ECE586 MDPs and Reinforcement Learning **University of Illinois at Urbana-Champaign**
Spring 2019 **Acknowledgment: R. Srikant's Notes and Discussions**

Lecture 9: Discounted Cost MDPs, Value and Policy Iteration, Monotone Policies

Instructor: Dimitrios Katselis *Scribe: Tianhao Wu, Xiaoyang Bai, Junchi Yang, Shen Yan*

9.1 Introduction

Sequential decision making under uncertainty: A probabilistic sequential decision model is represented as a stochastic process under the (partial) control of an observer. At each time instant or decision epoch, the state occupied by the process is observed and the controller takes an action, which affects the state occupied by the system at the next time instant or decision epoch. Also, at every time instant there is an incurred reward or cost associated with the selected action. In general, a *decision rule* or *policy* for choosing actions at every time instant may depend on the current state and the history of all past states and actions. Implementing a policy generates a sequence of rewards or costs. The sequential decision problem is to choose, prior to the first decision epoch, a policy to maximize a function of this reward sequence or to minimize a function of this cost sequence. Due to the difficulty of analyzing processes that allow arbitrarily complex dependencies between past and future events, it is customary to focus on Markov Decision Processes (MDP), which lead to tractable sequential decision making setups. In the MDP model, the set of available actions, the rewards or costs and the transition probabilities depend only on the current state and action and not on past states and past actions. The MDP model is sufficiently broad to allow modeling most realistic sequential decision-making problems.

A **Markov Decision Process (MDP)** is a Markov process with feedback control. More precisely, the underlying Markov chain is *controlled* by controlling the state transition probabilities. We denote the transition probabilities by $P_{ij}(u) = P(x_{k+1} = j | x_k = i, u_k = u)$, where $u \in \mathcal{U}$ is the control action taken at time k. We will assume that the state space $\mathcal{X}$ and the action space $\mathcal{U}$ are finite sets. When the action u is taken, a cost $c(x, u)$ (can be deterministic or random) is incurred, which is assumed to take values in a finite set. Without loss of generality we assume that[1] $c(x, u) \geq 0$, otherwise we can add a positive constant to make all of them nonnegative. We also note that immediate deterministic or immediate expected costs are known before the decision maker takes an action. Although our focus here will be on the infinite horizon discounted cost setup, we clarify that in general an MDP is characterized by a set $\mathcal{T} \subseteq [0, \infty]$, which is called **set of decision epochs**. Decision epochs are the time instants in which the agent takes actions. We mainly consider processes with countably many decision epochs yielding a discrete $\mathcal{T}$, which is usually taken to be $\mathcal{T} = \{0, 1, \ldots, N\}$ (finite-horizon problems) or $\mathcal{T} = \{0, 1, \ldots\}$ (infinite-horizon problems). Time is divided into time periods or stages and each decision epoch occurs at the beginning of a time period.

Note: $\mathcal{T}$ can be a continuum, in which case decisions are made either *continuously*, or at random time points when certain events occur, e.g., arrivals to a queueing system, or at opportunistic times chosen by the decision maker.

Infinite Horizon MDPs: In this course, we will focus on infinite-horizon MDPs with stationary (i.e., time-invariant) transition probabilities and costs. The set of decision epochs will be assumed to be $\mathcal{T} = \mathbb{Z}_{\geq 0}$. We have already discussed in a previous file that it is sufficient to consider the class of deterministic Markovian policies, i.e., policies where the optimal action(s) u_k^* at time instant k is a deterministic function $\mu_k^*(x_k)$ of the current state, to optimally solve different MDPs. In the context of infinite horizon problems, we will focus on *stationary policies*, i.e., policies

[1] In the literature, this is the central assumption to what is called *negative programming*. In relevance to discounted cost MDPs that we will focus on in this lecture, although in negative programming the discount factor α (that will be introduced shortly) is allowed to take any value in $[0, 1]$, α is usually taken to be 1.

