

# Overview

---

Today's topics:

- Average cost criterion
- Some basic results characterizing the optimal cost
- Vanishing discount approach
- Relation of discounted costs to average cost – a single policy case
- Blackwell optimal policy – definition

Notes:

- Our overall plan is to select only a few topics from the average cost theory to study in the seminar.
- The chain structure of MDP is best to be revisited and understood when we study the theory.
- Keep in mind various “threads” that we have or will talk about, and try to fuse them at the end to form a whole picture.

## Definition of Average Cost

---

About limits: *liminf* and *limsup*

$$\liminf_{j \rightarrow \infty} a_j \stackrel{def}{=} \lim_{j \rightarrow \infty} \inf_{k \geq j} a_k \qquad \limsup_{j \rightarrow \infty} a_j \stackrel{def}{=} \lim_{j \rightarrow \infty} \sup_{k \geq j} a_k$$

Example: For the sequence  $\{a_j\} = \{1, -1, 1, -1 \dots\}$ ,

$$\liminf_{j \rightarrow \infty} a_j = -1 \qquad \limsup_{j \rightarrow \infty} a_j = 1.$$

*Liminf average cost* and *limsup average cost* of  $\pi \in \Pi^{\text{HR}}$ :

$$J_{-, \pi}(s) \stackrel{def}{=} \liminf_{N \rightarrow \infty} \frac{1}{N} E_s^\pi \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\}$$

$$J_{+, \pi}(s) \stackrel{def}{=} \limsup_{N \rightarrow \infty} \frac{1}{N} E_s^\pi \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\}$$

Notes:

- relation of average costs to the expected *finite-stage* costs
- meaning of the definition: asymptotically, liminf and limsup ave. costs are the “best-case” and “worst-case” ave. costs, respectively.

A trivial MDP example where  $J_-^\pi \neq J_+^\pi$ :

- a single state  $s$ , two actions  $\{a, b\}$  with

$$g(s, a) = -1, g(s, b) = 1$$

- a non-stationary policy  $\pi$ :  
apply  $a$  for  $2^k$  times, then apply  $b$  for  $2^{k+1}$  times, and so on
- time-average of  $T_n$ -stage costs, with  $T_n = \sum_{k=0}^n 2^k$ :

$$n = 2m + 1 : \frac{\sum_{k=0}^m \left( (-1) \cdot 2^{2k} + 2^{2k+1} \right)}{\sum_{k=0}^m \left( 2^{2k} + 2^{2k+1} \right)} \xrightarrow{m \rightarrow \infty} \frac{1}{3}$$

$$n = 2m : \frac{-1 + \sum_{k=1}^m \left( 2^{2k-1} + (-1) \cdot 2^{2k} \right)}{1 + \sum_{k=1}^m \left( 2^{2k-1} + 2^{2k} \right)} \xrightarrow{m \rightarrow \infty} -\frac{1}{3}$$

By convention, define the average cost of  $\pi$  to be the limsup cost.

optimal average cost:  $J^*(s) \stackrel{def}{=} \inf_{\pi \in \Pi^{\text{HR}}} J_{+, \pi}(s).$

## Under-Selectiveness of Average Cost Criterion

---

For the previous example, consider two policies:

$\pi_1$  : apply  $n$  times  $b$ , and apply  $a$  afterwards

$\pi_2$  : apply  $a$  always

Both policies are average cost optimal, for any number  $n$ .

Reason of under-selectiveness:

- Finite-stage performance does not matter; never too late to perform optimally (with exceptions)

$$\frac{1}{N} E_s^{\pi} \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\} = \underbrace{\frac{1}{N} E_s^{\pi} \left\{ \sum_{k=0}^{n-1} g(S_k, U_k) \right\}}_{\text{diminishes to 0}} + \frac{1}{N} E_s^{\pi} \left\{ \sum_{k=n}^{N-1} g(S_k, U_k) \right\}$$

- If all policies have finite undiscounted total costs (e.g., they always reach some absorbing destination state, as in *stochastic shortest path problems*), then all policies are average cost optimal.
- Exceptions: on transient states the effect of suboptimal actions can be “irreversible,” when there are multiple recurrent classes.

