Controlled Markov Processes with Arbitrary Numerical Criteria

Naci Saldi

Department of Mathematics and Statistics

Queen's University

MATH 872 PROJECT REPORT

April 20, 2012

0.1 Introduction

In the theory of controlled Markov process with infinite discrete time horizon and with Borel state and control space, people investigated the various reward or cost functions or criterion. The case of total reward function was studied by D. Blackwell [1], [2] and R. Strauch [3] for the homogenous Markov chains. The results in these papers later generalized to the nonhomongenous Markov processes [4], [5], [6]. The case of mean reward function was first investigated in the papers [7], [8], [9], [10]. Furthermore, other type of reward functions [11], [12], [13] were also considered as well. In addition, instead of focusing on one particular reward function, some researchers studied the certain class of reward functions. One such example for this attempt is the expected utility class [14], [15], [16], [17], [18], [19]. This class contains the reward functions which are expectation of a functional on the trajectory space of the process. Total reward function is clearly in this class. However, this class does not contain all the reward functions that are used in the literature.

In this paper, the author studied the nonhomogenous infinite horizon discrete time controlled Markov process with Borel state and control spaces. He considered the most general class of reward function or criterion, i.e. the reward function is a functional of a probability measure on infinite product space of states and controls induced by the initial measure, transition probability of the chain and the control action. Clearly, this class contains total reward as well as mean reward function. Since obtaining some existence and optimality results are very difficult for this general class, the author introduced some properties of criteria in order to divide this class into smaller pieces of classes in which the problem is tractable. These properties are measurability, convexity and decomposability which will be defined in the next section. Based on the properties that criterion possesses, he gave some important results about the existence of optimal and ε-optimal policies. For example, it is shown that if the criterion is convex, then in each of the classes of Markov, semi-Markov and all strategies, for an arbitrary strategy there exists nonrandomized strategy in this class such that the latter is as good as the former. This implies that for convex criterion we can restrict ourselves just for nonrandomized strategies when we are searching for the optimal policy. This is a very important result which depends on the decomposability of randomized strategies.

The general results obtained for the smaller classes of criteria (e.g. the class of convex and measurable criterions or the class of convex, decomposable and measur-

able criterions) are applied to the concrete examples. For the case of total reward criterion, it is proved that this criterion is convex and decomposable. Hence one can always replace an arbitrary strategy, which has finite criterion value, with the one that is non-randomized Markov without any performance loss, if the criterion is total reward criterion. This is actually a very important result, because generally people presume that the value of the criterion is always finite. At the last section of the paper, there are bunch of examples that investigate the existence and optimality of strategies for the certain criteria.

The most important result that is used in this paper is the decomposability of the randomized strategies. This result was first proved by N.V. Krylov [20] for the case of countable state space and by I.I. Gikhman and A.V. Skorokhod [14] for the case of Borel state space. Later, this result was extended to the case of Markov and semi-Markov strategies [21].

0.2 Background

In this section, I briefly give some notations and then define the probability structure of the problem. The notations used in this paper is from [4].Let E be Borel space(Polish space), then $\mathbf{B}(E)$ denotes the Borel σ-algebra on E and $\Pi(E)$ denotes the set of all probability measures on $(E, \mathbf{B}(E))$. Additionally, $\boldsymbol{M}(E)$ denotes the minimum σ-algebra on $\Pi(E)$ such that $\forall E^{'} \in \mathbf{B}(E)$ the mappings $\mu \longmapsto \mu(E^{'})$, with $\mu \in \Pi(E)$, are measurable [22]. The pair $(\Pi(E), \boldsymbol{M}(E))$ forms a Borel space. The Borel σ-algebra assigned on the set of probability measures $\Pi(E)$ is especially very important for proving certain results and this Borel σ-algebra, I think, is smaller or equal to the Borel σ-algebra generated by the weak convergence of probability measures.

The probability structure of the problem is specified by the following collection : $\{X, A, j(.), p(. \mid .)\}$ where X space of states with $X = \cup_{t=1}^{\infty} X_{t-1}$, A space of control actions with $A = \cup_{t=1}^{\infty} A_t$, j mapping between A onto X and $p(. \mid .)$ regular conditional measure on X given A. In addition, we have the following assumptions:

- $X_{t-1} \in \mathbf{B}(X)$ and $X_{t-1} \cap X_{t^{'}-1} = \emptyset$, $\forall t \neq t^{'}$.
- $A_t \in \mathbf{B}(A)$ and $A_t \cap A_{t^{'}} = \emptyset$, $\forall t \neq t^{'}$.
- $j(A_t) = X_{t-1}$ or $j^{-1}(X_{t-1}) = A_t$ and specifies the admissible control space for state X_{t-1}, $\forall t$.