of the form $\{\mu, \mu, \ldots\}$ which use the same mapping of states to actions at every decision epoch. As mentioned earlier, our focus here will be on infinite horizon discounted cost problems. Nevertheless, in subsequent lectures we will look into infinite horizon average cost MDPs. We therefore provide here existing metrics to evaluate the performance of a given policy π in infinite horizon setups, where π can be stationary or in most generality it can be history-dependent and randomized:

- **Expected Total Cost of Policy** π: For a given initial state $x_0 = i$, the performance of a policy π is measured by the metric:

$$\tilde{J}_\pi(i) = \lim_{N\to\infty} E^\pi \left[\sum_{k=0}^{N} c(x_k, u_k) \middle| x_0 = i \right]. \tag{9.1}$$

  We note here that the limit exists (although it may be ∞) due to monotone convergence under our assumptions on the costs (nonnegativity). Also,

$$\lim_{N\to\infty} E^\pi \left[\sum_{k=0}^{N} c(x_k, u_k) \middle| x_0 = i \right] = E^\pi \left[\sum_{k=0}^{\infty} c(x_k, u_k) \middle| x_0 = i \right] \tag{9.2}$$

  due to *Lebesgue Monotone Convergence Theorem*, i.e., the order of the limit and the expectation can be interchanged.

- **Expected Total Discounted Cost of Policy** π: For a given initial state $x_0 = i$ and a discount factor $\alpha \in [0, 1)$, the performance of a policy π is measured by the metric:

$$J_\pi(i) = \lim_{N\to\infty} E^\pi \left[\sum_{k=0}^{N} \alpha^k c(x_k, u_k) \middle| x_0 = i \right]. \tag{9.3}$$

  The limit is guaranteed to exist and be finite if $c(x, u)$ is uniformly bounded, i.e., $0 \leq c(x, u) \leq C$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$, since then $J_\pi(i) \leq \frac{C}{1-\alpha}$ for any $i \in \mathcal{X}$, i.e., $J_\pi(i)$ is also uniformly bounded. This assumption holds in our case due to the finiteness of the set in which $c(x, u)$ takes values. Additionally, since (9.2) holds, we have that

$$\lim_{\alpha\uparrow 1} J_\pi(i) = \tilde{J}_\pi(i). \tag{9.4}$$

  Finally, under our assumptions on the costs

$$\lim_{N\to\infty} E^\pi \left[\sum_{k=0}^{N} \alpha^k c(x_k, u_k) \middle| x_0 = i \right] = E^\pi \left[\sum_{k=0}^{\infty} \alpha^k c(x_k, u_k) \middle| x_0 = i \right] \tag{9.5}$$

  due to *Lebesgue Monotone Convergence Theorem* or due to the *Bounded Convergence Theorem*, i.e., the order of the limit and the expectation can be interchanged.

- **Average Cost of Policy** π: For a given initial state $x_0 = i$, the performance of a policy π is measured by the metric:

$$\bar{J}_\pi(i) = \lim_{N\to\infty} \frac{1}{N+1} E^\pi \left[\sum_{k=0}^{N} c(x_k, u_k) \middle| x_0 = i \right]. \tag{9.6}$$

  If the limit does not exist (relaxing for the moment any constraints on the cost values), then $\liminf$ and $\limsup$ can instead be taken (which are guaranteed to exist, although they can be infinite) and they provide upper and lower bounds on the average cost achieved by policy π.

Remarks:

1. Consider (9.2), (9.5) and the exchange of the order of the limit and the expectation. The left-hand side in these two equations is the limit of an expectation, while the right-hand side is the expectation of a limit. The expectation of a limit is significantly more complicated than the limit of an expectation, because it requires an infinite-dimensional probability distribution. On the other hand, the limit of an expectation requires only taking successively larger and larger expectations of *finite-dimensional* distributions. Although under our assumptions on the costs, (9.2) and (9.5) hold, the left-hand sides in these two equations considered in isolation are in principle mathematically simpler.

2. We will investigate in later lectures average cost MDPs. In general, although infinite horizon discounted cost MDPs are simple, average cost problems require a lot of technicalities in their analyses. Depending on the structure of the transition matrices $P(u)$, an optimal stationary policy need not even exist for such problems.

