$ will result in faster changes in the mean than if no scaling were

TABLE II
DISCOUNTED MEAN VARIANCE RECURSION

	Mean-variance
Prediction	$\bar{p}_i(k+1\|k) = \bar{p}_i(k\|k)$ $\Sigma_{ii}^{-1}(k+1\|k) = \lambda_k\ \Sigma_{ii}^{-1}(k\|k)$
Combined updates	$\bar{p}_i(k+1\|k+1) = \bar{p}_i(k\|k) + 1/\lambda_k \Sigma_{ii}(k\|k) \frac{\delta_{i,i'} - \bar{p}_i(k\|k)}{\bar{p}_i(k\|k)(1-\bar{p}_i(k\|k))}$ $\Sigma_{ii}^{-1}(k+1\|k+1) = \lambda_k \gamma_{k+1} \Sigma_{ii}^{-1}(k\|k) + \frac{1}{\bar{p}_i(k\|k)(1-\bar{p}_i(k\|k))}$

applied. Table II shows this estimator also in terms of the individual prediction and measurement update steps.

A. Intuition on the Dirichlet model

There is a fairly natural counts-based interpretation of covariance scaling for the Dirichlet distribution. Note that the variance of the Dirichlet implies that the following holds,

$$\begin{aligned} 1/\lambda_k\ \Sigma_{ii}(k+1|k) &= 1/\lambda_k\ \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)} \\ &= \frac{\lambda_k\alpha_i(\lambda_k\alpha_0-\lambda_k\alpha_i)}{\lambda_k^3\alpha_0^2(\alpha_0+1)} \end{aligned} \quad (7)$$

When $\alpha_0 \gg 1$ (this holds true very early in the estimation process), the above expression is approximately equal to

$$\frac{\lambda_k\alpha_i(\lambda_k\alpha_0-\lambda_k\alpha_i)}{\lambda_k^3\alpha_0^2(\alpha_0+1)} \approx \frac{\alpha_i(\alpha_0-\lambda_k\alpha_i)}{\alpha_0^2(\lambda_k\alpha_0+1)} \quad (8)$$

But this is nothing more than the variance of a Dirichlet distribution where the parameters are chosen in the form $\alpha^{'}(k) = \lambda_k\ \alpha(k)$ instead of $\alpha(k)$. In fact, if the distribution is given by $f_D(\mathbf{p}|\alpha^{'}(\mathbf{k})) = K\ \prod_{i=1}^{N} p_i^{\lambda_k\alpha_i}$, the first two moments are given by

$$\begin{aligned} \bar{p}_i &= \lambda_k\alpha_i/\lambda_k\alpha_0 = \alpha_i/\alpha_0 \\ \Sigma_{ii} &= \frac{\lambda_k^2\alpha_i(\alpha_0-\alpha_i)}{\lambda_k^2\alpha_0^2(\lambda_k\alpha_0+1)} = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\lambda_k\alpha_0+1)} \end{aligned} \quad (9)$$

Hence, the discounted mean variance formulation can be interpreted as updating the counts in the following manner

$$\alpha_i(k+1) = \lambda_k\alpha_i(k) + \delta_{i,i'} \quad (10)$$

rather than $\alpha_i(k+1) = \alpha_i(k) + \delta_{i,i'}$ in the undiscounted version.

B. Switching Models

Now, consider a specialized case of a time-varying transition matrix: the case when the matrix switches at distinct (but unknown) set of times T_{sw}. In this case, it can be shown that the Mean-Variance estimator will eventually converge to the true model.

The discounted mean-variance estimator does not exhibit the same convergence properties as the undiscounted estimator for arbitrary $\lambda_k < 1$; this includes the case of constant λ_k, where $\lambda_k = \lambda < 1$. This is because the estimator has been modified to always maintains some level of *uncertainty* by rescaling the uncertainty. In particular, the estimator will constantly be responding to the most recent observations, and will only converge if the following proposition holds.

Proposition 2: The discounted mean-variance estimator will converge if $\lim_{k\to\infty} \lambda_k = 1$.

Proof: The asymptotic variance, $\Sigma_{ii}(\infty) = \lim_{k\to\infty} \Sigma_{ii}(k)$ is given by

$$\Sigma_{ii}(\infty) = \lim_{k\to\infty} \frac{(1-\lambda_k\ \gamma_{k+1})}{2-\lambda_k} \bar{p}_i(k+1)(1-\bar{p}_i(k+1))$$

and will asymptotically reach zero if both $\lim_{k\to\infty} \gamma_{k+1} = 1$ and $\lim_{k\to\infty} \lambda_k = 1$. If $\lambda_k = \lambda < 1$, the variance will not converge to 0; however, if $\lim_{k\to\infty} \lambda_k = 1$, the discounted mean estimator will converge to the undiscounted form, and hence the estimator will converge to the true parameter. ■

It is shown in the next simulations that using a constant λ_k still provides good estimates of the true parameter, but we caution that to achieve convergence, λ_k should be chosen such that $\lim_{k\to\infty} \lambda_k = 1$. Such a choice could for example be $\lambda_k = 1 - \lambda^k$, where $\lambda < 1$.