Advanced topics: more selective optimality criteria (Section 5.4.2, 5.4.3 of [Put94]), e.g.,

- over-taking optimality
- average over-taking optimality
- $n$ -discount optimality
- Blackwell optimality (to be mentioned today)

## Some Basic Results on the Optimal Ave. Cost $J^*$

---

We can already derive a few results for average cost problems using mostly what we have already studied:

Adequacy of Markov Policies (Theorem 1 of Lecture 1):

$$J^*(s) = \inf_{\pi \in \Pi^{\text{MR}}} J_{+, \pi}(s).$$

Indeed, stronger claims hold: for any given  $s$ ,

$$\forall \pi \in \Pi^{\text{HR}}, \exists \pi' \in \Pi^{\text{MR}} \text{ s.t.}$$

$$J_{+, \pi'}(s) = J_{+, \pi}(s), \quad J_{-, \pi'}(s) = J_{-, \pi}(s)$$

Next: derive two lower bounds of  $J^*(s)$

## A Lower Bound from the Finite-Stage Optimal Costs

Notation:  $J_{k,\pi}$ ,  $k$ -stage cost of  $\pi$ ;  $J_0(s) = 0, \forall s$ .

Recall:  $T^k J_0$  is the optimal  $k$ -stage cost function.

$$\begin{aligned} \forall \pi, s, \quad & (T^k J_0)(s) \leq J_{k,\pi}(s), \\ \implies & \frac{1}{k}(T^k J_0)(s) \leq \frac{1}{k}J_{k,\pi}(s) \\ \implies & \limsup_{k \rightarrow \infty} \frac{1}{k}(T^k J_0)(s) \leq \limsup_{k \rightarrow \infty} \frac{1}{k}J_{k,\pi}(s) = J_{+,\pi}(s) \\ \implies & \limsup_{k \rightarrow \infty} \frac{1}{k}(T^k J_0)(s) \leq \inf_{\pi} J_{+,\pi}(s) = J^*(s), \quad \forall s. \end{aligned}$$

Note: for finite-state and action MDPs,  $T^k J_0$  indeed converges to  $J^*$ .

## A Lower Bound from the Discounted Optimal costs

**A Tauberian theorem** (for a proof, see Lemma 8.10.6, pp. 417 of [Put94]):

Let  $\{a_j\}$  be a bounded sequence of scalars. Let  $s_n = \sum_{j=0}^{n-1} a_j$ . Then

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{s_n}{n} &\leq \liminf_{x \uparrow 1} (1-x) \sum_{j=0}^{\infty} x^j a_j \\ &\leq \limsup_{x \uparrow 1} (1-x) \sum_{j=0}^{\infty} x^j a_j \\ &\leq \limsup_{n \rightarrow \infty} \frac{s_n}{n}. \end{aligned}$$

Note: it is very useful in relating the limits of the time-average costs to the limits of the discounted costs.

Notation:  $J_\alpha^*$ , optimal cost of the  $\alpha$ -discounted problem

**Proposition 1.**

$$\limsup_{\alpha \uparrow 1} (1 - \alpha) J_\alpha^*(s) \leq J^*(s), \quad \forall s.$$

*Proof.* The discounted cost of any policy  $\pi$  is always no less than  $J_\alpha^*$ :

$$J_\alpha^*(s) \leq E_s^\pi \left\{ \sum_{k=0}^{\infty} \alpha^k g(S_k, U_k) \right\}.$$

