MARKOV DECISION PROCESSES

Vivek Borkar

IIT BOMBAY

June 6/13, 2024, LAAS Toulouse

CONTROLLED MARKOV CHAINS

A controlled Markov chain or Markov Decision Process (MDP for short) $\{X_n\}$ taking values in a finite state space S and controlled by a control process $\{Z_n\}$ taking values in a finite action or control space U satisfies the 'controlled Markov property' :

$$P(X_{n+1} = j | X_m, Z_m, m \leq n) = P(X_{n+1} = j | X_n, Z_n)$$

$$= p(j | X_n, Z_n) \text{ a.s.},$$

for a *controlled* transition probability function

$$(i, j, u) \in S^2 \times U \mapsto p(j|i, u) \in [0, 1],$$

$$\sum_j p(j|i, u) = 1 \ \forall \ i \in S, u \in U.$$

Intuitively, the choice of control can depend on the past history, i.e., $X_m, Z_m, m < n$, the current state X_n, and any independent extraneous randomization. $\{Z_n\}$ is called an admissible control.

Let $\mathcal{N} := \{0, 1, 2, \cdots\}$. If $Z_n = v(X_n, n)\ \forall\ n \in \mathcal{N}$ for some $v : S \times \mathcal{N} \mapsto U$, then we say that $\{Z_n\}$ or, by abuse of terminology, $v(\cdot, \cdot)$ itself, is a Markov policy (or strategy).

If $Z_n = v(X_n)\ \forall n \in \mathcal{N}$ for some $v(\cdot) : S \mapsto U$, then $\{Z_n\}$ or equivalently, $v(\cdot)$, is called a stationary policy. In other words, a stationary policy is a Markov policy, but with no explicit dependence on n.

Sometimes we allow for independent randomization in a Markov or stationary policy, i.e., specify a $\phi : S \times \{0, 1, 2, \cdots\} \mapsto \mathcal{P}(U)$ (resp., $\varphi : S \mapsto \mathcal{P}(U)$) so that Z_n is picked with conditional distribution ϕ (resp., φ) given X_n, conditionally independent of $X_m, Z_m, m < n$, given X_n. These are called resp., randomized Markov or randomized stationary policies.

In Markov or randomized Markov policies, the control choice depends only on current state and time which makes $\{X_n\}$ a possibly time-inhomogeneous Markov chain.

In randomized stationary or stationary policies, the control choice depends only on the current state, which makes $\{X_n\}$ a time-homogeneous Markov chain.

For simplicity and ease of notation, we assume that every action is feasible in every state, i.e. the action space U does not depend on the state.

More general scenario wherein the action space depends on the state is possible (e.g., in queues where departures are controlled, they cannot exceed the state, i.e., the current queue length). The theory we develop needs only minor modifications to accommodate this.

More precisely, we can have an action space U_i for state $i, i \in S$. But this can be reduced to the previous case by replacing each U_i by the common $U := \prod_i U_i$ and redefiing the transition probability as the transition probabiity

$$p'(j|i, [u_1, u_2, \cdots]) := p(j|i, u_i).$$

The objective is to maximize a suitable reward or minimize a suitable cost.

We shall mostly stick to the latter, the treatment of the former being analogous (with one exception we mention later).

Some common cost criteria are as follows.

1. **Finite horizon cost:** Let the 'time horizon' $T > 0$ be prescribed. Minimize

$$E\left[\sum_{t=0}^{T-1} k(t, X_t, Z_t) + h(X_T)\right].$$

Here $k : \{0, 1, 2, \cdots\} \times S \times U \mapsto \mathcal{R}$ is called the running cost function and $k(t, X_t, Z_t)$ the running cost.

Likewise, $h(X_T)$ is called the terminal cost with X_T as the terminal state.

We can also write the above as

$$E\left[\sum_{t=0}^{T} k(t, X_t, Z_t)\right]$$

by setting $k(T, X_T, Z_T) = h(X_T)$, but usually the previous format is preferred.

We can also consider running costs of the type $k(t, X_t, X_{t+1}, Z_t)$, which are equivalent to

$$\bar{k}(t, X_t, Z_t) := \sum_j p(j|X_t, Z_t) k(t, X_t, j, Z_t).$$

Similar remarks apply to other cost crteria as well.

2. **Infinite horizon discounted cost:** Minimize

$$E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)\right]$$

where $\alpha \in (0, 1)$ is the discount factor. In addition to ensuring the summability of the series above, it has practical interpretation: expenditure now is more costly than expenditure later because of interest accrued etc. Thus this criterion is popular when money is involved, usually with $\alpha = \frac{1}{1+r}$ where r is the interest rate.

3. **Average (or 'ergodic') cost:** Minimize

$$\limsup_{T\uparrow\infty} \frac{1}{T} E\left[\sum_{t=0}^{T-1} k(X_t, Z_t)\right].$$

This is used when short term transients are deemed unimportant and long term or 'equilibrium' behaviour is of primary interest.

4. **Risk-sensitive cost:** Minimize

$$\limsup_{T\uparrow\infty} \frac{1}{T} \log E\left[e^{\sum_{t=0}^{T-1} k(X_t, Z_t)}\right]$$

which seeks to minimize the mean exponential growth rate. This criterion is used when variation around the mean cost also needs to be accounted for, or when compounding effects make the criterion more natural, e.g., in finance.

We call a random variable τ taking values in $\{0, 1, 2, \cdots; \infty\}$ a *stopping time* if the event $\{\tau \leq n\}$ (equivalently, $\{\tau = n\}$) depends only on $\{X_m, Z_m, m \leq n\}$ for all $n \geq 0$.

(Convince yourself that the two definitions are equivalent.)

There are cost criteria that involve optimization over stopping times. We take these up later.

We shall denote by P_u for $u = [u_1, u_2, \cdots]$ the stochastic matrix whose (i, j)th element is $p(j|i, u_i)$.

By abuse of terminology, for a stationary policy v, we shall denote by P_v the stochastic matrix whose (i, j)th element is $p(j|i, v(i))$.

Similar notation is use for a randomized stationary policy φ where these get replaced by P_φ and $\sum_u p(j|i, u)\varphi(u|i)$ respectively.

Dynamic and Linear Programming

Dynamic programming is a general purpose approach to sequential, multi-stage decision making. In words, the dynamic programming principle says that the minimum cost to go at a given stage is the minimum of the expected sum of the immediate cost and the minimum cost to go from the next stage*.

The important thing here is the backward recursion implicit in this statement. This is best illustrated by working it out for the finite horizon problem.

*as perceived in the present stage

Recall the finite horizon cost

$$E\left[\sum_{t=0}^{T-1} k(t, X_t, Z_t) + h(X_T)\right].$$

Define the value function $V : (t, i) \in \{0, 1, \cdots, T\} \times S \mapsto V(t, i) \in \mathcal{R}$ as the 'minimum cost to go' when you are at state i at time t. That is,

$$V(t, i) := \min E\left[\sum_{m=t}^{T-1} k(m, X_m, Z_m) + h(X_T) | X_t = i\right]$$

where the minimum is over all admissible controls $Z_m, t \leq m < T$.

Then the dynamic programming principle leads to the equations

$$\begin{aligned}
V(t,i) &= \min E\left[k(t,X_t,Z_t) + V(t+1,X_{t+1})|X_t = i\right] \\
&= \min E[k(t,X_t,Z_t) + \\
&\quad E\left[V(t+1,X_{t+1})|X_m, Z_m, m \le t\right]|X_t = i] \\
&= \min E\left[k(t,X_t,Z_t) + \sum_j p(j|X_t,Z_t)V(t+1,j)|X_t = i\right] \\
&= \min_u \left[k(t,i,u) + \sum_j p(j|i,u)V(t+1,j)\right]
\end{aligned}$$

for $t < T$, with terminal condition $V(T,i) = h(i)$.

Here, the first equality is from the dynamic programming principle,

the second by additional conditioning,

the third by the controlled Markov property, and,

the fourth by the fact that $X_t = i$ and minimizing over Z_t reduces to minimizing over u.

The equality between the leftmost and the rightmost expression constitutes the dynamic programming equation for this problem, i.e.,

$$V(t,i) = \min_u \left[k(t,i,u) + \sum_j p(j|i,u) V(t+1,j) \right], 0 \leq t \leq T,$$

$$V(T,i) = h(i), i \in S.$$

Note that in the case of finite action spaces, the 'min' in all the equalities exists. If the number of actions is, say, countably infinite and the 'min' does not exist, the same arguments go through with 'inf'.

Formal proof: Suppose $X_t = i$. If we use control u at time t and from time $t+1$, use an ϵ-optimal control for initial condition X_{t+1}, then

$$\begin{aligned} V(t,i) &\leq k(t,i,u) + E\left[V(t+1, X_{t+1}) + \epsilon | X_t = i, Z_t = u\right] \\ &= k(t,i,u) + \sum_j p(j|i,u) V(t+1,j) + \epsilon \end{aligned}$$

$$\Longrightarrow V(t,i) \leq \min_u \left[k(t,i,u) + \sum_j p(j|i,u) V(t+1,j) \right] + \epsilon$$

$$\Longrightarrow V(t,i) \leq \min_u \left[k(t,i,u) + \sum_j p(j|i,u) V(t+1,j) \right] \qquad (*)$$

where we let $\epsilon \to 0$ using the fact that $\epsilon > 0$ was arbitrary.