V. NUMERICAL SIMULATIONS

This section presents some numerical simulations showing the responsiveness of the discounted mean-variance estimator. In the first set of examples, we show a set of runs showing the identification of an underlying (non-stationary transition matrix) that switches from A_1^- to A_1^+ at some unknown time T_{sw} and show that the discounted mean-variance estimator responds quicker to the change than other estimators, such as the undiscounted version or a finite memory estimator. In the second set of examples, we show an implementation of the discounted mean-variance formulation in an infinite horizon Markov Decision Process, where at each time that the transition matrix is identified, a new control policy is calculated. The optimal objective of each policy converges quicker when the discounted mean-variance approach is used to identify the transition matrix.

A. Transition Matrix Identification

This first example has an underlying transition matrix that switches at some unknown time T_{sw}. First, we show the benefit of using the discounted version of the estimator over the undiscounted estimator. This is shown in Figure 1 where the discounted estimator (blue) responds to the change in transition matrix almost instantly at $t = 50$ seconds, and after 20 seconds from the switch, has a 50% error ($\hat{p} = 0.7$) from the true parameter $p = 0.8$. The undiscounted estimator (red) has a 50% error after 50 seconds, and is much slower.

Next, compare the identification of this model with a finite memory estimator which calculated by storing all observed

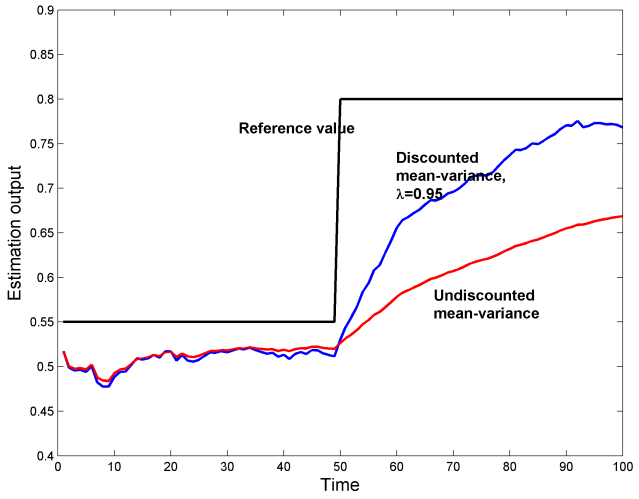

Fig. 1. Discounted estimator (blue) has a faster response at the switch time than undiscounted estimator (red).

transitions in previous M time steps,

$$\hat{\alpha}_i(k) = \sum_{j=k-M+1}^{k} \delta^j_{i,i'} \tag{11}$$

where $\delta^j_{i,i'}$ is unity if a transition occurred from state i to state i' at time j. The mean and variance are calculated using

$$\begin{aligned} \bar{p}_i(k) &= \frac{\hat{\alpha}_i}{\hat{\alpha}_0} \\ \Sigma_{ii}(k|k) &= \frac{\hat{\alpha}_i(\hat{\alpha}_0 - \hat{\alpha}_i)}{\hat{\alpha}_0^2(\hat{\alpha}_0 + 1)} \end{aligned}$$

where $\hat{\alpha}_0 = \sum_i \hat{\alpha}_i(k)$. Note that the finite memory estimator does not include information that is older than M time steps. The three estimators compared in the next simulations are

- Estimator #1: Undiscounted estimator
- Estimator #2: Discounted estimator (varying λ_k)
- Estimator #3: Finite memory estimator (varying M)

Table III presents the summary statistics of these simulations in terms of mean absolute error, standard deviation of absolute error, and min/max of the absolute error. A two-sided T-test showed that the difference between the discounted estimator and the finite memory estimator up to $\lambda = 0.925$ and $M = 20$ was statistically significant at $p < 0.01$. Also note that the use of a finite memory estimator generally requires that all the observed transitions in the previous M steps be stored. For large M and a large system, this may in fact be impractical; this memory storage is not required in the discounted mean-variance formulation, where only storing the $\alpha_i(k)$ is required.