By the Tauberian theorem,

$$\limsup_{\alpha \uparrow 1} (1 - \alpha) E_s^\pi \left\{ \sum_{k=0}^{\infty} \alpha^k g(S_k, U_k) \right\} \leq \limsup_{N \rightarrow \infty} \frac{1}{N} E_s^\pi \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\},$$

so

$$\limsup_{\alpha \uparrow 1} (1 - \alpha) J_\alpha^*(s) \leq \limsup_{N \rightarrow \infty} \frac{1}{N} E_s^\pi \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\},$$

$$\limsup_{\alpha \uparrow 1} (1 - \alpha) J_\alpha^*(s) \leq \inf_{\pi \in \Pi^{\text{HR}}} \limsup_{N \rightarrow \infty} \frac{1}{N} E_s^\pi \left\{ \sum_{k=0}^{N-1} g(S_k, U_k) \right\} = J^*(s).$$

□

## Vanishing Discount Approach

---

About this approach: from the  $\alpha$ -discounted optimality equations, we take  $\alpha \uparrow 1$  and obtain (under some conditions) in the limit an equation that is the optimality equation for the average cost problem.

Note: it is widely applied for general models; we illustrate a basic form of it for finite-state and action models.

The optimality equation in the discounted case:

$$J_{\alpha}^*(s) = \min_{u \in \mathcal{U}(s)} \left[ g(s, u) + \alpha \sum_{s'} p(s'|s, u) J_{\alpha}^*(s') \right]$$

Subtract  $\alpha J_{\alpha}^*(\bar{s})$ ,  $\bar{s}$  being some fixed reference state, from both sides:

$$(1-\alpha)J_{\alpha}^*(\bar{s}) + (J_{\alpha}^*(s) - J_{\alpha}^*(\bar{s})) = \min_{u \in \mathcal{U}(s)} \left[ g(s, u) + \alpha \sum_{s'} p(s'|s, u) (J_{\alpha}^*(s') - J_{\alpha}^*(\bar{s})) \right] \quad (*)$$

Boundedness conditions:  $\exists$  some constants  $C_1, C_2$ , for all  $s, \alpha < 1$ ,

$$|(1 - \alpha)J_{\alpha}^*(\bar{s})| < C_1 \quad (\text{i})$$

$$|J_{\alpha}^*(s) - J_{\alpha}^*(\bar{s})| < C_2 \quad (\text{ii})$$

Under (i) and (ii), we can **extract a sequence**  $\alpha_k \uparrow 1$  such that

$$(1 - \alpha_k)J_{\alpha_k}^*(\bar{s}) \rightarrow \lambda, \quad (\lambda \text{ is a scalar})$$

$$h_k(s) \stackrel{\text{def}}{=} J_{\alpha_k}^*(s) - J_{\alpha_k}^*(\bar{s}) \rightarrow h(s), \quad (h \text{ is a function})$$

From (\*):

$$\lambda + h(s) = \lim_{k \rightarrow \infty} \min_{u \in \mathcal{U}(s)} \left[ g(s, u) + \alpha_k \sum_{s'} p(s'|s, u) h_k(s) \right]$$

**Exchange lim and min:**

$$\lambda + h(s) = \min_{u \in \mathcal{U}(s)} \left[ g(s, u) + \sum_{s'} p(s'|s, u) h(s) \right]$$

Important: [where](#) have we used finiteness of the model

Average cost optimality equation (ACOE):

$$\lambda + h(s) = \min_{u \in \mathcal{U}(s)} \left[ g(s, u) + \sum_{s'} p(s'|s, u) h(s) \right]$$

To understand the ACOE (proofs in later classes, hopefully):

- If ACOE has a solution  $(\lambda, h)$ , then the optimal average cost  $J^*$  is a constant function and  $J^*(s) = \lambda$ .
- If ACOE has a solution  $(\lambda, h)$ , then any policy greedy with respect to  $h$  is average cost optimal.
- Meaning of  $h$ : from the derivation, it is the limit of the relative cost difference  $J_{\alpha_k}^*(s) - J_{\alpha_k}^*(\bar{s})$ ; if  $\lambda = 0$ , then  $h$  is indeed the total undiscounted cost function.
- $h$  is hard to interpret; we will understand it better later through other analyses.
- This ACOE is for the special case where the optimal average cost is constant. We will study the general form of ACOE later.