If the inequality in (*) is strict for some t, i, then there exists a $\delta > 0$ such that

$$V(t,i) + \delta \leq \min_u \left[k(t,i,u) + \sum_j p(j|i,u) V(t+1,j) \right]$$

$$\Longrightarrow V(t,i) + \delta \leq E\left[k(t,i,Z_t) + V(t+1, X_{t+1}) \right]$$

under any admissible $\{Z_n\}$. Iterating (*),

$$V(t,i) + \delta \leq E\left[\sum_{s=t}^{T-1} k(s, X_s, Z_s) + h(X_T) | X_t = i \right] \Longrightarrow$$

$$V(t,i) + \delta \leq \inf_{\{Z_m\}} E\left[\sum_{s=t}^{T-1} k(s, X_s, Z_s) + h(X_T) | X_t = i \right],$$

a contradiction. The claim follows.

Next, let $v(t,i)$ minimizer the RHS of the DP equation. Consider the Markov policy $Z_t = v(t, X_t)$. For $X_0 = i$,

$$V(t,X_t) = k(t,X_t,Z_t) + E\left[V(t+1,X_{t+1})|X_m, m \le t\right],\ t < T.$$

Taking expectations,

$$E\left[V(t,X_t)\right] = E\left[k(t,X_t,Z_t) + V(t+1,X_{t+1})\right],\ t < T.$$

Summing over $t = 0$ to $T-1$ and canceling common terms on both sides, we get

$$V(0,i) = E\left[\sum_{t=0}^{T-1} k(t,X_t,Z_t) + h(X_T)\right]$$

where we use $V(T,j) = h(j)\ \forall j$. So $\{Z_t\}$ is optimal.

Uniqueness of solution: For *any* solution $V(\cdot,\cdot)$ of the dynamic programming equations, we can establish the last equality for the corresponding choice of $\{Z_t\}$.

For any other admissible control $\{Z'_t\}$ and the corresponding state process $\{X'_t\}$, one has $\leq$ in place of equality above when $\{X_t, Z_t\}$ are replaced by $\{X'_t, Z'_t\}$.

Hence $V(0,i)$ is the minimum cost to go from state i at time 0 for $i \in S$.

A similar argument works for $V(t,i), i \in S, 1 \leq t < T$.

Summarizing,

Theorem The value function is the unique solution to the dynamic programming equations. Moreover, $\{Z_t\}$ is optimal if and only if for $0 \leq t < T$,

$$Z_t \in \text{ Argmin}(k(t, X_t, \cdot) + \sum_j p(j|X_t, \cdot)V(t+1, j)).$$

The Markov policy $v(t, i), 0 \leq t < T, i \in S$, is optimal for any initial time $t < T$ and initial state $i \in S$ if and only if $v(t, i)$ minimizes the RHS of the DP equation for all choices of (t, i).

Thus we have the following recipe for finding the optimal policy:

1. Solve the DP equation for $V(\cdot, \cdot)$ by backward recursion starting with the terminal condition $V(T, \cdot) = h(\cdot)$ and successively computing $V(t, \cdot)$ from $V(t+1, \cdot)$ using the DP equation.

2. Find the minimizer $v(t, i)$ by explicit minimization in the DP equation for $i \in S$ and $0 \leq t < T$, and set $Z_t = v(t, X_t), 0 \leq t < T$.

There is also a linear programming formulation for this problem. Define the 'occupation measure'

$$\mu(t, i, u) = P(X_t = i, Z_t = u).$$

Setting $k(T, i, u) = h(i)$, $\forall u$, we can then equivalently write the cost as:

$$\sum_{t,i,u} \mu(t, i, u) k(t, i, u).$$

Let $\lambda(i) := P(X_0 = i)$ $\forall i$, i.e., λ is the initial distribution of the state.

Then our optimization problem becomes the LP:

Minimize

$$\sum_{t,i,u} \mu(t,i,u)k(t,i,u)$$

subject to:

$$\begin{aligned}
\mu(t,i,u) &\geq 0, && (1)\\
\sum_{i,u} \mu(t,i,u) &= 1, && (2)\\
\sum_{u} \mu(t+1,i,u) &= \sum_{j,u} \mu(t,j,u)p(i|j,u), \ 0 \leq t < T, && (3)\\
\sum_{u} \mu(0,i,u) &= \lambda(i). && (4)
\end{aligned}$$

That the occupation measure satisfies (1)-(4) is obvious. For the converse, define the randomized Markov policy ϕ by

$$\phi(u|i,t) := \frac{\mu(t,i,u)}{\sum_u' \mu(t,i,u')}.$$

Then one can check by induction that the above constraints lead to

$$\mu(t,i,u) = P(X_t = i, Z_t = u) \text{ for } (X_t, Z_t), t \geq 0,$$

controlled by the randomized Markov policy ϕ.

On the other hand, one can also check by induction that for any state-action sequence $(X_t, U_t), 0 \leq t \leq T$, the corresponding occupation measure is exactly the same as that for the randomized Markov policy ϕ defined by

$$\phi(u|i,t) := P(Z_t = u|X_t = i).$$

This is another way of seeing that randomized Markov policies suffice.

Conditions (1)-(4) completely characterize possible μ, which therefore form a convex polytope.

Given an optimal solution $\mu(\cdot,\cdot,\cdot)$, an optimal randomized Markov policy is given by: pick $Z_t = u$ with probability $\phi_t(u|X_t)$ for $0 \leq t < T$, where

$$\phi_t(u|i) := \frac{\mu(t,i,u)}{\sum_{u'} \mu(t,i,u')}.$$

It can be shown that the extreme points (or 'corners') of this polytope correspond to Markov policies, implying existence of an optimal Markov policy.

The dual linear program turns out to be:

Maximize $\sum_i \lambda(i) V(0,i)$ subject to:

$$V(t,i) \leq k(t,i,u) + \sum_j p(j|i,u) V(t+1,j), \ \forall \ i,u, \qquad (\dagger)$$

$$V(T,i) \leq h(i).$$

The optimal V turns out to be precisely the value function. An optimal Markov policy then is $v(t,i) =$ any u for which $(\dagger)$ becomes an equality for the optimal V.

Discounted Cost

The infinite horizon discounted cost is given by

$$E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)\right] \tag{5}$$

with discount factor $0 < \alpha < 1$ and running cost $k : S \times U \mapsto \mathcal{R}$.

No 'terminal time' $\Longrightarrow$ no simple backward recursion.

But the possible future costs from initial state (say) i look the same regardless of when you arrive at i.

Thus we can define value function $V : S \mapsto \mathcal{R}$, i.e., a function of state alone, by

$$V(i) := \inf E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t) | X_0 = i\right] \qquad (6)$$

where the infimum is over all admissible controls.

Dynamic programming principle can be applied as before to write the corresponding DP equation as

$$V(i) = \min_u \left(k(i,u) + \alpha \sum_j p(j|i,u) V(j)\right), \ i \in S. \qquad (*)$$

This can also be proved formally as for the finite horizon case by using 'ϵ-optimal controls'.

Thus with $\epsilon > 0$, for $X_0 = i$, pick control u at time 0 and an ϵ-optimal control thereafter. Then

$$\begin{aligned}
V(i) &\leq k(i,u) + E[\sum_{t=1}^{\infty} \alpha^t k(X_t, Z_t) | X_0 = i] \\
&= k(i,u) + \alpha E[\sum_{t=1}^{\infty} \alpha^{t-1} k(X_t, Z_t) | X_0 = i] \\
&\leq k(i,u) + E[\alpha(V(X_1) + \epsilon) | X_0 = i] \\
&\leq k(i,u) + \alpha \sum_j p(j|i,u) V(j) + \epsilon \\
\Longrightarrow V(i) &\leq \min_u [k(i,u) + \alpha \sum_j p(j|i,u) V(j)] + \epsilon \\
\Longrightarrow V(i) &\leq \min_u [k(i,u) + \alpha \sum_j p(j|i,u) V(j)].
\end{aligned}$$

If the inequality is strict for some i, then with $X_0 = i$, for some $\delta > 0$ and any admissible $\{Z_t\}$,

$$\begin{aligned}
V(i) + \delta &\leq \min_u [k(i,u) + \alpha \sum_j p(j|i,u) V(j)] \\
&\leq E[k(X_0, Z_0) + \alpha E[V(X_1)|X_0]|X_0 = i] \\
&= E[k(X_0, Z_0) + \alpha V(X_1)|X_0 = i] \\
\text{(iterating)} \quad &\leq E[\sum_{t=0}^{T} \alpha^t k(X_t, Z_t) + \alpha^{T+1} V(X_{T+1}|X_0 = i] \\
&\to E[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)|X_0 = i] \\
\Longrightarrow & \\
V(i) + \delta &\leq \min_{\{Z_t\}} E[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)|X_0 = i] = V(i),
\end{aligned}$$

a contradiction. Hence the DP equation holds.