TABLE III
MEAN / STANDARD DEVIATION OF ABSOLUTE ERROR

λ	Mean	Variance	Min	Max
0.85	0.215	0.099	0.018	0.632
0.875	0.196	0.096	0.011	0.601
0.90	0.178	0.094	0.005	0.577
0.925	0.163	0.094	0.013	0.563
0.95	0.156	0.096	0.011	0.587
M	Mean	Variance	Min	Max
10	0.255	0.119	0.014	0.659
15	0.236	0.108	0.017	0.777
20	0.204	0.102	0.004	0.586
25	0.144	0.084	0.009	0.463
30	0.144	0.084	0.009	0.463

B. Online MDP Replanning

This section considers a machine repair/replacement problem [2] driven by a time-varying transition matrix, posed as a Markov Decision Process (MDP). Similar to the previous example, the transition model is assumed to switch from model A_1^- to model A_1^+ at an unknown time T_{sw}. The estimate of the transition matrix is updated at each time step with the most recent observations, and the optimal policy for the DP is re-calculated using the current estimate.[2]

1) Problem Statement: A machine can take on one of two states x_k at time k: i) the machine is either *running* ($x_k = 1$), or ii) broken (not running, $x_k = 0$). If the machine is running, a profit of \$100 is made. The control options available to the user are the following: if the machine is running, a user can choose to either i) perform maintenance (abbreviated as $u_k = m$) on the machine (thereby decreasing the likelihood the machine failing in the future), or ii) leave the maching running without maintenance ($u_k = n$). The choice of maintenance has cost, C_{maint}, e.g., the cost of a technician to maintain the machine.

If the machine is broken, two choices are available to the user: i) repair the machine ($u_k = r$), or ii) completely replace the machine ($u_k = p$). Both of these two options come at a price, however; machine repair has a cost C_{repair}, while machine replacement is $C_{replace}$, where for any sensible problem specification, the price of replacement is greater than the repair cost $C_{replace} > C_{repair}$. If the machine is replaced, it is *guaranteed* to work for at least the next stage.

For the case of the machine running at the current time step, the state transitions are governed by the following model

$$\begin{aligned} \Pr\,(x_{k+1} = \text{fails} \mid x_k\text{=running}, u_k\text{=m}) &= \gamma_1 \\ \Pr\,(x_{k+1} = \text{fails} \mid x_k\text{=running}, u_k\text{=n}) &= \gamma_2 \end{aligned}$$

For the case of the machine not running at the current time step, the state transition are governed by the following model

$$\begin{aligned} \Pr\,(x_{k+1} = \text{fails} \mid x_k\text{=fails}, u_k\text{=r}) &= \gamma_3 \\ \Pr\,(x_{k+1} = \text{fails} \mid x_k\text{=fails}, u_k\text{=p}) &= 0 \end{aligned}$$

Note that, consistent with our earlier statement that machine replacement guarantees machine function at the next time step, the transition matrix for the replacement is deterministic. From these two models, we can completely describe the transition matrix if the machine is running or not running at

[2]This problem is sufficiently small that the policy can be quickly recalculated. For larger problems, one might have to resort to Real-Time Dynamic Programming (RTDP) techniques [1].

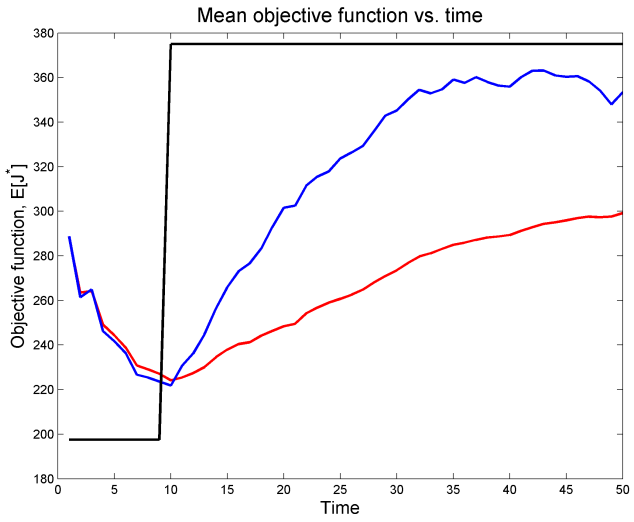

Fig. 2. At $t = 10$, the transition matrix changes from A_1^- to A_1^+, and the MDP solution (after replanning at each observation) using the discounted estimator (blue) converges in the neighborhood of the optimal objective J^* quicker than with using the undiscounted estimator (red)

the current time step:

$$\text{Machine Running } (x_k = 1), A_1 : \quad \begin{bmatrix} 1-\gamma_1 & \gamma_1 \\ 1-\gamma_2 & \gamma_2 \end{bmatrix}$$

$$\text{Machine Not Running } (x_k = 0), A_2 : \quad \begin{bmatrix} 1-\gamma_3 & \gamma_3 \\ 1 & 0 \end{bmatrix}$$

The objective is to find an optimal control policy such that $u_k(x_k = 0) \in \{$ r, p $\}$ if the machine is not running, and $u_k(x_k = 1) \in \{$ m, n $\}$ if the machine is running, for each time step. The state of the machine is assumed to be perfectly observable, and this can be solved via dynamic programming.

