EE 736: Convex Analytic Method for Infinite Horizon Average Cost Optimal Stochastic Control Problems

Vansh Kapoor
200100164

November 19, 2023

1 Introduction

There are 3 major approaches to solve an average cost problem

1. **Relative Value Iteration**: This method is very similar to the commonly used 'value iteration' apart from the fact that the scheme uses an 'offset' which is subtracted during every iteration so as to prevent the iterate from increasing unboundedly. There is a possibility for the iterate to increase unboundedly as the Bellman optimality operator is no longer a contraction mapping in the average cost setup unlike in infinite horizon discounted cost where it was indeed a contraction map.
2. **Vanishing Discount Method**: As the discounting factor $\alpha \to 0$, the Vanishing Discount Method ensures that the limit of the optimum policies for discounted cost problems is the same as the optimal policy for the average cost problem.In summary, the vanishing discount method involves solving a sequence of discounted cost problems for decreasing values of α, and then finding the value of α that minimizes the average cost function. This approach provides a way to solve average cost problems in dynamic programming and obtain an optimal policy that minimizes the long-run average cost.
3. **Convex Analytic Method**: It uses properties of sample path occupation measures and we will be applying this method for optimality results and performance of deterministic policies in average cost stochastic control. This is a versatile approach to optimization of infinite-horizon problems, which avoids the use of dynamic programming and leads to linear program.

In this paper we will be discussing the convex analytic approach to solve an average cost problem. Let us first list the advantages of this method over the other two.

- **Computational efficiency:** Convex analytical methods are generally more computationally efficient than the other two listed methods (especially the iterative methods). Another advantage is that convex optimization algorithms have well-developed numerical libraries and software packages that speed up the computation process.
- **Convergence**: Convex optimization algorithms always converge to the optimal solution. Iterative approaches, on the other hand, may converge to a local minimum or can take a while to converge to the optimal solution
- **Flexibility**: While iterative approaches may be more restricted in the types of issues they can solve than convex analytical methods, convex analytical methods can handle a wide range of problem types, including linear, quadratic, and nonlinear programming problems..

2 Setting Up Notation

1. Let $\boldsymbol{B}(\mathbf{X})$ be the Borel σ-field of a topological space $\mathbf{X}$ and $P(\mathbf{X})$ be the set of probability measures on $\boldsymbol{B}(\mathbf{X})$.
   A controlled Markov Chain is described by the tuple$(X, U, \boldsymbol{U}, T, c)$.
   (a) X is the state space
   (b) U is the control space
   (c) $\boldsymbol{U}$: $X \to \boldsymbol{B}(U)$ is a strict measureable multifunction
   (d) $K := \{(x,u) : x \in X, u \in \boldsymbol{U}(x)\}$ is the set of admissible state/action pairs
   (e) T: $K \to P(X)$ such that $T(\cdot|x,u)$ gives the transition probability for $(x,u) \in K$
   (f) $c : K \to \mathbb{R}_+$ is the running cost which is bounded from below in K and takes values in the interval in $[1, \infty)$
2. Let $M_b(X)(C_b(X)$ be the space of bounded Borel continuous real-valued functions on X
3. Let Γ_A be the set of all admissible policies and Γ_S be the set of all stationary policies
4. $$J^* := \inf_{\gamma \in \Gamma_A} J(x,\gamma) = \inf_{\gamma \in \Gamma_A} \limsup_{N \to \infty} \frac{1}{N} E_x^\gamma \left[\sum_{t=0}^{N-1} c(X_t, U_t) \right]$$

