An Introduction to Stochastic Control and Reinforcement Learning

Subhrakanti Dey and Simone Garatti

Signals and Systems, Dept of Electrical Engineering, Uppsala University
Politechnico Di Milano, Milan

SIDRA Summer School, Bertinoro, Italy

July 2025

Course Content

- *July 7*: Dynamic Programming and its solutions, Closed form solution for the Linear Quadratic Gaussian (LQG) control problem, Infinite horizon stochastic control problems (discounted and average cost with finite state and action space), Bellman optimality equation, existence of stationary control policy,
- *July 8*: Curse of dimensionality in solving Dynamic Programming algorithms, Approximate Dynamic Programming algorithms – approximation in policy space and value space, contraction properties and error bounds, simulation based implementation
- *July 9*: Advanced Reinforcement Learning: policy gradient methods, actor-critic based reinforcement learning and their applications to continuous control (such as LQG) problems

Course Material

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, volumes 1 and 2, Athena Scientific, 2012

[2] Sutton and Barto, *Reinforcement Learning: Second Edition*, MIT Press.

[3] D.P.Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996

[4] https://www.mit.edu/ dimitrib/dpbook.html (All resources related to Bertsekas's lecture slides and videos)

[4] Relevant research papers will be referred to and distributed during the lectures.

Dynamic Programming and how to solve it

- Finite Horizon Dynamic Programming:
  $x_{k+1} = f(x_k, u_k, w_k)$, $u_k \in U_k(x_k)$, Transition probability $P(x_{k+1}|x_k, u_k)$, policy $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$, such that $u_k = \mu_k(x_k) \in U_k(x_k)$.
- Expected cost starting at x_0 under policy π is

$$J_\pi(x_0) = E[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k)],$$

$$J^*(x_0) = \min_\pi J(x_0) = J_{\pi^*}(x_0)$$

- Consider the tail subproblem starting at time i at x_i and we want to minimize the *cost-to-go* to time N with the tail policy $\{\mu_i^*, \ldots, \mu_{N-1}^*\}$

$$E[g_N(x_N) + \sum_{k=i}^{N-1} g_k(x_k, \mu_k(x_k), w_k)]$$

.

- *Principle of Optimality*: The tail policy is optimal for the tail subproblem (unaffected by past controls)
- DP recursively solves all the tail subproblems starting from $k = N - 1$.

Dynamic Programming (DP) algorithm

- Start with Value Function $V_N(x_N) = g_N(x_N)$
- Go backwards with

$$V_k(x_k) = \min_{u_k} E[g_k(x_k, u_k, w_k) + V_{k+1}(x_{k+1})], k = N-1, N-2, \ldots, 1$$

- The optimal cost is given by $V_0^*(x_0) = V_0(x_0)$ with the optimal policy obtained by $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$
- Proof by induction : Let's do it!

Dynamic Programming: Computational difficulty

- Note that the value function has to be evaluated for each value of the state x_k at each time k. For finite state and action space, this is possible but computationally intensive for large state and action space, complexity $O(N^{|X||U|})$, where $|X|, |U|$ are the cardinalities of the state and action space (when they are finite).
- For continuous state and action space, there are no closed form solutions to the value function in general.
- Exception: Linear Quadratic Control

$$x_{k+1} = Ax_k + Bu_k + w_k$$
$$g_k(.) = {x_k}^T Q x_k + {u_k}^T R u_k$$

- **Derivation of optimal control law**

$$J^*(u_0^*, u_1^*, \ldots, u_{n-1}^*) = \min_{u_0,\ldots,u_{N-1}} E\left[x_N^T Q_f x_N + \sum_{t=0}^{N-1} (x_t^T Q x_t + u_t^T R u_t)\right]$$

Infinite Horizon Stochastic Control Problems

- Minimize a total discounted cost

$$J_\pi(x_0) = \lim_{N\to\infty} \left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right]$$

$$\textit{subject to } x_{k+1} = f(x_k, u_k, w_k)$$

- Here $0 < \alpha < 1$ is a discount factor to keep the total cost bounded. The policy $\pi = \{\mu_0, \mu_1, \dots\}$, the expectation is taken over the noise/disturbance sequence $\{w_0, w_1, \dots\}$
- One needs to consider $\limsup$ if the limit does not exist.
- $J^*(x) = \min_\pi J_\pi(x), \forall x$ is the optimal cost with optimal policy π^*
- A policy π is stationary if $\pi = \{\mu, \mu, \dots\}$, and the corresponding cost function is defined by $J_\mu(x)$. The stationary policy μ is optimal if $J_\mu(x) = J^*(x), \forall x$.
- WHY CONSIDER INFINITE HORIZON COSTS?