3. Discounted models have a complete theory and numerous computational algorithms are available for their solution. Discounting is a natural mechanism of weighting more current rewards or costs over delayed rewards or costs, which can be important, e.g., in economical or biological (evolutionary) problems. The discounting parameter α is also referred to as a *forgetting factor* in various adaptive filtering applications and in such contexts it gives exponentially less weight to older errors. An alternative way to interpret discounting is by considering the horizon N *random and independent* of the actions of the decision maker and the sequence of states occupied by the system up to the final decision epoch. Let $\tilde{J}_\pi^N(i)$ be the expected total cost by using policy π when the horizon N is random. We then have:

$$\tilde{J}_\pi^N(i) = E^\pi\left[E_N\left[\sum_{k=0}^{N} c(x_k,u_k)\right]\middle| x_0=i\right] = E^\pi\left[\sum_{n=0}^{\infty} P(N=n)\sum_{k=0}^{n} c(x_k,u_k)\middle| x_0=i\right].$$

Proposition 9.1.1. *Suppose that N has a geometric distribution with parameter $1-\alpha$, i.e., $P(N=n) = (1-\alpha)\alpha^n, n=0,1,\ldots$. Then, $\tilde{J}_\pi^N(i) = J_\pi(i)$ for any $i \in \mathcal{X}$.*

Proof. Since N is geometrically distributed and independent of the process $\{(x_k,u_k)\}_{k\geq 0}$ (recall that u_k can have an arbitrary dependency on the history $\mathcal{F}_k = \{x_0,u_0,x_1,u_1,\ldots,x_{k-1},u_{k-1},x_k\}$ up to time k per the choice of π), we have that:

$$\begin{aligned}
\tilde{J}_\pi^N(i) &= E^\pi\left[\sum_{n=0}^{\infty}(1-\alpha)\alpha^n\sum_{k=0}^{n} c(x_k,u_k)\middle| x_0=i\right]\\
&= E^\pi\left[\sum_{k=0}^{\infty} c(x_k,u_k)\sum_{n=k}^{\infty}(1-\alpha)\alpha^n\middle| x_0=i\right]\\
&= E^\pi\left[\sum_{k=0}^{\infty} \alpha^k c(x_k,u_k)\sum_{n=k}^{\infty}(1-\alpha)\alpha^{n-k}\middle| x_0=i\right]\\
&= E^\pi\left[\sum_{k=0}^{\infty} \alpha^k c(x_k,u_k)\middle| x_0=i\right] = J_\pi(i),
\end{aligned}$$

where we have used the fact that $\sum_{n=0}^{\infty}\alpha^n = \frac{1}{1-\alpha}$. The computation is justified if we can interchange the order of summation. This interchange is a consequence of the Fubini-Tonelli theorem for series[2] since

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}|(1-\alpha)\alpha^n c(x_k,u_k)| < \frac{C}{1-\alpha} < \infty,$$

[2] or simply due to Tonelli's theorem for series by the nonnegativity of the summands according to our assumptions.

where $0 \leq c(x,u) \leq C$ for all (x,u).

□

Our focus here is on **infinite-horizon discounted cost MDPs**. In general, the policy $\mu_k(\cdot)$ at time k can be deterministic or randomized and the action at time k may satisfy $u_k = \mu_k(\mathcal{F}_k)$ or $u_k \sim \mu_k(\mathcal{F}_k)$. As mentioned earlier, in the considered MDPs here, we will focus on stationary deterministic policies of the form $u_k = \mu(x_k)$ at any time k.

9.2 Stationary policies

Why stationary policies in stationary infinite horizon problems? Consider an infinite horizon problem with stationary transition probabilities and costs. Let $\{\mu_0^*, \mu_1^*, \ldots\}$ be an optimal policy for this problem. It is intuitive that as $N \to \infty$, the optimal policy μ_0^* at time 0 loses its dependence on N. Thus, μ_0^* is expected to be optimal for subsequent time instants.