3 Useful Definitions and Properties

3.1 Preliminary Definitions

1. For $\gamma \in \Gamma_S$, we define,
$$T^\gamma(A|x) := \int_{\boldsymbol{U}(x)} T(A|x,u)\gamma(du|x) \tag{1}$$
2. For $\mu \in P(K)$ and $f \in M_b(X)$ we define $\mu T \in P(X)$
$$\mu T(A) := \int_K \mu(dx,du)T(A|x,u), \qquad A \in \boldsymbol{B}(X) \tag{2}$$
and Tf: $K \to \mathbb{R}$
$$Tf(x,u) := \int_K f(y)T(dy|x,u), \qquad (x,u) \in K \tag{3}$$
3. Define integral of functions $\mu(f)$ for $\mu \in P(K), f \in M_b(X)$
$$\mu(f) = \langle \mu \ , \ Tf \rangle = \int_K f(x,u)\mu(dx,du) \tag{4}$$
4. The transition kernel is called *weakly continuous* if for $(x,u) \in K$ the map
$$(x,u) \to \textstyle\int_X f(z)T(dz|x,u)$$
is continuous for $f \in C_b(X)$

5. Let for $N \geq 1$ we define $v_T(D)$ s.t.

$$v_N^\gamma(D) = \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{1}_D(X_t, U_t) \tag{5}$$

$$\mu_N^\gamma(D) = E_v^\gamma[v_N^\gamma(D)] = \frac{1}{N} E_v^\gamma \left[\sum_{t=0}^{N-1} \mathbf{1}_D(X_t, U_t) \right] \tag{6}$$

where γ is a stationary policy, $D \in \boldsymbol{B}(X \times U)$
$\left\{\mu_T^\gamma\right\}_{N>0}$ is the family of *mean empirical occupation measures* under the policy $\gamma \in \Gamma_A$ with initial distribution μ

6. Let us now define the set of *invariant occupation measures* by

$$G := \{\mu \in P(k) : \mu(B \times U) = \mu T(B), B \in \boldsymbol{B}(X)\}$$

7. Let G_e denote the set of extreme points of G

8. Also now we define

$$H := \left\{\pi \in P(X) : \exists\ \gamma \in \Gamma_S \text{ such that } \pi(A) = \textstyle\int_X T^\gamma(A|x)\pi(dx), A \in \boldsymbol{B}(X)\right\}$$

9. Define $\delta^* := \inf_{\mu \in G} \langle \mu T\ ,\ c \rangle$

3.2 Fundamental Properties of Defined functions

1. $\langle \mu T\ ,\ f \rangle = \langle \mu\ ,\ Tf \rangle \quad$ for $\mu \in P(K), f \in M_b(X)$

   Proof. $\langle \mu T\ ,\ f \rangle = \int_K \int_X \mu(dx, du) T(dy|dx, du) f(y)$

   $\Rightarrow \langle \mu T\ ,\ f \rangle = \int_K \int_X \mu(dx, du) T(dy|dx, du) f(y)$

   $\Rightarrow \langle \mu T\ ,\ f \rangle = \int_K \mu(dx, du) \int_X T(dy|dx, du) f(y)$

   $\Rightarrow \langle \mu T\ ,\ f \rangle = \int_K \mu(dx, du) T f(dx, du)$

   $\Rightarrow \langle \mu T\ ,\ f \rangle = \langle \mu\ ,\ Tf \rangle$ □

2. Let $\mu \in G$. Then $\exists\ \phi$ on $X \times \boldsymbol{B}(X)$ an d $\pi \in P(X)$ such that

$$\mu(dx, du) = \phi(du|x)\ \pi(dx)$$

   This is denoted by $\mu = \phi \circledast \pi$. Therefore, if $\gamma \in \Gamma_S$ is any policy which agrees with π almost surely with ϕ then for $A \in \boldsymbol{B}(X)$.

$$\pi(A) = T^\gamma(A|x)\pi(dx)$$

   Note: The converse statement is also true

3. The map $\mu \to \langle \mu\ ,\ c \rangle$ is lower semi-continuous. This is because c is left continuous and bounded from below(by the assumption made on c)

4 Optimality under Weakly Continuous Kernels

In this section we will first find a lower bound on expected cost, based on the assumption that the Transition kernel T is weakly continuous as defined in **Section 3.1** and then establish the conditions under which the lower bound is exactly equal to the expected cost. We extend the application of the theorem to Economics and then rephrase the theorem in Economics jargon.from an state the applications

4.1 Lower Bound on Expected Cost

Assumption (1) The Transition kernel T be a weakly continuous map.