Let V be *some* solution of this equation. If $v(i), i \in S$, is a minimizer on the RHS, then for $Z_t = v(X_t), t \geq 0$, by arguments analogous to those for the finite horizon case,

$$\begin{aligned} E\left[V(X_t)\right] &= E\left[k(X_t, Z_t) + \alpha V(X_{t+1})\right] \\ &\Longrightarrow \\ E\left[V(X_0)\right] &= E\left[\sum_{t=0}^{T} \alpha^t k(X_t, Z_t)\right] + \alpha^{T+1} E\left[V(X_{T+1})\right] \\ &\qquad \text{(by iterating)} \\ &\overset{T\uparrow\infty}{\rightarrow} E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)\right]. \end{aligned}$$

For arbitrary $(X'_t, Z'_t), t \geq 0$, we similarly get

$$E\left[V(X_0)\right] \leq E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, Z_t)\right].$$

Thus $V(i) =$ the minimum discounted cost starting from $i \implies V$ is the unique solution to the DP equation.

Also, a stationary policy $Z_t = v(X_t), t \geq 0$, is optimal for any intial condition if and only if $v(X_t)$ minimizes $k(X_t, \cdot) + \alpha \sum_j p(j|X_t, \cdot)V(j)$ for $t \geq 0$.

In particular, if we solve the DP equation for V, an optimal stationary policy can be found by minimizing its right hand side for each i.

The dynamic programming equation gives V in terms of itself, i.e., it is a fixed point equation. One algorithm for solving it is 'value iteration' that begins with a guess V_0 and successively computes

$$V_{n+1}(i) = \min_u \left(k(i,u) + \alpha \sum_j p(j|i,u) V_n(j) \right), \ i \in S, n \geq 0.$$

Subtract from the LHS, resp. RHS of this equation the LHS, resp. RHS of the DP equation and take absolute values.

Then

$$
\begin{aligned}
|V_{n+1}(i) - V(i)| &\leq |\min_u(k(i,u) + \alpha \sum_j p(j|i,u)V_n(j)) - \\
&\quad \min_u(k(i,u) + \alpha \sum_j p(j|i,u)V(j))| \\
&\leq \max_u |(k(i,u) + \alpha \sum_j p(j|i,u)V_n(j)) - \\
&\quad (k(i,u) + \alpha \sum_j p(j|i,u)V(j))| \\
&= \alpha \max_u |\sum_j p(j|i,u)(V_n(j) - V(j))| \\
&\leq \alpha \max_j |V_n(j) - V(j)|, \ \forall i
\end{aligned}
$$

$$
\Longrightarrow \max_i |V_{n+1}(i) - V(i)| \leq \alpha \max_i |V_n(i) - V(i)|
$$

$$
\Longrightarrow \max_i |V_n(i) - V(i)| \leq \alpha^n \max_i |V_0(i) - V(i)| \to 0 \text{ as } n \to \infty.
$$

An alternative scheme is 'policy iteration', which begins with an initial guess $v_0 : S \mapsto U$ for an optimal stationary policy and at nth iterate, does the following:

Given the current guess $v_n(\cdot)$ for the optimal stationary policy,

1. Solve the linear system

$$V_n(i) = k(i, v_n(i)) + \alpha \sum_j p(j|i, v_n(i)) V_n(j), \ i \in S, \quad (7)$$

for V_n, the value function corresponding to policy v_n.

2. Find a stationary policy v_{n+1} such that

$$k(i, v_{n+1}(i)) + \alpha \sum_j p(j|i, v_{n+1}(i)) V_n(j) =$$

$$\min_u \left(k(i,u) + \alpha \sum_j p(j|i,u) V_n(j) \right) \quad (8)$$

$$\leq \; k(i, v_n(i)) + \alpha \sum_j p(j|i, v_n(i)) V_n(j) \;\; \forall i. \quad (9)$$

3. Stop if equality holds in (9) and declare v_n as an optimal stationary policy, otherwise repeat with $n \to n+1$.

The legitimacy of the termination criterion follows from the fact that in case of equality, $V_n = V$.

By arguments similar to those used above, iterating (7), we see that $V_n(i) =$ the cost under the stationary policy v_n, i.e.,

$$E\left[\sum_{t=0}^{\infty} \alpha^t k(X_t, v_n(X_t)) | X_0 = i\right].$$

Likewise it follows from (9) that except in the termination step,

$$V_n(i) \geq k(i, v_{n+1}(i)) + \alpha \sum_j p(j|i, v_{n+1}(i)) V_n(j) \ \forall i$$

with at least one inequality strict, say the ith.

Iterating, we then get

$$V_n(i) > E\left[\sum_{t=0}^{\infty} \alpha^t k(X'_t, v_{n+1}(X'_t))|X_0 = i\right] = V_{n+1}(i),$$

where $\{X'_t\}$ is controlled by the stationary policy v_{n+1}.

Thus $v_n(\cdot)$ has non-increasing cost as n increases, which is strictly decreasing in at least one component if v_n is not optimal.

Since there are only finitely many stationary policies $\left(|U|^{|S|}\right.$ to be precise$\left.\right)$, v_n converges to an optimal policy in a finite time.

Next define the discounted occupation measure $\mu : (i,u) \in S \times U \mapsto \mu(i,u) \in \left[0, \frac{1}{1-\alpha}\right]$ by:

$$\mu(i,u) := \sum_{t=0}^{\infty} \alpha^t P(X_t = i, Z_t = u).$$

Then the cost becomes $\sum_{i,u} \mu(i,u)k(i,u)$, leading to the linear program: for initial distribution $\lambda(\cdot)$,

Minimize $\sum_{i,u} \mu(i,u)k(i,u)$ subject to:

$$\begin{aligned} \mu(i,u) &\geq 0 \ \forall \ i,u, && (10)\\ \sum_{i,u} \mu(i,u) &= \frac{1}{1-\alpha}, && (11)\\ \sum_{u} \mu(i,u) &= \lambda(i) + \alpha \sum_{j,u} p(i|j,u)\mu(j,u) \ \forall i. && (12) \end{aligned}$$

As for finite horizon problem, (10)-(12) characterize μ.

If μ^* is an optimal μ, then the randomized stationary policy that chooses at state i the control u with probability $\frac{\mu^*(i,u)}{\sum_a \mu^*(i,a)}$ is optimal.

The set of μ is a nonempty convex polytope whose extreme points can be shown to correspond to stationary policies. This implies the existence of an optimal stationary policy.

The dual linear programme is, for initial distribution λ,

Maximize $\sum_i \lambda(i)V(i)$ subject to

$$V(i) \leq k(i,u) + \alpha \sum_j p(j|i,u)V(j).$$

Again, when $\lambda(i) > 0 \ \forall i$, the optimal solution coincides with the value function. The optimal choices of v are those that achieve equality in the above constraints.

Note that in absence of any irreducibility assumption, if we drop the condition $\lambda(i) > 0 \ \forall i$, some states may never be visited and the control choice there is irrelevant, as is the value function.

Stochastic shortest path

Consider the controlled Markov chain $\{(X_t, Z_t)\}$ as before with finite state space S decomposed as $S = S_0 \cup A$ with $S_0 \cap A = \phi$.

The states in S_0 are referred to as nonterminal states while the states in A are called terminal states.

The set A may be taken to be a set of absorbing states as will become clear later.

Define the first passage time τ to A as

$$\tau := \min\{t \geq 0 : X_t \in A\},$$

with the convention $\tau = \infty$ if the set on the right is empty. Then $\tau = 0$ when $X_0 \in A$.

We assume for simplicity that under any stationary policy, there is a path from any $i \in S_0$ to some $j \in A$.
It follows that there exists an integer $K \geq 0$ and $1 > \delta > 0$ such that under any stationary policy,

$$\max_i P(\tau > K | X_0 = i) < \delta.$$

If not, then for each $\delta = 1 - \frac{1}{n}$ and $K = n$, there exists a stationary policy $v_n(\cdot)$ such that for some $i = i_n$,

$$P(\tau > n | X_0 = i_n) \geq 1 - \frac{1}{n}$$

under $v_n(\cdot)$.

Passing to the limit along a suitable subsequence, we can show the existence of a stationary policy $v(\cdot)$ under which

$$P(\tau > n | X_0 = j) = 1$$

for some j and all $n \geq 1$.

That is, there would exist a stationary policy such that for some j, $\tau = \infty$ with probability 1, a contradiction. Thus for $n \geq 0$,

$$\begin{aligned} P(\tau > (n+1)K) &= P(\tau \geq (n+1)K, \tau > nK) \\ &= E[P_{X_{Kn}}(\tau > K) I\{\tau > nK\}] \\ &\leq \delta P(\tau > nK). \end{aligned}$$

Iterating, we have

$$P(\tau > nK) \leq \delta^n.$$

An important consequence of this is that

$$E[\tau] = \sum_{t=0}^{\infty} P(\tau \geq t)$$

$$
\begin{aligned}
&= \sum_{t=0}^{\infty} \sum_{m=Kt}^{K(t+1)-1} P(\tau \geq m) \\
&\leq \sum_{t=0}^{\infty} \sum_{m=Kt}^{K(t+1)-1} P(\tau \geq Kt) \\
&= K \sum_{t=0}^{\infty} P(\tau \geq Kt) \\
&\leq K \sum_{t=0}^{\infty} \delta^t \\
&= \frac{K}{1-\delta} \quad < \quad \infty.
\end{aligned}
$$

Let $k : S_0 \times U \mapsto [0, \infty)$ and $h : A \mapsto \mathcal{R}$ be prescribed running cost and terminal cost functions. The objective is to minimize

$$E\left[\sum_{t=0}^{\tau-1} k(X_t, Z_t) + h(X_\tau)\right].$$

This is finite under any stationary policy. Just as we noted for the infinite horizon discounted cost problem, the cost to go here will depend only on the current state and not on the clock time t. Therefore we define the value function as

$$V(i) := \min E\left[\sum_{t=0}^{\tau-1} k(X_t, Z_t) + h(X_\tau) | X_0 = i\right].$$

Here the minimum is over all admissible controls, finite because it is so for stationary policies.