9.2.1 Performance of a Stationary Deterministic Markovian Policy

We are interested in the performance of a given stationary policy $\{\mu, \mu, \ldots\}$. Let

$$J_\mu(i) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N} \alpha^k c(x_k, u_k) \middle| x_0 = i\right] = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N} \alpha^k c(x_k, \mu(x_k)) \middle| x_0 = i\right],$$

where we have dropped the superscript $\cdot^\mu$ from the expectation for easiness. We have already argued that $J_\mu(i) < \infty$ for any $i \in \mathcal{X}$, i.e., the limit exists and is finite (cf. (9.3)) due to monotone convergence. We further note that:

$$\begin{aligned} J_\mu(i) &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N} \alpha^k c(x_k, \mu(x_k)) \middle| x_0 = i\right] = \lim_{N\to\infty} E\left[c(x_0, \mu(x_0)) + \sum_{k=1}^{N} \alpha^k c(x_k, \mu(x_k)) \middle| x_0 = i\right] \\ &= \bar{c}(i, \mu(i)) + \sum_j P_{ij}(\mu(i)) \lim_{N\to\infty} E\left[\sum_{k=1}^{N} \alpha^k c(x_k, \mu(x_k)) \middle| x_1 = j\right] \\ &= \bar{c}(i, \mu(i)) + \alpha \sum_j P_{ij}(\mu(i)) J_\mu(j), \end{aligned} \tag{9.7}$$

where $\bar{c}(i, \mu(i)) = E[c(x_0, \mu(x_0)) | x_0 = i]$.

Let $J_\mu = [J_\mu(1), J_\mu(2), \ldots, J_\mu(X)]^T$ where $\mathcal{X} = \{1, 2, \ldots, X\}$ and $\bar{c}_\mu = [c(1, \mu(1)), c(2, \mu(2)), \ldots, c(X, \mu(X))]^T$. Then, (9.7) can be compactly written as

$$J_\mu = \bar{c}_\mu + \alpha P_\mu J_\mu, \tag{9.8}$$

where $P_\mu = [P_{ij}(\mu(i))]$.

Theorem 9.2.1. *There exists a unique solution J_μ to (9.8):*

$$J_\mu = (I - \alpha P_\mu)^{-1} \bar{c}_\mu. \tag{9.9}$$

Proof. Obtaining (9.9) from (9.8) is straightforward if $I - \alpha P_\mu$ is nonsingular. Additionally, the uniqueness of J_μ is guaranteed by the invertibility of $I - \alpha P_\mu$. This invertibility is equivalent to $I - \alpha P_\mu$ having only nonzero eigenvalues. We note that P_μ is a stochastic matrix. We will prove the existence and uniqueness of J_μ by showing that for a stochastic matrix P we have $|\lambda_i(P)| \leq 1, \forall i$, where $\lambda_i(P)$ is the ith eigenevalue of P. Combining this property with

the fact that $\alpha < 1$ will yield the conclusion that all the eigenvalues of $I - \alpha P_\mu$, which are equal to $1 - \alpha\lambda_i(P_\mu), i = 1, 2, \ldots, X$, are nonzero.

We first show that the eigenvalues of $I - \alpha P_\mu$ are $1 - \alpha\lambda_i(P_\mu), i = 1, 2, \ldots, X$. Let A be a square matrix and $B = I - aA$ for some $a \in \mathbb{R}$. Assume that the eigenvector of A associated with its ith eigenvalue is u: $Au = \lambda_i(A)u$. Then:

$$Bu = (I - aA)u = u - aAu = u - a\lambda_i(A)u = (1 - a\lambda_i(A))u = \lambda_i(B)u.$$

Therefore, $\lambda_i(B) = 1 - a\lambda_i(A), \forall i$, and A, B have the same set of eigenvectors.

We now need to show that $|\lambda_i(P_\mu)| \leq 1, \forall i$.

Lemma 9.2.2. *For a stochastic matrix P, $|\lambda_i(P)| \leq 1, \forall i$. Here, $|\cdot|$ denotes complex modulus.*

Proof. To prove Lemma 9.2.2, we first need to introduce some definitions.