Lemma 4.1. *Under Assumption (1), the limit of any weakly converging sub sequence of mean empirical occupation measures is in G.*

Proof. Using equation (2) and (6), we can conclude that for $\gamma \in \Gamma_A$

$$\begin{aligned} \mu_M^\gamma T(A) &= \frac{1}{N} E_v^\gamma \left[\sum_{t=1}^{N} \mathbf{1}_D(X_t, U_t) \right] \\ |\mu_M^\gamma(A \times U) - \mu_M^\gamma T(A)| &= \frac{1}{M} \left| \left[\sum_{t=1}^{N} \mathbf{1}_D(X_t, U_t) - \sum_{t=1}^{N} \mathbf{1}_D(X_t, U_t) \right] \right| \\ &\leq \frac{1}{N} \to 0 \text{ as } N \to 0 \end{aligned} \tag{7}$$

let along some subsequence t_k, $\mu_t^\gamma \implies \mu$, i.e., μ_t^γ weakly converges to $\mu \in P(K)$
By using triangular inequality,

$$|\mu(f) - \mu T(f)| \leq |\mu(f) - \mu_{t_k}^\gamma(f)| + |\mu_{t_k}^\gamma(f) - \mu_{t_k}^\gamma T(f)| + |\mu_{t_k}^\gamma T(f) - \mu T(f)| \tag{8}$$

(a) $|\mu(f) - \mu_{t_k}^\gamma(f)| \to 0$ as $k \to \infty$ by weak convergence of $\mu_{t_k}^\gamma$.
(b) By the same reasoning as in (a), we can conclude that $|\mu_{t_k}^\gamma T(f) - \mu T(f)| \to 0$ as $k \to \infty$.
(c) Further using equation (7) we can conclude that $|\mu_{t_k}^\gamma(f) - \mu_{t_k}^\gamma T(f)| \to 0$.

$\implies \mu(A, U) = \mu T(A)\ \forall A \in \boldsymbol{B}(X)$
Thus $\mu \in G$ by definition of G.
Note: Though it may seem that the notation $\mu_{t_k}^\gamma(f)$ doesn't make sense since $\mu_{t_k}^\gamma$ is defined on $\boldsymbol{B}(X \times U)$ the notation is consistent since f may be viewed as an element of $C_b(K)$ □

The expected cost
$J^*(x, \gamma) := \limsup_{N \to \infty} \langle \mu T^\gamma\ ,\ c \rangle$
$\implies J(x, \gamma) = \liminf_{t_k \to \infty} \langle \mu_{t_k}^\gamma\ ,\ c \rangle \geq \left\langle \lim_{t_k \to \infty} \mu_{t_k}^\gamma\ ,\ c \right\rangle$
$\implies J(x, \gamma) \geq \langle \mu\ ,\ c \rangle$ (weak convergence)
$\implies J(x, \gamma) \geq \delta^*$ (definition of δ^*)

4.2 Establishing Equality on Lower Bound

Assumption (2)
(A) The state and action spaces X and U are Polish. The set-valued map $U : X \to B(U)$ is upper semicontinuous and closed-valued.
(A') The state and action spaces X and U are compact. The set-valued map $U : X \to B(U)$ is upper semi-continuous and closed-valued.
(B) The non-negative running cost function $c(x, u)$ is lower semi-continuous and $c : K \to R$ is inf-compact, i.e. $(x, u) \in K : c(x, u) \leq \alpha$ is compact for every $\alpha \in \mathbb{R}_+$.
(B') The cost function c is bounded and l.s.c.
(C) There exists a policy and an initial state leading to a finite cost $\eta \in \mathbb{R}_+$.
(D) (H1) holds.
(E) Under every stationary policy, the induced Markov chain is Harris recurrent.