Observe that if we know a_t, control action at time t, then we also know the state x_{t-1} through the mapping $j(a_t) = x_{t-1}$. Let define the following infinite and finite product spaces $L = X_0 \times A_1 \times X_1 \times A_2 \times X_2 \times \cdots \times A_{t-1} \times X_{t-1} \times \cdots\cdots$, $H_{t-1} = X_0 \times A_1 \times X_1 \times A_2 \times X_2 \times \cdots \times A_{t-1} \times X_{t-1}$ (information space for the controller at time t) and $H = \cup_{t=1}^{\infty} H_{t-1}$. A conditional measure $\pi(. \mid h)$ on A where $h \in H$, concentrated on $A(x_{t-1})$ at time t, satisfying certain regularity conditions is called a "strategy π"; i.e. $\pi(a_t \mid x_0 \times a_1 \times x_1 \times \cdots \times a_{t-1} \times x_{t-1})$. We have the following categorization of the strategies : π is said to be nonrandomized if $\pi(. \mid h) = f(h)$, π is said to be Markov if $\pi(. \mid h) = \pi(. \mid x_{t-1})$ and π is said to be semi-Markov if $\pi(. \mid h) = \pi(. \mid x_{t-1}, x_0)$. Let denote Δ_3^N, Δ_3, Δ_2^N, Δ_2, Δ_1^N and Δ_1, respectively, the set of all "nonrandomized Markov", "Markov", "nonrandomized semi-Markov", "semi-Markov", "nonrandomized" and "randomized" strategies.

It is clear that initial measure μ and strategy π induce a probability measure on $(L, \mathbf{B}(L))$. Let denote this probability measure as P_μ^π. If our initial state is concentrated on the value x then we denote the induced probability measure as P_x^π. Probability measure P on $(L, \mathbf{B}(L))$ is **strategic** if $\exists \mu \in \Pi(X_0)$ and $\pi \in \Delta_1$ s.t. $P = P_\mu^\pi$. We denote S as the set of all strategic measures and $S(x) = \{P \in S : P = P_x^\pi, \pi \in \Delta_1\}$ and $S_0 = \cup_{x \in X_0} S(x)$. The function $w : S \longmapsto [-\infty, +\infty]$ defined on S is called "criterion" and we write $w(P_\mu^\pi) = w(\mu, \pi)$ and $w(P_x^\pi) = w(x, \pi)$. The function w penalizes or gives reward to the probability measure on L induced by the initial measure μ and strategy π. Furthermore, the function v defined by $v(x) = sup\{w(P) : P \in S(x)\}$ is called "value".

As we mentioned in the first section, the problem for general criterion case cannot be tractable if we do not put any restriction on the criterion. For this purpose we have the following definitions which will later help us to classify the class of criteria.

definition 1. *A criterion w is called **measurable**, if $w(P)$ is a measurable function on S.*

definition 2. *A measurable criterion w is **convex**, if for all $\nu \in \Pi(S)$ we have $w(P^\nu) \leq \int_S w(P)\nu(dP)$ where $P^\nu = \int_S P(.)\nu(dP)$.*

definition 3. *A criterion w is called **decomposable**, if for all P and $P' \in S$ we have $P(a_t \in A^t) = P'(a_t \in A^t)$ for all t,$A^t \in \boldsymbol{B}(A_t)$, then $w(P) = w(P')$ holds.*

The measurability of criterion is a very natural restriction on the class of criteria. Furthermore, if our criterion is a function of probability space through control actions only, which is the case in total reward criterion, then the criterion is decomposable.

The convexity property is also very crucial. As we mentioned in the first section the randomized policies can be decomposed as nonrandomized policies or in other words randomized policies are uncountable convex combination of nonrandomized policies. Hence, if the criterion is convex, then by the decomposability result one can show that we can replace any randomized policy with the nonrandomized policy without any performance loss. This implies that the optimal policy lives in the class of nonrandomized policies.