By the dynamic programming principle, we write the dynamic programming equation by inspection as

$$V(i) = \min_u [k(i,u) + \sum_j p(j|i,u)V(j)], \quad i \in S_0, \text{ (13)}$$

$$V(i) = h(i), \quad i \in A. \text{ (14)}$$

To prove this formally, let V solve (13)-(14). Let $X_0 = i$, $\epsilon > 0$ and consider $Z_0 = u$ and $Z_t, t \geq 1$, that is ϵ-optimal for initial condition X_1. Then

$$\begin{aligned} V(i) &\leq E[k(i,u)I\{\tau > 0\} + h(i)I\{\tau = 0\} \\ &\quad + V(X_1)I\{\tau > 0\} + \epsilon] \\ &= E[k(i,u)I\{i \in S_0\} + h(i)I\{i \in A\} \\ &\quad + V(X_1)I\{i \in S_0\} + \epsilon]. \end{aligned}$$

This leads to $V(i) \leq k(i,u) + \Sigma_{j \in S}\, p(j|i,u)V(j) + \epsilon, i \in S_0$, with $V(i) = h(i), i \in A$.

Letting $\epsilon \downarrow 0$,

$$V(i) \leq \min_u [k(i,u) + \sum_j p(j|i,u)V(j)], \ i \in S_0, \quad (15)$$

$$V(i) = h(i), \ i \in A. \quad (16)$$

The DP equation using the DP principle is

$$V(i) = \min_u [k(i,u) + \sum_j p(j|i,u)V(j)], \ i \in S_0, \quad (17)$$

$$V(i) = h(i), \ i \in A. \quad (18)$$

If the inequality in (17) is strict for some i, there exists an $\eta > 0$ such that

$$V(i) + \eta \leq \min_u [(k(i,u) + \sum_j p(j|i,u)V(j))I\{i \in S_0\} + h(i)I\{i \in A\}].$$

Iterating (17)-(18), we get,

$$\begin{aligned} V(i) + \eta &\leq E[\sum_{t=0}^{T} k(X_t, Z_t)I\{\tau > t\} + h(X_\tau)I\{\tau \leq T\}] \\ &\quad + E[V(X_1)I\{\tau > T\}] + \epsilon \\ &\leq E[\sum_{t=0}^{T} k(X_t, Z_t)I\{\tau > t\} + h(X_\tau)I\{\tau \leq T\}] \\ &\quad + \max_j |V(j)|P(\tau > T) + \epsilon \end{aligned}$$

$$\begin{aligned} & \overset{T\uparrow\infty}{\to} \quad E[\sum_{t=0}^{\infty} k(X_t, Z_t) I\{\tau > t\} + h(X_\tau) I\{\tau \leq \infty\}] + \epsilon \\ & = \quad E\left[\sum_{t=0}^{\tau-1} k(X_t, Z_t) + h(X_\tau)\right] + \epsilon, \end{aligned}$$

implying, letting $\epsilon \downarrow 0$,

$$V(i) + \eta \leq \min E\left[\sum_{t=0}^{\tau-1} k(X_t, Z_t) + h(X_\tau) | X_0 = i\right].$$

This contradicts the definition of $V(i)$, so equality must hold in (17)-(18). This proves the DP equation.

Let $v(i)$ minimize the right hand side of (13). Then under the stationary policy, an iterative argument analogous to the above leads to

$$V(i) = E\left[\sum_{t=0}^{\tau-1} k(X_t, v(X_t)) + h(X_\tau)|X_0 = i\right],$$

implying that v is optimal. The converse, viz., v is optimal for all initial conditions only if it minimizes the right hand side of (13), is proved similarly (check this).

Any solution of the DP equation must have this representation, hence is unique.

The policy iteration algorithm can be written down for this problem and justified along the same lines as before.

The value iteration algorithm becomes: beginning with an initial guess $V_0(\cdot)$ with $V_0(i) = h(i)$ for $i \in A$,

$$\begin{aligned} V_{n+1}(i) &= \min_u [k(i,u) + \sum_j p(j|i,u) V_n(j)], \quad i \in S_0, \\ V_n(i) &\equiv h(i) \qquad \forall\ i \in A, n \geq 0. \end{aligned}$$

Since $V_n(i)$ is frozen at $h(i)$ for $i \in A$, we are effectively iterating $\mathcal{V} := V$ restricted to S_0 as

$$\begin{aligned} \mathcal{V}_{n+1}(i) = \min_u [k(i,u) &+ \sum_{j \in S_0} p(j|i,u) \mathcal{V}_n(j) \\ &+ \sum_{j \in A} p(j|i,u) h(j)], \quad i \in S_0. \end{aligned} \tag{19}$$

Iterating and using (13)-(14), we get, for $n \geq 1$,

$$\begin{aligned}
\mathcal{V}_n(i) &= \min_u E[k(X_0,u)I\{\tau>0\} + h(X_0)I\{\tau=0\} \\
&\quad + \mathcal{V}_{n-1}(X_1)I\{\tau>1\} + h(X_1)I\{\tau=1\} \mid X_0 = i] \\
&= \min E[\sum_{t=0}^{n-1}(k(X_t,Z_t)I\{\tau>t\} + h(X_t)I\{\tau=t\}) \\
&\quad + \mathcal{V}_0(X_n)I\{\tau>n\} + h(X_n)I\{\tau=n\}|X_0=i].
\end{aligned}$$

Here and later, the 'min' is over all admissible controls.

Thus

$$\begin{aligned}
&|V_n(i) - \min E[\sum_{t=0}^{n-1}(k(X_t,Z_t)I\{\tau>t\} \\
&\quad + h(i)I\{\tau=t\})|X_0=i]| \\
&\leq \max_{i\in S_0}|V_0(i)|P(\tau>n) + \max_i |h(i)|P(\tau=n) \overset{n\uparrow\infty}{\to} 0.
\end{aligned}$$

We have

$$\min E[\sum_{t=0}^{n-1}(k(X_t,Z_t)I\{\tau>t\}+h(X_t)I\{\tau=t\})|X_0=i]$$

$$= \min E[\sum_{t=0}^{\tau\wedge(n-1)} k(X_t,Z_t)+h(X_{\tau\wedge n})|X_0=i],$$

and

$$|\min E[\sum_{t=0}^{\tau\wedge(n-1)} k(X_t,Z_t)+h(X_{\tau\wedge n})|X_0=i]$$

$$-\min E[\sum_{t=0}^{\tau} k(X_t,Z_t)+h(X_\tau)|X_0=i]|$$

$$\leq \max E[|\sum_{t=0}^{\tau} k(X_t,Z_t)+h(X_\tau)-$$

$$\sum_{t=0}^{\tau\wedge(n-1)} k(X_t,Z_t)+h(X_{\tau\wedge n})||X_0=i]$$

$$= \max E[|\sum_{t=\tau\wedge n+1}^{\tau} k(X_t, Z_t) + h(X_\tau) - h(X_{\tau\wedge n})||X_0 = i]$$
$$\to 0 \text{ as } n \to \infty.$$

Hence

$$\min E[\sum_{t=0}^{\tau\wedge(n-1)} k(X_t, Z_t) + h(X_{\tau\wedge n})|X_0 = i] \overset{n\uparrow\infty}{\to}$$
$$\min E[\sum_{t=0}^{\tau} k(X_t, Z_t) + h(X_\tau)|X_0 = i].$$

This proves that the value iteration converges to the value function as desired.

Exercises:

1. Write down the dynamic programming equation for the finite horizon cost

$$E\left[\sum_{m=0}^{T}\left(\prod_{k=0}^{m-1} c(X_k, Z_k)\right)k(X_m, Z_m)\right]$$

for prescribed $T > 0$ and $c, k : S \mapsto \mathcal{R}$.

2. Write down the LP formulation of the shortest path problem.

Average cost

Consider again a controlled Markov chain (X_n, Z_n) as before. The objective now is to minimize the long run average cost (or simply the 'average cost') defined by

$$\limsup_{N\uparrow\infty} \frac{1}{N} E\left[\sum_{m=0}^{N-1} k(X_n, Z_n)\right]. \tag{20}$$

We shall assume that under any stationary policy $Z_n = v(X_n)$ $\forall n$, the chain is irreducible. Then it has a unique stationary distribution π_v and (20) is in fact a limit a.s., which equals $\sum_i \pi_v(i)k(i, v(i))$ with probability one by the strong law of large numbers for Markov chains.

It turns out that an optimal stationary policy exists that is optimal among all policies, so we can confine our attention to stationary policies alone.

One problem is that the cost involves an asymptotic arithmetic mean and therefore no finite segment matters. In particular the 'one step analysis' that facilitated dynamic programming so far does not work.

The classical approach is to consider this problem as a limiting case of the infinite horizon discounted cost problem as the discount factor approaches 1 (i.e., no discount).