9.2.1.1 Norms

Definition 9.2.3. *Let $x \in \mathbb{R}^n$. A norm $\|\cdot\|$ is a function from $\mathbb{R}^n$ to $\mathbb{R}_{\geq 0}$ satisfying:*

(i) $\|x\| = 0 \iff x = 0$

(ii) $\|ax\| = |a|\,\|x\|, \forall a \in \mathbb{R}$ *and* $\forall x \in \mathbb{R}^n$ *(absolute homogeneity)*

(iii) $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n$ *(subadditivity or triangle inequality)*

Note: A **seminorm** satisfies (ii) and (iii) in the above definition, i.e., it is allowed to assign zero length to some nonzero vectors in addition to the zero vector.

Generalization: In full generality, norms are defined on vector spaces $\mathcal{V}$ over the field $\mathbb{C}$ of complex scalars. In this case $\|\cdot\| : \mathcal{V} \to \mathbb{R}_{\geq 0}$ and the vector space $\mathcal{V}$ together with a norm $\|\cdot\|$ is called a **normed linear space** or simply a **normed space**. Additionally, if $\mathcal{W}$ is a subspace of a normed space $\mathcal{V}$, then $\mathcal{W}$ is also a normed space with respect to the norm on $\mathcal{V}$ (restricted on $\mathcal{W}$). Finally, if $\mathcal{V}$ is a normed space, then $d(u, v) = \|u - v\|$ for $u, v \in \mathcal{V}$ defines a metric on $\mathcal{V}$, called **induced metric**. Not every metric is induced by a norm. A direct implication of the induced metric is that all normed spaces are metric spaces. In particular, the notions of *convergent* and *Cauchy sequences* apply in any normed space.

Implication: From the triangle inequality we easily get that $|\|x\| - \|y\|| \leq \|x - y\|$.

Definition 9.2.4. ***p-norms** or **ℓ_p-norms**:*

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p} = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}, \quad 1 \leq p \leq \infty.$$

The following norms are of particular interest:

- The ℓ_1**-norm**: $\|x\|_1 = |x_1| + \cdots + |x_n|$. Sometimes, this norm is referred to as the **taxicab norm** or **Manhattan norm**.
- The ℓ_2**-norm**: $\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{x^T x}$. It is also called **Euclidean norm**.
- The ℓ_∞**-norm**: $\|x\|_\infty = \max_i |x_i|$. It is also called **infinity norm** or **maximum norm** or **sup-norm**.

Note: It is generally easy to show that ℓ_p-norms are true norms for $p = 1, \infty$. It is nontrivial to show the triangle inequality for $1 < p < \infty$. The ℓ_2-norm satisfies the *Cauchy-Bunyakovsky-Schwarz inequality* (or simply Cauchy-Schwarz inequality) $|x^T y| \leq \|x\|_2 \|y\|_2$, which can be used to prove the triangle inequality for the ℓ_2-norm. The Cauchy-Bunyakovsky-Schwarz inequality is a special case of *Hölder's inequality*: $|x^T y| \leq \|x\|_p \|y\|_q, \ \frac{1}{p} + \frac{1}{q} = 1$. The numbers p and q are said to be *Hölder conjugates* of each other or *conjugate exponents*. Hölder's inequality can be used to show the triangle inequality for $1 < p < \infty$. We finally note that the triangle inequality for the ℓ_p-norm is also known as the *Minkowski inequality*.

Definition 9.2.5. ***Equivalence of vector norms****: We say that two vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are* equivalent *if there exist constants $0 < C_1 \leq C_2$ such that for any vector $x \in \mathbb{R}^n$*

$$C_1 \|x\|_a \leq \|x\|_b \leq C_2 \|x\|_a,$$

where C_1, C_2 are independent of x.

Remark: Two norms on a vector space $\mathcal{V}$ are equivalent if they define the same open subsets of $\mathcal{V}$.