## Summary on vanishing discount approach

- Verifying conditions (i) and (ii)

$$|(1 - \alpha)J_{\alpha}^*(\bar{s})| < C_1 \quad (\text{i})$$

$$|J_{\alpha}^*(s) - J_{\alpha}^*(\bar{s})| < C_2 \quad (\text{ii})$$

– (i) is true whenever  $g(s, u)$  is bounded (obviously true in finite-state and action models).

– (ii) can be verified easily in some cases:

A machine repair example

$\bar{s}$ , machine in perfect status; for any  $\alpha$ ,  $\bar{s}$  is always the state with the smallest discounted total cost.

$\bar{a}$ , an action to replace the machine with a new perfect one

$$\forall s, J_{\alpha}^*(s) - J_{\alpha}^*(\bar{s}) \leq g(s, \bar{a}) - (1 - \alpha)J_{\alpha}^*(\bar{s}).$$

- Usefulness of the approach in general space models:

We can exploit the cost structure of the problem, in addition to the chain structure, to characterize the optimal cost and policy.

## Re-examine Discounted Costs of a Stationary Policy

Purpose: relation to average cost and understanding  $h$

Consider a stationary policy  $\mu$ ;  $\mu$  induces a time-homogeneous Markov chain, so its average cost function  $J_\mu$  is

$$J_\mu = P_\mu^* g_\mu, \quad \text{where } P_\mu^* = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P_\mu^k.$$

Notes:

- Its liminf and limsup average costs are equal, since the limit exists.
- On each recurrent set  $C_i$  of states,  $J_\mu$  has the same value  $\lambda_i$ ; on a transient state  $s$ ,  $J_\mu = \sum_i p_i \lambda_i$  where  $p_i$  is the probability of being absorbed into the  $i$ th class when start from  $s$ .

By the Tauberian theorem, we can already conclude that

$$(1 - \alpha)J_{\alpha, \mu}^* \rightarrow J_\mu.$$

What is the difference  $J_{\alpha, \mu}^* - \frac{J_\mu}{(1-\alpha)}$ ?

There exists a function  $h_\mu$ , which together with  $J_\mu$  satisfy

- for the discounted costs:

$$J_{\alpha,\mu} = \frac{J_\mu}{(1-\alpha)} + h_\mu + O(1-\alpha)$$

- for the finite-stage costs, denote by  $J_{N,\mu}$ , (assume aperiodicity):

$$J_{N,\mu} = NJ_\mu + h_\mu + o(1)$$

Interpreting  $h_\mu$ :

- As  $\frac{1}{1-\alpha}$  or  $N$  increases,  $J_{\alpha,\mu}(s)$  or  $J_{N,\mu}(s)$  grows at the rate of  $J_\mu(s)$ .
- Within one recurrent set of states,  $h_\mu(s)$  characterizes the relative amount of “advantage” in the total cost by starting from  $s$ .

Rigorous derivations: Laurent series expansion (study later, hopefully)

Terminologies:  $J_\mu$ , average cost, gain;  $h_\mu$ , differential cost, bias

## Blackwell Optimality

---

Definition: A stationary policy  $\mu$  is called *Blackwell optimal*, if there exists some positive  $\bar{\alpha} < 1$  such that  $\mu$  is optimal for all  $\alpha$ -discounted problems with  $\alpha \in (\bar{\alpha}, 1)$ .

Next time (when we have a chance):

- existence of Blackwell optimal policy in finite-state and action models
- derive from the Blackwell optimal policy the general form of ACOE, a pair of optimality equations