We take a direct approach here which is less intuitive, but is simpler with fewer technicalities. It is based on the so called Poisson equation

$$V(i) = f(i) - \beta + \sum_j p(j|i)V(j), \ i \in S,$$

where $P = [[p(\cdot|\cdot)]]$ is the transition probability matrix for an irreducible Markov chain with stationary distribution π and $f : S \mapsto \mathcal{R}$ is a given function.

Both the vector V and the scalar β are unknowns.

Multiplying the equation on both sides by $\pi(i)$ and summing over i, we get

$$\sum_i \pi(i)V(i) = \sum_i \pi(i)f(i) - \beta + \sum_i \pi(i) \sum_j p(j|i)V(j).$$

Using the fact $\pi(j) = \sum_i \pi(i)p(j|i)\ \forall j$, we get $\beta = \sum_i \pi(i)f(i)$. Thus β is uniquely specified by the equation.

Clearly, adding a fixed constant to each $V(i)$ does not affect the equation, so V is not unique. If V' is another solution, subtracting the equation for V' from that for V and keeping in mind that β is unique, we get

$$(V - V') = P(V - V')$$

Iterating, assuming aperiodicity,

$$V - V' = P^n(V - V') \;\rightarrow \quad \text{a constant vector.}$$

In general, we can use

$$V - V' = \frac{1}{n} \sum_{m=0}^{n-1} P^m(V - V') \;\rightarrow \quad \text{a constant vector.}$$

Thus V is unique up to the addition of a constant and therefore can be rendered unique, e.g., by fixing the value of one component or fixing its minimum, maximum or average value, etc.

Now consider the problem of minimizing the cost only over stationary ramdomized policies. Under a stationary randomized policy $i \in S \mapsto q(\cdot|i) \in \{$the probability vectors on $U\}$, the chain is an irreducible Markov chain with transition probabilities $\sum_u p(j|i,u)q(u|i)$.

The problem reduces to an LP over $\mu(i,u) = \pi_q(i)q(u|i)$, where π_q is the unique stationary distribution under the stationary randomized policy q, as decribed below: There is a one-one correspondence between q and μ satisfying

$$\mu(i,u) \geq 0, \ \sum_{i,u} \mu(i,u) = 1,$$

$$\sum_{u} \mu(j,u) = \sum_{i,u} \mu(i,u)p(j|i,u), \ \forall j \in S.$$

because given any such μ, it can be written as $\mu(i,u) = \pi(i)q(u|i)$ where $\pi(i) = \sum_a \mu(i,a)$ and $q(u|i) = \mu(i,u)/\pi(i)$. Then the above equation becomes

$$\pi(j) = \sum_i \pi(i) \sum_u q(u|i)p(j|i,u) \ \forall j \Longrightarrow \pi = \pi_q.$$

Thus the problem reduces to the linear program:

Minimize $\sum_i \mu(i,u)k(i,u)$ subject to

$$\begin{aligned} \sum_{i,u} \mu(i,u) &= 1, \\ \sum_{u} \mu(j,u) &= \sum_{i,u} \mu(i,u)p(j|i,u), \ j \in S, \\ \mu(i,u) &\geq 0. \end{aligned}$$

This has an optimal solution μ^*. Let q^*, π^* denote the corresponding q, π and V^*, β^* a solution to the corresponding Poisson equation. That is,

$$V^*(i) = \sum_u k(i,u)q^*(u|i) - \beta^* + \sum_{u,j} p(j|i,u)q^*(u|i)V^*(j), \ i \in S. \tag{21}$$

Then $\beta^* = \sum_i \pi^*(i) \sum_u k(i,u) q^*(u|i)$ is the optimal cost (among stationary randomized policies).

Claim: V^*, β^* satisfy the dynamic programming equation for average cost given by

$$V^*(i) = \min_u [k(i,u) - \beta^* + \sum_j p(j|i,u) V^*(j)], \ i \in S. \quad (22)$$

From (21), we have

$$V^*(i) \geq \min_u [k(i,u) - \beta^* + \sum_j p(j|i,u) V^*(j)], \ i \in S.$$

Suppose the inequality is strict for some i.

Then for the stationary policy $v(\cdot)$ such that $v(i)$ attains the minimum on the right hand side of (22), we have

$$V^*(i) \geq k(i, v(i)) - \beta^* + \sum_j p(j|i, v(i))V^*(j), \ \ i \in S,$$

with a strict inequality for at least one i. Multiplying both sides of the equation by $\pi_v(i)$ and summing over i, we get

$$\sum_i \pi_v(i)k(i, v(i)) < \beta^*,$$

which contradicts the fact that β^* is the optimal cost. Thus equality must hold, i.e., (22) holds.

This also shows that a stationary policy v such that $v(i)$ minimizes the right hand side of (22) is optimal among all stationary randomized policies.

Conversely, suppose v is a stationary policy that is optimal among all stationary randomized policies. Then

$$V^*(i) \leq k(i, v(i)) - \beta^* + \sum_j p(j|i, v(i))V^*(j), \ \ i \in S.$$

If the inequality is srict for some i, multiplying both sides by $\pi_v(i)$ and summing over i, we get

$$\sum_i \pi_v(i)k(i, v(i)) > \beta^*.$$

Then v is not optimal among stationary randomized policies, a contradiction. Thus a stationary policy v is optimal among all stationary randomized policies if and only if it attains the minimum on the right hand side of (22).

Furthermore, since this minimum is indeed attained at some $u \in U$ for each i, setting $v(i) =$ this u yields a stationary policy optimal among all stationary randomized policies. That is, there exists a stationary policy optimal among all stationary randomized policies.

Now consider an arbitrary admissible control $\{Z_n\}$. Then from (22),

$$E[V^*(X_m)] \leq E[k(X_m, Z_m)] - \beta^* + E[V^*(X_{m+1})] \ \forall \ m \geq 0.$$

Summing both sides over $0 \leq m < n$, rearranging terms and dividing through by n, we get

$$\frac{1}{n} E\left[\sum_{m=0}^{n-1} k(X_m, Z_m)\right] - \beta^* \geq \frac{E[V^*(X_0) - V^*(X_n)]}{n} \to 0$$

as $n \uparrow \infty$. Thus no admissible control can give a cost lower than β^*, and therefore a stationary policy optimal among all stationary randomized policies is also optimal among all admissible policies.

Suppose V', β' is another solution pair for (22). That is,

$$V'(i) = \min_u [k(i,u) - \beta' + \sum_j p(j|i,u)V'(j)], \ i \in S.$$

Argue as for (V^*, β^*) to conclude that $\beta' = \beta^*$. The fact that a stationary policy v is optimal if and only if $v(i)$ minimizes the right hand side of (22) also follows as before. Then taking v to be a particular optimal stationary policy, we have

$$\begin{aligned} V^*(i) &= k(i,v(i)) - \beta^* + \sum_j p(j|i,v(i))V^*(j), \\ V'(i) &= k(i,v(i)) - \beta^* + \sum_j p(j|i,v(i))V'(j), \quad i \in S. \end{aligned}$$

Subtracting the second equation from the first, we get

$$V^*(i) - V'(i) = \sum_j p(j|i, v(i))(V^*(j) - V'(j)),$$

which by familiar arguments, implies that $V^*(\cdot) - V'(\cdot) \equiv$ a constant. Thus V^* is unique up to an additive constant.

That this cannot be improved upon follows as for the Poisson equation: (22) does not change if you add a constant to all components of V^*.

The policy iteration algorithm can be written down and its convergence in finitely many steps can be established the usual way. Thus one starts with an initial stationary policy $v_0(\cdot)$ and at step n, does:

1. Solve the Poisson equation

$$V_n(i) = k(i, v_n(i)) - \beta_n + \sum_j p(j|i, v_n(i))V_n(j), \ \ i \in S.$$

2. Pick $v_{n+1}(\cdot)$ such that $\forall\ i \in S$, $v_{n+1}(i)$ minimizes

$$k(i, \cdot) + \sum_j p(j|i, \cdot)V_n(j).$$

The value iteration algorithm is replaced by the so called 'relative value iteration' in absence of any contractivity. This is given by:

$$V_{n+1}(i) = \min_u [k(i,u) + \sum_j p(j|i,u) V_n(j)] - f(V_n), \ n \geq 0,$$

where $f(V_n)$ is the 'offset' that prevents the scheme from becoming numerically unstable. Choices of $f(V)$ are: $f(V) = V(i_0)$ for a fixed state i_0, $f(V) = \frac{1}{|S|} \Sigma_{i \in S} V(i)$, $\min_i V(i)$, $\max_i V(i)$, etc. Convergence proof of this algorithm is very technical and is omitted.

Alternative approaches:

Can show

$$V(i) = \inf_{\{Z_n\}} E_i \left[\sum_{m=0}^{\tau-1} (k(X_m, Z_m) - \beta) \right].$$

One can use RHS as a definition of V and shown that it satisfies the DP equation by 'one step analysis' as in the case of stochastic shortest path problem. The rest goes as before.

The classical approach is the vanishing discount argument, where one uses the dynamic programming equation for the infinite horizon discunted cost with discount factor $\alpha \in (0,1)$ and define the corresponding value function $V_\alpha(\cdot)$.