Theorem 4.2. *a) Under Assumption (2): (A, B, C, D) there exists an optimal measure in G. b) Under Assumption 2.1 (A', B', D, E), there exists a stationary policy which is optimal for the control problem*

$$\inf_{\gamma \in \Gamma_A} \limsup_{N \to \infty} \frac{1}{N} E^{\gamma}_{x_0} \left[\sum_{t=1}^{N} c(X_t, U_t) \right]$$

for every initial condition.

Proof. Our first aim is to prove that under Assumptions 2:(A,B,C,D) the below equation holds

$$\langle \mu \ , \ c \rangle = \delta^* \tag{9}$$

Under Assumptions (2): (B,C) its straightforward that $\exists$ set of policies γ such that
$\langle \mu^{\gamma}_N \ , \ c \rangle \leq M < \infty$
Thus along some subsequence $\mu_{t_k} \to \mu \in P(K)$ and using Lemma 4.1 we conclude that $\mu \in G$.

It can be concluded that from hypothesis (A) using Portmanteau theorem that every weak limit of a converging sequence of probability measures on K is also supported on K. (Proof for this subpart has been skipped)
Note: Informally speaking, Portmanteau theorem gives equivalence conditions of weak convergence of a sequence of measures.

The sequence μ_{t_k} by Assumption 2 (B), is tight by inf-compactness and $\mu^* \in G$ where $\mu_{t_k} \to \mu$ with $\mu^*(K) = 1$. Thus we then have an optimal policy ϕ hence $\langle \mu^{\gamma}_N \ , \ c \rangle = \delta^*$ which concludes our first half of our proof.

Moving to the next half of the proof we define a stationary policy γ s.t.

$$\mu^*(dx, du) = \gamma^*(du|x)\pi^*(dx) \tag{10}$$

□

4.3 Can it be interpreted from an Economics Standpoint ?

In economics , a kernel is a function that assigns weights to a set of neighboring points. Informally speaking, a weakly continuous kernel is a function that assigns weights to neighboring points in a way that is continuous and smooth.

In MISO(multi input single output) system the average cost is defined as the cost per unit of output which the total cost of production divided by the total output.
The optimality of average cost under weakly continuous kernels is a result in the theory of production that shows that under certain restrictions, the average cost is the optimal cost per unit of output. More formally, if the production function satisfies certain regularity conditions, and if the kernel is weakly continuous, then the average cost is the minimum cost per unit of output that can be achieved by any production process.

This seems intuitive as if the kernel is weakly continuous, then neighboring points in the input space will be assigned similar weights. This means that if we slightly change the inputs used in the production process, the resulting change in the cost will be small. This property of the kernel(weak continuity) ensures that the production process is robust to small changes in the inputs, and hence the average cost is a good measure of the cost per unit of output. This theory helps us understand how businesses can optimize their production processes in a competitive market.

5 Optimality of Deterministic Stationary Policies

Informally, a deterministic policy is one which explicitly states the action to be performed on the state, i.e., it maps states to actions. Whereas a stochastic policy is one that gives probability of each action in each state. In this section we first discuss the need for optimality of deterministic policies, then move to prove the conditions under which the solution to an optimal average cost stochastic control problem is a deterministic stationary policy.

5.1 Why restrict ourselves to Deterministic Policies ?

The optimality of deterministic stationary policies is an important concept in the decision theory and optimization.It has wide ranging theoretical as well as practical applications. Deterministic policies simplify our decision-making problems by reducing the number of parameters/variables involved. This makes the problem **more tractable** and easier which provide insights to optimal decision making strategies.

Moreover, the optimality of deterministic stationary policies also has several practical applications. In many real-world applications such as engineering, finance, trading it is often difficult or expensive to obtain real-time information about the state of the system. Deterministic stationary policies can provide a simple and effective way of making decisions in such situations, as they do not rely on real-time information and can be implemented easily.