Important Property for norms on finite-dimensional spaces: Any two norms on a finite-dimensional real or complex vector space are equivalent.

Note: It follows that if two norms are equivalent, then a sequence of vectors that converges to a limit with respect to one norm will converge to the same limit with respect to the other norm. All ℓ_p-norms are *equivalent* by the previous property.

To measure the magnitude of a matrix or the distance between matrices we introduce matrix norms:

Definition 9.2.6. ***Matrix norm****: A* matrix norm *is a function $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}_{\geq 0}$, where $\mathbb{R}^{m \times n}$ is the vector space of all $m \times n$ matrices with real entries, with the following properties:*

1. $\|A\| = 0 \iff A = 0$
2. $\|aA\| = |a| \, \|A\|, \ \forall a \in \mathbb{R}$ *and* $\forall A \in \mathbb{R}^{m \times n}$ *(absolute homogeneity)*
3. $\|A + B\| \leq \|A\| + \|B\|, \ \forall A, B \in \mathbb{R}^{m \times n}$ *(subadditivity or triangle inequality)*

A property that is often, but not always, included in the definition of a matrix norm is the submultiplicative property*: if $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ we require that*

$$\|AB\| \leq \|A\| \|B\|.$$

This property is very useful when A, B are square matrices. The submultiplicative property is satisfied by some but not all matrix norms

Any vector norm induces a matrix norm. Focusing on matrix norms induced by ℓ_p-norms, we have the following definition:

Definition 9.2.7. ***Induced norm*** *or* ***Operator norm by an ℓ_p-norm****: A matrix $A \in \mathbb{R}^{m \times n}$ induces a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$ with respect to the standard basis. Let a p-norm for some $1 \leq p \leq \infty$ be used for both $\mathbb{R}^m$ and $\mathbb{R}^n$. We can then define the following* induced norm *or* operator norm *on the space $\mathbb{R}^{m \times n}$ of all $m \times n$ matrices with real entries:*

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup\{\|Ax\|_p : x \in \mathbb{R}^n \ \text{ such that } \ \|x\|_p = 1\}.$$

Note: Any induced operator norm is a submultiplicative matrix norm.

Intuition: $\|A\|_p$ quantifies in the ℓ_p-norm the effect of the action of A on vectors of unit length, when this length is measured in the ℓ_p-norm.

The following matrix norms are of particular interest:

- The ℓ_1**-norm**: $\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |A_{ij}|$, i.e., the maximum column sum of absolute elements.
- The ℓ_2**-norm**: $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$, where $\lambda_{\max}(\cdot)$ corresponds to the maximum eigenvalue of the matrix argument and $\sigma_{\max}(\cdot)$ corresponds to the maximum singular value of the matrix argument.
- The ℓ_∞**-norm**: $\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |A_{ij}|$, i.e., the maximum row sum of absolute elements.

Definition 9.2.8. *Treating $A \in \mathbb{R}^{m\times n}$ as vector in $\mathbb{R}^{mn}$ we define the **Frobenius** norm or **Hilbert-Schmidt** norm or **Schur** norm:*

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\mathrm{tr}(AA^T)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)},$$

where $\{\sigma_i(A)\}$ are the singular values of A.

Note: The Frobenius norm is not induced by any vector ℓ_p-norm. Moreover, it is submultiplicative (this can be easily proved by an application of the Cauchy-Schwarz inequality).

Equivalence of matrix norms: Like vector norms, matrix norms are equivalent.

A final result of importance:

Theorem 9.2.9. *For every norm on $\mathbb{R}^n$ (or $\mathbb{C}^n$) and every matrix $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$), there exists a real constant $C_A > 0$ such that $\|Au\| \leq C_A\|u\|, \forall u \in \mathbb{R}^n$ (or $\mathbb{C}^n$).*

In plain words: Every linear map on a finite-dimensional space is **bounded**. This is a consequence of the fact that the unit ball in finite-dimensional spaces is compact. Moreover, this implies that every linear map on a finite dimensional space is (uniformly) **continuous**. To see this, apply the above theorem with $u = x - y$ for any $x, y \in \mathbb{R}^n$ or $\mathbb{C}^n$.

