Induced Stochastic Process and Expectations under a Fixed Policy

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Induced Stochastic Process under a Fixed Policy

In this section we formalize the stochastic process generated by a Markov decision process (MDP) once a policy is fixed. This construction corresponds to Section 2.1.6 of Puterman and provides the probabilistic foundation for policy evaluation.

1.1 Fixing a Policy

An MDP consists of:

- a state process $\{X_t\}_{t\geq 1}$ taking values in a finite state space S,
- an action process $\{Y_t\}_{t\geq 1}$ taking values in an action space $A(X_t)$,
- transition probabilities $p_t(x' \mid x, a)$,
- reward functions $r_t(x, a)$.

Once a policy $\pi = (d_1, d_2, \dots)$ is fixed, decision making is removed. Actions are no longer chosen optimally; instead, they are generated as random variables according to the policy. The MDP therefore induces an uncontrolled stochastic process with a uniquely defined probability measure $\mathbb{P}^\pi$ on the sample space Ω.

1.2 Sample Space and Sample Paths

For a finite horizon of length N, we define the sample space

$$\Omega = (S \times A)^{N-1} \times S.$$

For an infinite-horizon model,

$$\Omega = (S \times A)^{\infty}.$$

A typical element $\omega \in \Omega$ is a *sample path* of the form

$$\omega = (s_1, a_1, s_2, a_2, \dots, a_{N-1}, s_N),$$

or, in the infinite-horizon case,

$$\omega = (s_1, a_1, s_2, a_2, \dots).$$

1.3 Canonical Random Variables

We define the coordinate random variables:

$$X_t(\omega) = s_t, \qquad Y_t(\omega) = a_t,$$

for $t = 1, 2, \dots$.

Thus, X_t denotes the system state at decision epoch t, and Y_t denotes the action taken at that epoch.

1.4 History Process

Define the history process $\{Z_t\}$ by

$$Z_1(\omega) = s_1,$$

and for $t \geq 2$,

$$Z_t(\omega) = (s_1, a_1, s_2, a_2, \ldots, s_{t-1}, a_{t-1}, s_t).$$

The history Z_t contains all information available to the decision maker at time t.

1.5 Initial Distribution

Let $\mathbb{P}_1$ denote the initial distribution of the system state:

$$\mathbb{P}(X_1 = x) = \mathbb{P}_1(x), \qquad x \in S.$$

In most applications, $\mathbb{P}_1$ is degenerate, meaning the initial state is known with probability one.

2 Decision Rules and Conditional Distributions

2.1 History-Dependent Randomized Decision Rules

A randomized history-dependent decision rule at time t is a conditional distribution

$$q_{d_t(z)}(a) := \mathbb{P}^\pi(Y_t = a \mid Z_t = z),$$

defined for all histories z and feasible actions $a \in A(z)$.

2.2 Markov Decision Rules

A randomized Markov decision rule is a special case where actions depend only on the current state:

$$q_{d_t(x)}(a) := \mathbb{P}^\pi(Y_t = a \mid X_t = x).$$

This restriction induces the Markov property in the state-action process.

2.3 State Transition Dynamics

The system dynamics are specified by the transition probabilities

$$p_t(x' \mid x, a) := \mathbb{P}^\pi(X_{t+1} = x' \mid X_t = x, Y_t = a).$$

These probabilities are independent of the policy choice.

3 Dynamic Bayesian Network Representation

Once a policy is fixed, the induced stochastic process admits a Dynamic Bayesian Network (DBN) representation.

3.1 Graphical Structure

For $t = 1, \ldots, N-1$, the DBN contains the edges:

$$X_t \to Y_t, \qquad (X_t, Y_t) \to X_{t+1}.$$

Under Markov decision rules, the following conditional independence relations hold:

$$Y_t \perp (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}) \mid X_t,$$

$$X_{t+1} \perp (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}) \mid (X_t, Y_t).$$

3.2 Joint Distribution Factorization (Finite Horizon)

For $\omega = (x_1, a_1, \ldots, a_{N-1}, x_N)$, the joint distribution induced by a policy π factorizes as

$$\boxed{\mathbb{P}^\pi(\omega) = \mathbb{P}_1(x_1) \prod_{t=1}^{N-1} q_{d_t(z_t)}(a_t) p_t(x_{t+1} \mid x_t, a_t),}$$

where z_t denotes the realized history at time t.

For Markov decision rules, $q_{d_t(z_t)}(a_t)$ reduces to $q_{d_t(x_t)}(a_t)$.

3.3 Deterministic Decision Rules

If the decision rule is deterministic, $a_t = \mu_t(x_t)$, then

$$q_{d_t(x)}(a) = \mathbf{1}\{a = \mu_t(x)\},$$

and the joint distribution simplifies to

$$\mathbb{P}^\mu(x_1, \ldots, x_N) = \mathbb{P}_1(x_1) \prod_{t=1}^{N-1} p_t(x_{t+1} \mid x_t, \mu_t(x_t)).$$

4 Expectations under the Induced Measure

4.1 Random Variables on the Sample Space

Let $W : \Omega \to \mathbb{R}$ be any real-valued random variable defined on the sample space. In particular, for a finite-horizon MDP, the total reward is

$$W(\omega) = \sum_{t=1}^{N-1} r_t(x_t, a_t) + r_N(x_N).$$

4.2 Expectation of the Total Reward (Policy Evaluation)

For a finite-horizon problem, define the total reward random variable

$$W(\omega) = \sum_{t=1}^{N-1} r_t(X_t(\omega), Y_t(\omega)) + r_N(X_N(\omega)), \qquad \omega \in \Omega.$$

By definition, the expectation of W under a fixed policy π is

$$\mathbb{E}^{\pi}[W] = \sum_{\omega \in \Omega} W(\omega)\, \mathbb{P}^{\pi}(\omega).$$

This expression represents the expected total reward obtained by following policy π, computed by averaging the cumulative reward over all possible sample paths weighted by their probabilities. Conceptually, this is the most direct form of *policy evaluation*.

Using the linearity of expectation, the expectation of the total reward can be decomposed into a sum of stage-wise expectations:

$$\mathbb{E}^{\pi}[W] = \mathbb{E}^{\pi}\left[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right] = \sum_{t=1}^{N-1} \mathbb{E}^{\pi}[r_t(X_t, Y_t)] + \mathbb{E}^{\pi}[r_N(X_N)].$$

Each term $\mathbb{E}^{\pi}[r_t(X_t, Y_t)]$ represents the expected reward collected at decision epoch t, where the expectation is taken with respect to the joint distribution of the random variables (X_t, Y_t) induced by the policy π. The terminal term $\mathbb{E}^{\pi}[r_N(X_N)]$ represents the expected terminal reward.

This decomposition shows that policy evaluation amounts to computing expected rewards at each stage under the induced stochastic process, rather than summing over entire sample paths. Subsequent developments exploit this structure using conditional expectations and recursive representations.

5 Remarks

- Fixing a policy transforms an MDP into a Markov chain with an explicitly defined probability measure.
- The Dynamic Bayesian Network representation makes the factorization of the joint distribution transparent.
- This construction provides the probabilistic foundation for value functions, policy evaluation, and reinforcement learning.