Define

$$\overline{V}_\alpha(i) = V_\alpha(i) - V_\alpha(i_0)$$

for some $i_0 \in \mathcal{S}$. Suppose the corresponding chain is aperiodic. Let $v_\alpha^*(\cdot)$ denote an optimal stationary policy for this discounted problem.

Let $\{X_n\}, \{X'_n\}$ be independent Markov chains initiated at i, i_0, resp., controlled by $v^*(\cdot)$ and, glued together when they first meet, i.e, at $\tau := \min\{n \geq 0 : X_n = X'_n\}$ denote this coupling time. Assume aperiodicity. Then for some $K > 0$,

$$\begin{aligned}
&\overline{V}_\alpha(i) \\
&= V_\alpha(i) - V_\alpha(i_0) \\
&= E\left[\sum_{m=0}^{\infty} \alpha^m (k(X_m, v^*_\alpha(X_m)) - k(X'_m, v^*_\alpha(X'_m))\right] \\
&= E\left[\sum_{m=0}^{\tau-1} \alpha^m (k(X_m, v^*_\alpha(X_m)) - k(X'_m, v^*_\alpha(X'_m))\right] \\
&\leq KE[\tau] < \infty.
\end{aligned}$$

Thus $\overline{V}_\alpha(\cdot)$ converges to some $V(\cdot)$ along a subsequence as $\alpha \uparrow 1$.

Since $(1-\alpha)V_\alpha(i_0)$ clearly remains bounded by $\max_{i,u} |k(i,u)|$ as $\alpha \uparrow 1$, we can take a further subsequence along which $(1-\alpha)V_\alpha(i_0) \to \beta'$ (say). Recall the DP equation for discounted cost :

$$V_\alpha(i) = \min_u \left[k(i,u) + \alpha \sum_j p(j|i,)V_\alpha(j) \right].$$

Then $\overline{V}_\alpha(\cdot)$ satisfies

$$\overline{V}\alpha(i) = \min_u \left[k(i,u) + \alpha \sum_j p(j|i,u)\overline{V}(j) \right] - (1-\alpha)V_\alpha(i_0).$$

Letting $\alpha \uparrow 1$ along the above subsequence, we have

$$V(i) = \min_u \left[k(i,u) - \beta' + \sum_j p(j|i,u)V(j) \right].$$

which is the DP equation for average cost.

The proof can be easily adapted if $\{X_n\}$ is periodic.

Multichain problems: Drop irreducibility. Let β_j denote the minimum cost starting from j. Then we have the following DP:

$$\begin{aligned} \beta_i &= \min_u p(j|i,u)\beta_j \ \forall i,u, \\ V(i) &= \min_{u \in B_i} \left[k(i,u) - \beta_i + \sum_j p(j|i,u)V(j) \right] \ \forall i,u, \end{aligned}$$

where $B_i :=$ the set of minimizers in the first equation. The primal LP is: Maximize $\sum_i \beta_i$ subject to

$$\begin{aligned} \beta_i &\leq \sum_j p(j|i,u)\beta_j \ \forall i,u, \\ V(i) &\leq k(i,u) - \beta_i + \sum_j p(j|i,u)V(j) \ \forall i,u. \end{aligned}$$

The dual program is: Minimize $\sum_{i,u} \mu(i,u)k(i,u)$ subject to

$$\begin{aligned}
\sum_u \mu(i,u) &= \sum_{j,u} p(i|j,u)\mu(j,u) \ \forall i, \\
\sum_u \mu(i,u) + \sum_u \nu(i,u) &= \sum_{j,u} p(i|j,u)\nu(j,u) + 1 \ \forall i, \\
\mu(i,u), \nu(i,u) &\geq 0 \ \forall i,u, \ \sum_{i,u} \mu(i,u) = 1.
\end{aligned}$$

Risk-sensitive control

Risk-sensitive control seeks to minimize or maximize the exponential growth rate of a multiplicative cost/reward. We shall consider the cost minimization problem, where the objective is to minimize

$$\limsup_{n\uparrow\infty} \frac{1}{n} \log E\left[e^{\sum_{m=0}^{n-1} c(X_m, Z_m)}\right]. \tag{23}$$

We shall assume as before that the chain is irreducible under any stationary policy and in addition, that it is aperiodic.

For a stationary policy v, consider the matrix Q_v whose (i,j)th element is $e^{c(i,v(i))}p(j|i,v(i))$. This is a non-negative irreducible matrix and therefore by the Perron-Frobenius theorem, has a unique eigenvalue-eigenvector pair (λ_v, V_v) with the properties: $\lambda_v > 0$, $V_v > 0$ componentwise, and for any other eigenvalue λ' of Q_v, $|\lambda'| < \lambda_v$ (because of irreducibility and aperiodicity).

We first prove that $\log \lambda_v$ is precisely the risk-sensitive cost associated with the policy v. Consider the eigenvalue equation $Q_v V_v = \lambda_v V_v$, i.e.,

$$\lambda_v V_v(i) = \sum_j e^{c(i,v(i))} p(j|i,v(i)) V_v(j), \ \ i \in S. \qquad (24)$$

Defining

$$\tilde{p}(j|i) := \frac{e^{c(i,v(i))}p(j|i,v(i))V_v(j)}{\lambda_v V_v(i)}, \ i \in S,$$

we have $\tilde{p}(\cdot|\cdot) \geq 0$ and $\sum_j \tilde{p}(j|i) = 1 \ \forall i$. Thus $\tilde{p}(\cdot|\cdot)$ is a legitimate transition probability function. It will be irreducible aperiodic because $p(\cdot|\cdot, v(\cdot))$ was so. Thus it has a unique stationary distribution $\tilde{\pi}$ such that if $\{\widetilde{X}_n\}$ is a Markov chain with transition probability function $\tilde{p}$, then by aperiodicity

$$E[f(\widetilde{X}_n)] \to \sum_i \tilde{\pi}(i) f(i)$$

for all $f : S \mapsto \mathcal{R}$, regardless of the initial distribution.

Let $\{X_n\}$ be controlled by the stationary policy v with $X_0 = i_0$ (say). Then

$$\begin{aligned}
&\frac{1}{n}\log E\left[e^{\sum_{m=0}^{n-1} c(X_m, v(X_m))} | X_0 = i_0\right] \\
&= \frac{1}{n}\log\left[\sum_{\{i_1,\cdots,i_n\}} \prod_{m=0}^{n-1} e^{c(i_m,v(i_m))} p(i_{m+1}|i_m, v(i_m))\right] \\
&= \frac{1}{n}\log\Big[\sum_{\{i_k\}} \prod_{m=0}^{n-1} \left(\frac{e^{c(i_m,v(i_m))} p(i_{m+1}|i_m, v(i_m)) V_v(i_{m+1})}{\lambda_v V_v(i_m)}\right) \\
&\qquad\qquad \times \lambda_v^n \left(\frac{V_v(i_0)}{V_v(i_n)}\right)\Big] \\
&= \log\lambda_v + \frac{1}{n}\left(\log V_v(i_0) + \log E_{i_0}\left[\frac{1}{V_v(\widetilde{X}_n)}\right]\right) \\
&\overset{n\uparrow\infty}{\to} \log\lambda_v.
\end{aligned}$$

Thus $\log \lambda_v$ is the cost associated with the stationary policy v. Since there are finitely many stationary policies, there exists a stationary policy v^* such that $\lambda_{v^*} = \min_v \lambda_v$. Let $V := V_{v^*}, \lambda := \lambda_{v^*}$. Then

$$\lambda V(i) = e^{c(i,v^*(i))} \sum_j p(j|i, v^*(i)) V(j), \ \ i \in S. \tag{25}$$

We claim that

$$\lambda V(i) = \min_u \left(e^{c(i,u)} \sum_j p(j|i,u) V(j) \right), \ \ i \in S. \tag{26}$$

The proof is similar to that for the additive costs.

If not, there exists a stationary policy v (obtained as the argmin of the right hand side) such that

$$\lambda V(i) \geq e^{c(i,v(i))} \sum_j p(j|i,v(i))V(j), i \in S, \qquad (27)$$

with a strict inequality for at least one i, say $i = i_0$. Then for some $\delta > 0$,

$$\lambda V(i_0) \geq e^{c(i_0,v(i_0))} \sum_j p(j|i_0,v(i_0))V(j) + \delta.$$

On the other hand,

$$\lambda_v V_v(i) = e^{c(i,v(i))} \sum_j p(j|i,v(i))V_v(j), i \in S. \qquad (28)$$

Divide the LHS, resp., the RHS of (27) by the LHS, resp., the RHS of (28).

This leads to:

$$\begin{aligned}\frac{\lambda V(i)}{\lambda_v V_v(i)} &\geq \frac{\sum_j p(j|i,v(i))V_v(j)\left(\frac{V(j)}{V_v(j)}\right)}{\sum_j p(j|i,v(i))V_v(j)} \\ &= \sum_j \bar{p}(j|i)\left(\frac{V(j)}{V_v(j)}\right),\end{aligned}$$

for

$$\bar{p}(j|i) := \frac{p(j|i,v(i))V_v(j)}{\sum_k p(k|i,v(i))V_v(k)}.$$

This too is an irreducible transition probability with a unique stationary distribution (say) $\bar{\pi}$.