9.2.1.2 Proof of Lemma 9.2.2

Proof. A well-known result in linear algebra states that:

$$\underbrace{\rho(A) = \max_i |\lambda_i(A)|}_{\text{spectral radius of } A} \leq \|A\|,$$

where $\|\cdot\|$ is any matrix norm and A is a square matrix. Applying the above result to a stochastic matrix P for the induced norm by the ℓ_∞-norm we obtain:

$$|\lambda_i(P)| \leq \|P\|_\infty = 1, \forall i,$$

where the last equality is due to the stochasticity of P. □

Back to the proof of Theorem 9.2.1:

Thus, $I - \alpha P_\mu$ is nonsingular and (9.8) has a unique solution given by (9.9). □

9.2.2 Proof 2 of Theorem 9.2.1

Theorem 9.2.10. ***Contraction Mapping Theorem (Banach)****: Let $\tilde{T} : \mathcal{D} \to \mathcal{D}$, where $\mathcal{D} \subseteq \mathbb{R}^n$ is a closed set. Assume that $\tilde{T}$ satisfies a "contraction property", i.e, $\exists\, a \in [0, 1)$ such that*

$$\|\tilde{T}(x) - \tilde{T}(y)\| \leq a\|x - y\|, \, \forall x, y \in \mathcal{D}$$

where $\|\cdot\|$ can be any norm (e.g., $\|\cdot\|_p$ for $1 \leq p \leq \infty$).

Then:

(a) $\exists$ a unique fixed point $x^* \in \mathcal{D}$ of $\tilde{T}$, i.e, $x^* = \tilde{T}(x^*)$.

(b) Starting at any $x_0 \in \mathcal{D}$:

$$x_{k+1} = \tilde{T}(x_k) \xrightarrow[k\to\infty]{} x^*$$

at a geometric rate.

Note: Contractions are **Lipschitz maps** with a Lipschitz constant smaller than 1. Therefore, they are (uniformly) *continuous* mappings.

Proof. **(Proof of the Contraction Mapping Theorem)**: Fix $x_0 \in \mathcal{D}$ and let $\{x_n\} \subset \mathcal{D}$ be a sequence of vectors defined by the recursion:

$$x_{n+1} = \tilde{T}(x_n).$$

Then, using the triangle inequality we have:

$$\|x_n - x_{n+l}\| \leq \|x_n - x_{n+1}\| + \|x_{n+1} - x_{n+2}\| + \cdots + \|x_{n+l-1} - x_{n+l}\|. \tag{9.10}$$

Also,

$$\|x_n - x_{n+1}\| = \|\tilde{T}(x_{n-1}) - \tilde{T}(x_n)\| \leq a\|x_{n-1} - x_n\| \leq \cdots \leq a^n\|x_1 - x_0\|. \tag{9.11}$$

Combining (9.10) and (9.11) we obtain:

$$\begin{aligned} \|x_n - x_{n+l}\| &\leq \left(a^n + a^{n+1} + \cdots + a^{n+l-1}\right)\|x_1 - x_0\| \\ &= a^n\left(1 + a + \cdots + a^{l-1}\right)\|x_1 - x_0\| \\ &\leq \frac{a^n}{1-a}\|x_1 - x_0\|, \end{aligned} \tag{9.12}$$

which is independent of l. Clearly, (9.12) implies that $\forall \epsilon > 0$, $\exists N_\epsilon > 0$ such that

$$\|x_n - x_{n+l}\| < \epsilon, \quad \forall n \geq N_\epsilon \text{ and } \forall l \geq 1.$$

Equivalently,

$$\|x_n - x_m\| < \epsilon, \quad \forall n, m \geq N_\epsilon,$$

i.e., $\{x_n\}$ is a Cauchy sequence. Hence, $x_n \xrightarrow[n\to\infty]{} x^* \in \mathcal{D}$, since $\mathcal{D}$ is closed. By the continuity of $\tilde{T}$, we further have that

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \tilde{T}(x_{n-1}) = \tilde{T}(\lim_{n\to\infty} x_{n-1}) = \tilde{T}(x^*),$$

i.e., x^* is a fixed point of $\tilde{T}$.