0.3 Results

In this section I will first give very important result that was proven in the paper [21]. It is about the decomposability of randomized strategies.

theorem 1. *Let* $\Omega = [0,1]^{\infty}$, $\boldsymbol{B}^{\infty} = \boldsymbol{B}(\Omega)$. *For an arbitrary strategy* $\pi \in \Delta_i$, $\exists(\boldsymbol{B}^{\infty} \times \boldsymbol{B}(Y^i))$*-measurable mapping* $\phi^i(\omega, y) : \Omega \times Y^i \longmapsto A$ *s.t.*

a) $\phi^i(\omega, y) \in A(y)$, $\forall(\omega, y) \in \Omega \times Y^i$.

b) $\forall \mu \in \Pi(X_0)$ *and* $f(l)$ *on* L

$$\int_L f(l) P_{\mu}^{\pi}(dl) = \int_{\Omega} m(d\omega) \int_L f(l) P_{\mu}^{\phi^i(\omega)}(dl)$$

where $\phi^i(\omega) \in \Delta_i^N$.

This theorem is important for proving the following theorem about the class of convex criteria.

theorem 2. *Let* w *be a convex criterion, then* $\forall \mu \in \Pi(S)$, $K < \infty$ *and* $\pi \in \Delta_i, i = 1, 2, 3$, $\exists$ *a nonrandomized strategy* $\varphi \in \Delta_i^N$ *s.t.*

$$w(\mu, \varphi) \geq \begin{cases} w(\mu, \pi), \text{ if } w(\mu, \pi) < \infty \\ K, \text{ if } w(\mu, \pi) = \infty \end{cases}$$

As stated before, for convex criterion the optimal strategy is nonrandomized if the maximum value of the criterion is finite. Furthermore, we have the following result for the class of decomposable criteria.

theorem 3. *Let* w *be a decomposable criterion, then*

a) For any measure $\mu \in \Pi(X_0)$, *any* $\pi \in \Delta_1$ $\exists$ *Markov strategy* $\sigma \in \Delta_3$ *s.t.* $w(\mu, \sigma) = w(\mu, \pi)$.

b) For any strategy $\pi \in \Delta_1$ $\exists$ *semi-Markov strategy* $\sigma \in \Delta_2$ *s.t.* $w(\mu, \sigma) = w(\mu, \pi)$ $\forall \mu \in \Pi(X_0)$.

If we combine these two theorems, we will get the following corollary which is important for certain class of criteria, in particular for total reward criterion.

corollary 1. *Let* w *be convex decomposable criterion, then*

a) For any measure $\mu \in \Pi(X_0)$, $\pi \in \Delta_1$ *s.t.* $w(\mu, \pi) < \infty$, $\exists$ *nonrandomized Markov strategy* φ *s.t.* $w(\mu, \varphi) \geq w(\mu, \pi)$.

b) For any π *s.t.* $w(x, \pi) < \infty$ $\forall x \in X_0$, $\exists$ *a nonrandomized semi-Markov strategy* φ *s.t.* $w(x, \varphi) \geq w(x, \pi)$ $\forall x \in X_0$.

One can actually show that the total reward criterion given by the following formula $w(P_\mu^\pi) = E_\mu^\pi \sum_{t=1}^{\infty} q(a_t)$ is both decomposable and convex and so the assertions of corollary are valid for it.

0.4 Conclusion

In this paper, author investigates the existence and the optimality results of non-homogenous controlled Markov processes with arbitrary numerical criterion, i.e. the criterion is a function of probability measure on infinite product space induced by the transition probability and control actions. Since the problem is not tractable for general criterion, he introduces three properties of criteria : measurability, convexity and decomposability. These properties are used to classify the criterion set, e.g. the set of convex criteria, the set of decomposable criteria, the set of convex and decomposable criteria etc. With this classification, he is able to give several existence optimality results.

In theorem 2, it is proven that any randomized strategy π can be replaced by the nonrandomized strategy where these two strategy live in the same class, if the criterion is convex. Hence for convex criteria the optimal strategy is in the class of nonrandomized strategies. The proof of this theorem depends on the result given by theorem 1 which states that any randomized policy can be decomposable into nonrandomized policies. This means that randomized policies are uncountable convex combination of nonrandomized policies. In the paper, it is also stated that the space of strategies is convex. If we combine these two results, we can say that the extreme points of the strategy space are nonrandomized policies. Furthermore, if the criterion

is convex, then it is not difficult to see that the optimal policy is nonrandomized.

In theorem 3, it is basically stated that any strategy can be replaced by Markov strategy without any performance loss, if the strategy is decomposable. Hence, it is enough to use just the current state for the control action. If the criterion is a function of control action variables, then it is easy to show that the criterion is decomposable. As a result, if the criterion is both convex and decomposable, then the optimal policy is nonrandomized Markov. The total reward criteria and limit supremum of the average reward criterion are example of convex and decomposable criteria.