2) Results: The transition matrix for time $t < T_{sw}$ was

$$A_1^- = \begin{bmatrix} 0.05 & 0.95 \\ 0.3 & 0.7 \end{bmatrix},$$

while for $t \geq T_{sw}$, the transition matrix was

$$A_1^+ = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}.$$

The response speeds of the two types of estimators can be calculated by evaluating the difference in the mean objective function and. The optimal policy $u^*(k,s)$ and optimal cost $J^*(k,s)$ are calculated at each time step k and simulation s using $i)$ the discounted estimator $(u_d^*(k,s), J_d^*(k,s))$ and the undiscounted estimator $(u_u^*(k,s), J_u^*(k,s))$. The mean of the objective function is calculated as follows

$$\bar{J}_u(k) \quad = \quad \frac{1}{N_s} \sum_{s=1}^{N_s} J_u^*(k,s),$$

The mean of the objective function for $\lambda_k = 0.90$ is shown in Figure 2. The discounted estimator response (blue) is shown to be much faster than the undiscounted response (red) at the switch time of T_{sw} 10 seconds.

VI. Conclusions

This paper has presented a formulation for the identification on non-stationary Markov Chains that uses filtering insight to speed up the response of classical ML-based estimator. We have shown that the addition of an artificial pseudonoise like term is equivalent to a fading of the transition observations using the Dirichlet model; this fading of the observations is similar to fading mechanisms proposed in time-varying parameter estimation techniques, but our pseudonoise-based derivation provides an alternative motivation for actually fading these Dirichlet counts in a perfectly observable system.

Our future work will investigate other forms of non-stationary transition matrices, such as slowly-varying models. Also we will connect the estimation techniques of the transition model to our robust MDP formulations [4] to obtain less conservative robust solutions.

Acknowledgements

Research supported by AFOSR grant FA9550-04-1-0458.

References

[1] A. Barto, S. Bradtke, and S. Singh. Learning to Act using Real-Time Dynamic Programming. *Artificial Intelligence*, 72:81–138, 1993.
[2] D. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 2005.
[3] L. F. Bertuccelli. *Robust Decision-Making with Model Uncertainty in Aerospace Systems.* PhD thesis, MIT, 2008.
[4] L. F. Bertuccelli and J. P. How. Robust Decision-Making for Uncertain Markov Decision Processes Using Sigma Point Sampling. *IEEE American Controls Conference*, 2008.
[5] P. Borkar and P. Varaiya. Adaptive Control of Markov Chains, I: Finite Parameter Set. *IEEE Trans. on Automatic Control*, AC–24(6), 1979.
[6] G. Iyengar. Robust Dynamic Programming. *Math. Oper. Res.*, 30(2):257–280, 2005.
[7] R. Jaulmes, J. Pineau, and D. Precup. Active Learning in Partially Observable Markov Decision Processes. *European Conference on Machine Learning (ECML)*, 2005.
[8] R. Jaulmes, J. Pineau, and D. Precup. Learning in Non-Stationary Partially Observable Markov Decision Processes. *ECML Workshop on Reinforcement Learning in Non-Stationary Environments*, 2005.
[9] V. Jilkov and X. Li. Online Bayesian Estimation of Transition Probabilities for Markovian Jump Systems. *IEEE Trans. on Signal Processing*, 52(6), 2004.
[10] V. Konda and J. Tsitsiklis. Linear stochastic approximation driven by slowly varying Markov chains. *Systems and Control Letters*, 50, 2003.
[11] V. Krishnamurthy and J. B. Moore. On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure. *IEEE Trans on Signal Processing*, 41(8), 1993.
[12] P. R. Kumar and W. Lin. Simultaneous Identification and Adaptive Control of Unknown Systems over Finite Parameters Sets. *IEEE Trans. on Automatic Control*, AC-28(1), 1983.
[13] S. Mannor, D. Simester, P. Sun, and J. Tsitsiklis. Bias and Variance Approximation in Value Function Estimates. *Management Science*, 52(2):308–322, 2007.
[14] P. Marbach. *Simulation-based methods for Markov Decision Processes*. PhD thesis, MIT, 1998.
[15] R. W. Miller. Asymptotic behavior of the Kalman filter with exponential aging. *AIAA Journal*, 9, 1971.
[16] A. Nilim and L. El Ghaoui. Robust Solutions to Markov Decision Problems with Uncertain Transition Matrices. *Operations Research*, 53(5), 2005.
[17] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE Trans.*, 77(2), 1990.
[18] M. Sato, K. Abe, and H. Takeda. Learning Control of Finite Markov Chains with Unknown Transition Probabilities. *IEEE Trans. on Automatic Control*, AC-27(2), 1982.
[19] Y. Bar Shalom, X. Rong Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation.* Wiley Interscience, 2001.