Infinite Horizon Stochastic Control Problems

- Similar to the finite horizon scenario, one can consider a dynamic programming formulation under a stationary policy μ

$$(T_\mu V)(x) = E\left[g(x, \mu(x), w) + \alpha V(f(x, \mu(x), w))\right]$$

whereas

$$(TV)(x) = \min_{u \in U(x)} E_w\left[g(x, u, w) + \alpha V(f(x, u, w))\right]$$

- Here $T_\mu V$ may be viewed as the cost function associated with policy μ for a one-stage problem that has a stage cost g and terminal cost αV.
- For a k-stage problem one can similarly define a mapping $T_\mu^k V$ by applying the mapping T_μ recursively to $T_\mu^{k-1} V$, starting from $(T_\mu^0 V)(x) = V(x)$.
- MONOTONICITY LEMMA: One can easily show that for any two functions $V(x) \leq V'(x), \forall x$,
$(T_\mu^k V)(x) \leq (T_\mu^k V')(x)$, $(T^k V)(x) \leq (T^k V')(x)$, $\forall x, \forall k$.
- WHEN CAN WE EXPECT TO HAVE STATIONARY OPTIMAL POLICIES?

Infinite Horizon Stochastic Control Problems

- Generally, the existence of stationary optimal control policies depends on the nature of the cost function g, and the transition probability kernel $P(x_{k+1}|x_k, \mu(x_k))$ (or equivalently, the probability distribution of w_k)
- We will consider the simple case of finite state and action space and bounded cost per stage $|g(.)| \leq M$ for all values of (x, u, w). (Boundedness is not necessarily a very restrictive assumption!)
- *Convergence of the DP algorithm*: For the total discounted cost problem, one can show that for any bounded funvtion $V(x)$, the optimal cost function satisfies

$$V^*(x) = \lim_{N \to \infty} (T^N V)(x), \forall x$$

- It also follows that for a given stationary policy μ, one can show that

$$V_\mu(x) = \lim_{N \to \infty} (T_\mu^N V)(x), \forall x$$

Infinite Horizon Stochastic Control Problems

- Bellman's Equation of Optimality

$$V^*(x) = \min_{u \in U(x)} E\{g(x,u,w) + \alpha V^*(f(x,u,w)))\}$$

Or, equivalently, $V^* = TV^*$

- Furthermore, V^* is the unique solution of this equation within the class of bounded functions.
- One can similarly show that for every stationary policy μ, the associated cost function satisfies

$$V_\mu(x) = E\{g(x,\mu(x),w) + \alpha V_\mu(f(x,\mu(x),w)))\}$$

Or, equivalently, $V_\mu = T_\mu V_\mu$

Infinite Horizon Stochastic Control Problems: Discounted costs

- **Necessary and Sufficient Condition for Optimality** A stationary policy μ is opitmal if and only if $\mu(x)$ attains the minimum in Bellman's equation for each x, i.e.

$$TV^* = T_\mu V^*$$

- In case of a finite action set for each state x, an optimal stationary policy is guaranteed to exist.
- Finally, for any two bounded functions $V(x), V'(x)$, and any stationary policy μ, one can also show the following convergence rate for $k = 0, 1, \ldots$

$$\max_x |(T^k V)(x) - (T^k V')(x)| \leq \alpha^k \max_x |V(x) - V'(x)|$$

$$\max_x |(T_\mu^k V)(x) - (T_\mu^k V')(x)| \leq \alpha^k \max_x |V(x) - V'(x)|$$

- For finite state systems, where the state space $S = \{1, 2, \ldots, n\}$ and $u \in U(i)$, one can write

$$(TV)(i) = \min_{u \in U(i)} [g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) V(j)], \forall i = 1, 2, \ldots, n$$

and a similar equation for $(T_\mu V)$.

Infinite Horizon Stochastic Control Problems: Average Cost per stage (briefly)

- Finite state and action space
- Here we minimize the cost

$$J_\pi(x_0) = \lim_{N\to\infty} \frac{1}{N} E\{\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k))\}$$

  over all policies $\{\mu_0, \mu_1, \ldots,\}$ starting from a given initial state x_0.
- Optimality Conditions: If a scalar λ and an n-dimensional vector h satisfy

$$\lambda + h(i) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right], i = 1, 2, \ldots, n$$

  then λ is the optimal average cost per stage for all i, i.e.
  $\lambda = \min_\pi J_\pi(i) = J^*(i)$.
- $h(.)$ is known as the differential or relative cost vector since one can show that $h(i) - h(j) = (T^N h)(i) - (T^N h)(j)$ for all i, j.

Infinite Horizon Stochastic Control Problems: Average Cost per stage (briefly)

- For a stationary policy μ, one can show similarly, that if a scalar λ_μ and an n-dimensional vector h_μ satisfy

$$\lambda_\mu + h_\mu(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) h_\mu(j), \forall i$$

  then $\lambda_\mu = J_\mu(i)$.
- Whwn μ is a unichain policy (i.e the transition probability $P(\mu)$ has a single recurrent class), one can prove the existence and uniqueness of λ_μ, h_μ.
- Question: How to find such optimal policies?