Furthermore,

$$\frac{\lambda V(i_0)}{\lambda_v V_v(i_0)} \geq \frac{\sum_j p(j|i_0, v(i_0)) V_v(j) \left(\frac{V(j)}{V_v(j)}\right)}{\sum_j p(j|i_0, v(i_0)) V_v(j)} + \frac{\delta}{\lambda_v V_v(i_0)}$$

$$\Longrightarrow \qquad \text{(since } \lambda \leq \lambda_v\text{)}$$

$$\begin{aligned}
\frac{\lambda V(i)}{\lambda_v V_v(i)} &\geq \frac{\sum_j p(j|i, v(i)) V_v(j) \left(\frac{V(j)}{V_v(j)}\right)}{\sum_j p(j|i, v(i)) V_v(j)} + \frac{\delta}{\lambda_v V_v(i_0)} I\{i = i_0\} \\
&= \sum_j \bar{p}(j|i) \left(\frac{V(j)}{V_v(j)}\right) + \frac{\delta}{\lambda_v V_v(i_0)} I\{i = i_0\} \\
&\geq \sum_j \bar{p}(j|i) \left(\frac{\lambda V(j)}{\lambda_v V_v(j)}\right) + \frac{\delta}{\lambda_v V_v(i_0)} I\{i = i_0\}
\end{aligned}$$

Then iterating the inequality, we have, for $C := \frac{\delta}{\lambda_v V_v(i_0)}$,

$$\frac{\lambda V(i)}{\lambda_v V_v(i)} \geq E\left[\frac{\lambda V(\bar{X}_n)}{\lambda_v V_v(\bar{X}_n)} + C I\{\bar{X}_n = i_0\}\right]$$

$$\rightarrow \sum_j \bar{\pi}(j)\left(\frac{\lambda V(j)}{\lambda_v V_v(j)}\right) + C\bar{\pi}(i_0).$$

Taking $i :=$ the minimizer of $\frac{V(\cdot)}{V_v(\cdot)}$, we get a contradiction. So (26), the 'DP equation for risk-sensitive control', holds. This is an equation in unknowns $(V(\cdot), \lambda)$.

Take any other solution pair (λ', V'), i.e.,

$$\lambda' V'(i) = \min_u \left(e^{c(i,u)} \sum_j p(j|i,u) V'(j) \right), \ i \in S.$$

Let v' be the stationary policy such that $v'(i)$ minimizes the right hand side. Then $\lambda' = \lambda_{v'}$. Hence $\lambda \leq \lambda'$ and therefore

$$\lambda V'(i) \leq \lambda_v V'(i) \leq e^{c(i,v^*(i))} \sum_j p(j|i, v^*(i)) V'(j).$$

Hence

$$\frac{V(i)}{V'(i)} \leq \frac{\sum_j p(j|i, v^*(i)) V'(j) \left(\frac{V(j)}{V'(j)} \right)}{\sum_j p(j|i, v^*(i)) V'(j)}.$$

Defining

$$\breve{p}(\cdot|\cdot) = \frac{p(j|i, v^*(i))V'(j)}{\sum_j p(j|i, v^*(i))V'(j)},$$

we get

$$\frac{V(i)}{V'(i)} \geq \sum_j \breve{p}(j|i) \left(\frac{V(j)}{V'(j)} \right).$$

Now picking i to be the minimizer on the left leads to a contradiction unless $\frac{V(\cdot)}{V'(\cdot)}$ is a constant. Hence V, V' differ at most by a multiplicative constant. It follows that $\lambda = \lambda'$.

Claim: A stationary policy v is optimal if and only if $v(i)$ achieves the minimum on the right hand side of (26).

The 'if' case is easy, because if so, (25) holds with v in place of v^* and the uniqueness of the principal eigenvalue guaranteed by the Perron-Frobenius theorem implies that $\lambda_v = \lambda$.

The 'only if' part follows by a familiar argument using contradiction. If not, we have

$$\lambda V(i) \;\leq\; e^{c(i,v(i))} \sum_j p(j|i,v(i)) V(j) \;\; \forall i,$$

with the inequality strict for at least one i, which then leads to $\lambda_v > \lambda$, a contradiction. Finally, under a general admissible $\{Z_n\}$ with $X_0 = i_0$ (say), we have

$$\begin{aligned} V(X_n) \;&\leq\; \lambda^{-1} e^{c(X_n,Z_n)} E\left[V(X_{n+1})|X_n, Z_n\right] \\ &=\; \lambda^{-1} e^{c(X_n,Z_n)} E\left[V(X_{n+1})|X_m, Z_m, m \leq n\right] \forall n, \end{aligned}$$

where the equality follows from the controlled Markov property.

Iterating and taking expectation, we have

$$\lambda^n V(i_0) \;\leq\; E\left[e^{\sum_{m=0}^{n-1} c(X_m,Z_m)} V(X_n)\right]$$

Hence, since $V(X_{n+1}) \leq \max_i |V(i)|$, we have

$$\log\lambda \;+\; \frac{1}{n}\log V(i_0) \leq \frac{1}{n}\log E\left[e^{\sum_{m=0}^{n-1} c(X_m,Z_m)}\right]$$

$$+\;\frac{1}{n}\log(\max_i |V(i)|).$$

Letting $n \uparrow \infty$,

$$\log\lambda \leq \liminf_{n\uparrow\infty} \frac{1}{n}\log E\left[e^{\sum_{m=0}^{n-1} c(X_m,Z_m)}\right].$$

That is, no admissible policy can do better than the optimal stationary policy.

Policy iteration for risk-sensitive control starts with a stationary policy v_0 and at step n,

a) solves the eigenvalue problem

$$\lambda_n V_n(i) = e^{c(i,v_n(i))} \sum_j p(j|i, v_n(i)) V_n(j), \ \ i \in S,$$

for $(V_n(\cdot), \lambda_n)$, both > 0, and,

b) picks $v_{n+1}(i) \in$ Argmin $\left(e^{c(i,\cdot)} \sum_j p(j|i,\cdot) V_n(j) \right) \ \ \forall i.$

Convergence in finitely many steps can be proved as before.

The 'relative value iteration' algorithm starts with an initial guess V_0 and does, for a fixed $i_0 \in S$,

$$\begin{aligned} \widetilde{V}_{n+1}(i) &= \min_u \left(e^{c(i,u)} \sum_j p(j|i,u) V_n(j) \right), \\ V_{n+1}(i) &= \frac{\widetilde{V}_{n+1}(i)}{\widetilde{V}_{n+1}(i_0)}, \ n \geq 0. \end{aligned}$$

Then $V_n \to V$ for the V such that $V(i_0) = \lambda$. (Recall that V is unique only up to a multiplicative scalar. This condition renders the limit unique.)

This is a nonlinear analog of the 'power iteration'.

Some general facts about risk-sensitive control:

Looking at the Taylor series for the exponential function, one sees that the risk-sensitive cost captures all moments of the cumulative cost $\sum_{m=0}^{n} c(X_m, Z_m)$, and hence can be viewed as an extension of the 'mean-variance' criterion. Unlike the latter, it is amenable to dynamic programming, hence has gained some popularity in financial problems. It also has relations to robust control.

One often puts an additional parameter κ in the exponent, i.e., the cost is

$$\limsup_{n\uparrow\infty} \frac{1}{n} \log E\left[e^{\kappa \sum_{m=0}^{n-1} c(X_m, Z_m)}\right].$$

Assuming $c(\cdot,\cdot) > 0$, the case $\kappa > 0$ corresponds to the more common risk-averse behavior, whereas $\kappa < 0$ corresponds to risk-seeking behavior. Dividing the above expression by κ and letting $\kappa \to 0$, one formally recovers the average cost as a limiting case, which therefore is called the risk-neutral case.

Exponentiation also arises naturally in many cases due to compounding effects.

Also note that in classical criteria, cost minimization is equivalent to reward maximization if you set running reward equal to negative of the running cost. In risk-sensitive control, this is not so, you get a different problem.

Optimal stopping:

If you stop the process at the stopping time n, you pay stopping cost $h(X_n)$. Otherwise you pay cost $k(X_n)$ and continue.

Thus the cost is

$$E\left[\sum_{m=0}^{\tau-1} k(X_m) + h(X_\tau)\right]$$

and the objective is to minimize this over stopping times τ. The value function $V(i)$ is the minimum of the above over all τ when $X_0 = i, i \in S$.

The dynamic programming equation then is

$$V(i) = \min\left(k(i) + \sum_j p(j|i)V(j), h(i)\right)$$

$$= \min_{u\in\{0,1\}}\left(u(k(i) + \sum_j p(j|i)V(j)) + (1-u)h(i)\right)$$

and the optimal stopping time is

$$\tau := \min\{n \geq 0 : k(X_n) + \sum_j p(j|X_n)V(X_n) \geq h(X_n)\},$$

i.e., the first exit time from the set

$\{i \in S : k(i) + \sum_j p(j|i)V(i) < h(i)\}$.

This can be written as

$$V(i) \leq k(i) + \sum_j p(j|i)V(j),$$

$$V(i) \leq h(i),$$

$$(V(i) - k(i) - \sum_j p(j|i)V(j))(V(i) - h(i)) = 0.$$

This is called a system of 'variational inequalities'. Mathematically, this is an 'obstacle problem'. LP formulation is also possible.