5.2 Optimality of Deterministic Policies Under Countable State/Action Space Setup

We already have proved in **Lemma 4.1** that G is closed under weak convergence, we can also show that G is convex, i.e., if $\mu^1, \mu^2 \in G$ then for every $\kappa \in (0,1)$
$\mu(dx, du) := \kappa\mu^1(dx, du) + (1-\kappa)\mu^2(dx, du)$
$\mu(dx, du) \in G$
If ϕ is a non-deterministic policy, we claim (without proving)that we can select $\alpha \in X, \theta \in (0,1)$ and probability measures γ_1, γ_2 on U s.t.

$$\phi(du|\alpha) = \theta\gamma_1(du) + (1-\theta)\gamma_2(du) \tag{11}$$

Lemma 5.1. *We assume that the chain is controlled by some $\phi \in \Gamma_S$ has an invariant probability measure π_ϕ. Suppose that ϕ is non-deterministic on some set that has positive π_ϕ measure. Then the corresponding invariant occupation measure μ_ϕ cannot lie in G_e.*

Proof. Let ϕ be a non-deterministic policy and let $\phi_i, i \in (0,1)$ be two Markov Policies that select action wrt the probability distribution given by γ_i at state α and is consistent with ϕ at every other state.
Now define τ_α be the first hitting/return time to state α Under a given policy, the MDP becomes a Markov Chain. By **Renewal Reward Theory** of Markov Chains

$$\pi_{\phi_i}(x) = \frac{E_\alpha^{\phi_i}\left[\sum_{k=0}^{\tau_\alpha - 1} \mathbf{1}_{\{X_k = x\}}\right]}{E_\alpha^{\phi_i}[\tau_\alpha]} \qquad i = 1, 2 \tag{12}$$

Similarly, by using condition expectation on the expected reward obtained, in this case which is the expected number of hits to state x in an excursion, we obtain

$$\pi_\phi(x) = \frac{\theta E_\alpha^{\phi_1}\left[\sum_{k=0}^{\tau_\alpha - 1} \mathbf{1}_{\{X_k = x\}}\right] + (1-\theta) E_\alpha^{\phi_2}\left[\sum_{k=0}^{\tau_\alpha - 1} \mathbf{1}_{\{X_k = x\}}\right]}{\theta E_\alpha^{\phi_1}[\tau_\alpha] + (1-\theta) E_\alpha^{\phi_2}[\tau_\alpha]} \tag{13}$$

Thus $\pi_\phi = \kappa \pi_{\phi_1}(x) + (1-\kappa)\pi_{\phi_2}$
Hence, $\mu_\phi \notin G_e$ □

5.2.1 Establishing Optimality

To show the optimality of the Deterministic policies, we characterize the extreme points of the convex set G. We use the fact that the optimal policy can be found over the extreme points of the set G(due to LP formulation). Using this result as well as Lemma 5.1 we can conclude that for countable state/action space setup an optimal policy is *stationary as well as deterministic*(provided convex analytic method can be applied)

6 Denseness of Performance of Stationary Policies

In many applications it is essential to know not only that the optimal policies are deterministic, but also if they are "dense". We first define denseness in this context both formally and intuitively. We then motivate the need for denseness of stationary deterministic policies in the context of real life applications and then finally state the denseness result.
Furthermore, the dense set of deterministic and stationary policies can be assumed to have finite range

6.1 Denseness of Policies

We refer a policy as dense if , for any given $\epsilon > 0$, $\exists$ a policy γ such that under γ, it achieves a performance (Value function) within ϵ of that of the optimal performance (for all the states). Intuitively this means that we can find a policy that achieves a performance arbitrarily close to the optimal one. If we consider a sequence of dense policies, each one being more optimal than the previous one, we can construct a sequence whose difference in performance becomes arbitrarily small as the sequence progresses and approaches the optimal policy.

6.2 Applications

Denseness of policies is essential because with dense policies there are no fundamental limitations to achieving arbitrarily high performance up to the optimal performance.Instead, if a policy is dense it means that with sufficient effort we can get as close to optimal performance as we want though never quite achieve it. It has wide ranging applications in almost every field: from machine learning to finance.