In the class, we studied the minimization of average cost criterion or limit supremum of average cost criterion. This problem can be converted to a maximization problem, if limit supremum of average cost criterion is replaced by the negative of limit infimum of average cost criterion. This criterion is both measurable and decomposable but it is not convex. This criterion is actually concave under some further assumptions. Hence the optimal strategy lives in the space of Markov strategies by decomposability but not necessarily in the nanrandomized ones. Actually, since the criterion is concave and the randomized strategies are convex combinations of nonrandomized strategies, it can be shown that the optimal strategy is randomized Markov strategy.

The results given in this paper are very important and universal, because they are applicable to any criterion that is used in the literature and they give good characterization of the optimal strategies. By using these results, one can restrict the search space for the optimal strategy just by verifying that the criterion possesses some certain properties. For instance, if the criterion is convex and decomposable, then the optimal strategy is in the set of nonrandomized strategies. With this information in hand, the solution of the problem may become very easy.

One further attempt to improve this paper is to characterize the set of probability measures induced by the randomized (nonrandomized) Markov or semi-Markov strategies (i.e. proving compactness, completeness of these subsets). Or different topologies(topology induced by total variation metric or topology induced by setwise convergence) can be put on the space of probability measures and they may give better results for certain class of problems.

Bibliography

[1] D. Blackwell, "Dynamic programming in problems with damped action," *Matematika*, vol. 11, pp. 151–160, 1967.

[2] D. Blackwell, "Positive dynamic programming," *Fifth Berkeley Symp. Math. Stat. and Prob.*, vol. 1, pp. 415–418, 1966.

[3] R. E. Strauchl, "Negative dynamic programming," *Amer. Math. Statist.*, vol. 37, pp. 871–890, 1966.

[4] A. A. Y. E. B. Dykin, *Controlled Markov Process*. Springer, 1979.

[5] K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*. Springer-Verlag, 1970.

[6] N. Furukawa, "Markov desicion process with non-stationary laws," *Bull. Math. Statist.*, vol. 13, pp. 41–52, 1968.

[7] R. A. Howard, *Dynamic Programming and Markov Processes*. Wiley, 1960.

[8] C. Derman, "Markov desicion process with non-stationary laws on sequental desicions and markov chains," *Management Sci.*, vol. 9, pp. 16–24, 1962.

[9] C. Derman, "On sequental control processes," *Ann. Math. Statist.*, vol. 35, pp. 341–349, 1964.

[10] A. N. S. O. V. Viskov, "On controls leading to optimal stationary regimes," *Trudy Matem. Steklov Inst.*, vol. 71, pp. 35–45, 1964.

[11] R. Y. Chitashvili, "A finite controlled markov chain with small termination probability," *Theory Prob. Appl.*, vol. 21, pp. 158–163, 1976.

[12] L. J. S. L. E. Dubins, *How to Gamble if You Must*. McGraw-Hill, 1965.

[13] T. P. Hill, "On the existence of good markov strategies," *Trans. Amer. Math. Soc.*, vol. 247, pp. 157–176, 1979.

[14] A. V. S. I. I. Gikhman, *Controlled Stochastic Processes*. Springer, 1979.

[15] S. I. N. Furukawa, "Markovian desicion processes with recursive reward functions," *Bull. Math. Statist.*, vol. 15, pp. 79–91, 1973.

[16] A. Nowak, "On a general dynamic problem," *Colloq. Math.*, vol. 37, pp. 131–138, 1977.

[17] D. M. Kreps, "Desicion problems with expected utility criteria, i: Upper and lower convergent utility, ii:stationarity," *Math. Oper. Res.*, vol. 2, 1, pp. 45–53, 1977.

[18] D. C. N. R. P. Kertz, "Persistently optimal lans for nonstationary dynamic programming: the topology of weak convergence case," *Ann. Probab.*, vol. 7, pp. 811–826, 1979.

[19] M. Schal, "Existence of optimal polcies in stochastic dynamic programming," *University of Bonn*, 1979.

[20] N. V. Krylov, "The construction of an optimal strategy for a finite controlled chain," *Theory Prob. Appl.*, vol. 10, pp. 45–54, 1965.

[21] E. A. Fainberg, "Nonrandomized markov and semi-markov strategies in dynamic programming," *Theory Prob. Appl.*, vol. 27, pp. 116–126, 1982.

[22] D. F. L. Dubins, "Measurable sets of measures," *Pasific J. Math.*, vol. 14, pp. 1211–1222, 1964.