A 'mixed problem' with additional classical control leads to

$$V(i) = \min\left(k(i,u) + \sum_j p(j|i,u)V(j), h(i)\right).$$

Equivalently, we have the 'quasi-variational inequalities'

$$V(i) \leq \min_u(k(i,u) + \sum_j p(j|i,u)V(j)),$$

$$V(i) \leq h(i),$$

$$(V(i) - \min_u(k(i,u) + \sum_j p(j|i,u)V(j)))(V(i) - h(i)) = 0.$$

Impulse control:

Here one can reset the trajectory from i to j with cost $c(i,j)$. Consider the discounted cost with discount factor $\alpha \in (0,1)$ given by

$$E\left[\sum_{m=0}^{\infty} \alpha^m k(X_m) + \sum_{m=0}^{\infty} \alpha^{\tau_m} c(X_{\tau_m}^-, X_{\tau_m}^+)\right].$$

The optimization is over the stopping times $\{\tau_n\}$.

One assumes $c(i,i) = 0$ and

$$c(i,k) < c(i,j) + c(j,k) \ \forall \ i,j,k.$$

This avoids some pathologies. The dynamic programming equation is

$$V(i) = \min \left(k(i) + \alpha \sum_j p(j|i) V(j), \min_j (c(i,j) + V(j)) \right).$$

The optimal decision is to continue if the first term in the outer parentheses is $\leq$ the second and reset the state to argmin$(c(i,\cdot) + V(\cdot))$ otherwise. Extensions to mixed problems and other cost criteria are possible.

Variational inequalities:

$$
\begin{aligned}
V(i) &\leq k(i) + \alpha \sum_j p(j|i) V(j), \\
V(i) &\leq \min_j (c(i,j) + V(j)), \\
0 &= (V(i) - (k(i) + \alpha \sum_j p(j|i) V(j)) \times \\
&\quad (V(i) - \min_j (c(i,j) + V(j))).
\end{aligned}
$$

Quasi-variational inequalities for mixed problems can be written analogously.

Switching control:

Switching cost $c(u, u')$ associated with changing control from u to u' satisfies

$$c(u, u') < c(u, u'') + c(u'', u') \ \forall \ u, u', u''.$$

The dynamic programming equation is

$$V(i, u) \ = \ \min \left(k(i, u) + \alpha \sum_j p(j|i, u) V(j, u), \right.$$
$$\left. \min_v (c(u, v) + V(i, v)) \right).$$

The optimal decision is to continue with the current control u if the first term in the outer parentheses is $\leq$ the second, otherwise switch to $v :=$ the minimizer of the second term. Extensions to mixed problems and other cost criteria are possible. Variational/quasi-variational inequality formulations are also possible.

Partially observed MDPs

Partially observed Markov Decision Processes (POMDPs) have, in adition to (X_n, Z_n), an observation process $\{Y_n\}$ taking values in a finite observation space H, with an associated transition kernel

$$(i, u, j, y) \in S \times U \times S \times H \mapsto p(j, y|i, u) \in [0, 1]$$

with $\sum_{j,y} p(j, y|i, u) = 1 \ \forall \ i, u$, and for $n \geq 0$,

$$P(X_{n+1} = j, Y_{n+1} = y|X_m.Y_m, Z_m, m \leq n) = p(j, y|X_n, Z_n).$$

Examples: 1. $X_n = (X_n^1, X_n^2), Y_n = X_n^2$.
2. $p(j, y|i, u) = p(j|i, u)q(y|j)$. ($q(\cdot|\cdot) \approx$ 'communication channel').

The idea is that only $\{Y_n\}$ is observed, $\{X_n\}$ is not, and the control Z_n at time n therefore can depend only on past observations and controls $Y_m, m \leq n; Z_m, m < n$, and possibly some independent extraneous randomization, but not on $\{X_m\}$.

A natural thing to do is to consider

$$\pi_n(i) := P(X_n = i | Y_m, m \leq n; Z_m, m < n)$$

for $i \in S, n \geq 0$. Then $\pi_n := [\pi_n(1), \cdots, \pi_n(s)]$ is a random probability vector that is the conditional distribution of X_n given past observations $Y_m, m \leq n$, and controls $Z_m, m < n$.

$\{\pi_n\}$ (written as a row vector) is given recursively by

$$\pi_{n+1} = \frac{\pi_n P(Z_n, y_{n+1})}{\pi_n P(Z_n, y_{n+1})\mathbf{1}}, \ n \geq 0, \tag{29}$$

where π_n is written as a row vector and,

- y_{n+1} is the observed value of Y_{n+1},
- $P(u, y)$ for $u \in U, y \in H$, is a substochastic matrix whose (i, j)th element is $p(j, y|i, u)$,
- $\mathbf{1} := [1, 1, \cdots, 1]^T \in \mathcal{R}^s$,

- $\pi_0 :=$ the distribution of X_0.

Equivalently,

$$\pi_{n+1}(i) = \frac{\sum_j \pi_n(j) p(i, y_{n+1}|j, Z_n)}{\sum_{j,k} \pi_n(j) p(k, y_{n+1}|j, Z_n)}.$$

Numerator of the RHS

$$\begin{aligned}
&= \sum_i P(X_n = j | Y_m, m \le n; Z_m, m < n) \times \\
&\qquad P(X_{n+1} = i, Y_{n+1} = y_{n+1} | X_n = j, Z_n) \\
&= \sum_i P(X_n = j | Y_m, m \le n; Z_m, m < n) \times \\
&\quad P(X_{n+1} = i, Y_{n+1} = y_{n+1} | X_n = j, Z_m, m \le n; Y_m, m \le n) \\
&= P(X_{n+1} = i, Y_{n+1} = y_{n+1} | Z_m, m \le n; Y_m, m \le n).
\end{aligned}$$

Here we use the conditional independence of X_n and Z_n given $Y_m, m \leq n; Z_m, m < n$. Similarly, the denominator of RHS

$$\begin{aligned} &= \sum_i P(X_{n+1} = i, Y_{n+1} = y_{n+1} | Z_m, Y_m, m \leq n) \\ &= P(Y_{n+1} = y_{n+1} | Z_m, Y_m, m \leq n). \end{aligned}$$

Thus the ratio is

$$\begin{aligned} &\frac{P(X_{n+1} = i, Y_{n+1} = y_{n+1} | Z_m, Y_m, m \leq n)}{P(Y_{n+1} = y_{n+1} | Z_m, Y_m, m \leq n)} \\ &= P(X_{n+1} = i | Z_m, m \leq n; Y_m, m \leq n+1) \\ &= \pi_{n+1}(i) \ = \ \text{the LHS.} \end{aligned}$$

The formula (29) is an example of a nonlinear filter.

Note also that

$$P(Y_{n+1} = y | Y_m, Z_m, m \leq n) = \sum_{i,j} p(j, y | i, Z_n) \pi_n(i),$$

which depends on past, i.e., $Y_m, Z_m, m \leq n$, only through π_n. Along with (29), this implies that $\{\pi_n\}$ is a controlled Markov chain taking values in $\mathcal{P}(S)$:= the simplex of probability vectors on S.

Consider, e.g., the discounted cost

$$\begin{aligned} & E\left[\sum_{m=0}^{\infty} \alpha^m k(X_m, Z_m)\right] \\ = \; & E\left[\sum_{m=0}^{\infty} \alpha^m E[k(X_m, Z_m)|Y_i, Z_i, i \le m]\right] \\ = \; & E\left[\sum_{m=0}^{\infty} \alpha^m \bar{k}(\pi_m, Z_m)\right] \end{aligned}$$

for

$$\bar{k}(\pi, u) := \sum_i \pi(i) k(i, u).$$

Let $V(\pi)$ denote the value function, i.e., the infimum of the above cost over all admissible controls when $\pi_0 = \pi$.

The corresponding dynamic programming equation becomes

$$V(\pi) = \min_u \left[\bar{k}(\pi, u) + \alpha \sum_{i,j} \pi(i) p(j, y|i, u) V\left(\frac{\pi P(u, y)}{\pi P(u, y)\mathbf{1}} \right) \right]. \tag{30}$$

Dynamic programming equation for finite horizon control can be written analogously.

Standard arguments show that if $v(\pi)$ attains the minimum on the right, then the 'stationary policy' $Z_n = v(\pi_n)$ is optimal.

There are some technicalities involved in ensuring that the v here is a nice (i.e., measurable) function. These can be taken care of easily in the finite set-up here. One can also prove by induction that the Z_n defined thus depends only on $Y_m, m \leq n$, so that it is admissible.

The more complicated criteria such as average and risk-sensitive costs are harder to analyze and have been handled only in special cases, if at all. The key difficulty is the counterpart of 'irreducibility' for this controlled Markov chain, which is generally unavailable.

Constrained MDPs

Example: Minimize

$$\limsup_{N\uparrow\infty} \frac{1}{N} \sum_{m=0}^{N-1} E[k(X_m, Z_m)]$$

subject to

$$\limsup_{N\uparrow\infty} \frac{1}{N} \sum_{m=0}^{N-1} E[c(X_m, Z_m)] \leq C.$$

More than one constraint is also possible.

Using LP formulation, one sees that this simply adds one more constraint to the LP.

On the other hand, using Lagrange multiplier λ, this can be reduced to the unconstrained MDP with running cost

$$k(i,u) + \lambda c(i,u).$$

These can be solved by 'primal-dual' methods.