- **Robotics**: Denseness in this context would mean that we would be able to design efficient control policies that maximizes the robot's performance while simultaneously minimizing the cost of control, thus approaching the optimal performance.
- **Gaming**: When building AI agents for video games, denseness of performance might be useful in achieving optimal or almost ideal gameplay. We can build AI agents that can perform close to the best possible in games by employing a well-designed stationary deterministic policy.
- **Operations research**: Denseness of performance/policies in operations research is extremely helpful in optimizing complex systems, such as supply chain management,transportation and scheduling.
- **Finance**: While creating investing strategies in finance, denseness of performance can be helpful in maximising returns while lowering risk. In our investing decisions, we can obtain a high level of performance by utilising a well-designed stationary deterministic strategy.
- **Transportation**: Designing traffic control strategies that reduce congestion and travel time in transportation can benefit from performance density.

6.3 Stating the Theorem

Theorem 6.1. *Suppose that*

- *G is weakly compact. Furthermore $X = \mathbb{R}^n$ for some finite n, and for all $x \in \mathbb{R}$, $U(x) = U$ is compact.*
- *For some $\alpha \in [0, 1)$, under every stationary policy γ the induced kernel P_γ of the Markov chain given by*

$$P_\gamma(\pi)(\cdot) := (\pi T_\gamma)(\cdot) = f_\pi(dx)\gamma(du|x)f_T(\cdot|x, u)$$

satisfies

$$||P_\gamma(\pi) - P_\gamma(\bar{\pi})||_{TV} \leq \alpha||\pi - \bar{\pi}||_{TV}, \tag{14}$$

for any pair of probability measures $(\pi, \bar{\pi})$. This condition implies, naturally, that every stationary policy leads to a unique invariant probability measure.

- *The kernel $T(dy|x, u)$ is such that, the family of conditional probability measures $T(dy|x, u), x \in X, u \in U$ admit densities $f_{x,u}(y)$ with respect to a reference measure and all such densities are bounded and equicontinuous (over $x \in X, u \in U$).*
- *One of the following holds: T is weakly continuous holds and the bounded cost function $c(x, u)$ is continuous;*
  or
  For any $x \in X$, the map $u \to \int f(z)T(dz|x, u)$ is continuous for every bounded measurable function f and the bounded cost function $c(x, u)$ is continuous in u for every x.

Then, deterministic and stationary policies are dense among those that are randomized and stationary, in the sense that the cost under any randomized stationary policy can be approximated arbitrarily well by deterministic and stationary policies. Furthermore, the dense set of deterministic and stationary policies can be assumed to have finite range.

Note: There is a small typo in (26) in the paper where ϕ^1 should be replaced with ϕ^i

7 Summary

We have begun this report by stating the three popular approaches to solve an average cost problem and stated the advantages of Convex Analytic Method over the other two. We then found a lower bound on average cost and established the conditions under which the equality between the two hold. We then moved on to state the conditions under which deterministic policies are optimal and also listed the applications of this result in various fields. We finally present a denseness result of costs induced under deterministic and stationary policies among those that are attained by randomized and stationary policies.

References

[1] Ari Arapostathis, Serdar Yüksel, Convex analytic method revisited: Further optimality results and performance of deterministic policies in average cost stochastic control, Journal of Mathematical Analysis and Applications, Elsevier, 2023

[2] V.S. Borkar, S.K. Mitter, S. Tatikonda, Optimal sequential vector quantization of Markov sources, SIAM J. Control Optim. 40 (2001) 135–148.

[3] V.S. Borkar, Convex analytic methods in Markov decision processes, in: E.A. Feinberg, A. Shwartz (Eds.), Handbook of Markov Decision Processes, Kluwer, Boston, MA, 2001, pp. 347–375.

[4] A. Arapostathis, V.S. Borkar, Average cost optimal control under weak ergodicity hypotheses: relative value iterations, Arxiv preprints, arXiv:1902.01048, 2019

[5] A. Arapostathis, Some new results on sample path optimality in ergodic control of diffusions, IEEE Trans. Autom. Control 62 (10) (2017) 5351–5356.