To finish the proof, we need to show that x^* is the unique fixed point of $\tilde{T}$. Let $y^* \neq x^*$ be a fixed point of $\tilde{T}$, i.e., $y^* = \tilde{T}(y^*)$. Then,

$$\|y^* - x^*\| = \|\tilde{T}(y^*) - \tilde{T}(x^*)\| \leq a\|y^* - x^*\| < \|y^* - x^*\|,$$

where the last inequality is due to $a < 1$. This is a contradiction. Therefore, x^* is the unique fixed point of $\tilde{T}$. □

A few more words about the Contraction Mapping Theorem: The above theorem corresponds to the most basic fixed-point theorem in analysis. It appeared in **Banach's** Ph.D. thesis (1920, published in 1922). The theorem is stated in full generality for **complete metric spaces**.

Definition 9.2.11. *Let $(\mathcal{X}, d)$ be a metric space. A sequence $\{x_n\} \subset \mathcal{X}$ is said to be a **Cauchy sequence** if $\forall \epsilon > 0$ there exists $N_\epsilon > 0$ such that $d(x_n, x_m) < \epsilon$ for all $n, m \geq N_\epsilon$.*

Definition 9.2.12. *A metric space (X, d) is **complete**, if every Cauchy sequence converges in it.*

Examples: Examples of complete metric spaces are $\mathbb{R}^n$ with the (usual) Euclidean metric and all closed subsets of $\mathbb{R}^n$ with this metric. More generally, $\mathbb{R}^n$ with $\|\cdot\|_p$ for any $p \geq 1$ is a complete metric space (the corresponding metric is the induced one).

Definition 9.2.13. *A mapping $\tilde{T} : \mathcal{X} \to \mathcal{X}$, where $(\mathcal{X}, d)$ is a metric space, is a **contraction** if there exists $K < 1$ such that*

$$d(\tilde{T}(x), \tilde{T}(y)) \leq K d(x, y), \quad \forall x, y \in \mathcal{X}.$$

Theorem 9.2.14. ***Contraction Mapping Theorem (Metric space terminology)**: Let $\mathcal{X}$ be a complete metric space and $\tilde{T} : \mathcal{X} \to \mathcal{X}$ be a contraction. Then, $\tilde{T}$ has a unique fixed point and under the action of iterates of $\tilde{T}$, all points converge exponentially fast to this unique fixed point.*

Back to the proof of Theorem 9.2.1:

Define the operator:

$$T_\mu(J) = \bar{c}_\mu + \alpha P_\mu J.$$

Then, by (9.8) J_μ is the solution of the fixed point equation

$$J = T_\mu(J).$$

We observe that:

$$T_\mu(x) - T_\mu(y) = \alpha P_\mu (x - y).$$

Taking the maximum norm to both sides of this equation, we obtain:

$$\begin{aligned}
\|T_\mu(x) - T_\mu(y)\|_\infty &= \max_i |[T_\mu(x) - T_\mu(y)]_i| = \alpha \max_i |[P_\mu(x - y)]_i| \\
&= \alpha \max_i \left| \sum_j P_{ij}(x_j - y_j) \right| \\
&\leq \alpha \max_i \sum_j P_{ij} |x_j - y_j| \leq \alpha \|x - y\|_\infty.
\end{aligned}$$

Therefore, T_μ is a contraction mapping and by the previous theorem the fixed point equation $J = T_\mu(J)$ has a unique solution J_μ. We can iteratively compute J_μ by the iteration:

$$J^{(k+1)} = T_\mu(J^{(k)}) \quad \text{for an arbitrary } J^{(0)}.$$

□

9.3 Optimal Policy

Let

$$J_0^N(i) = \min_{\mu_0,\mu_1,\ldots,\mu_N} E\left[\sum_{k=0}^{N} \alpha^k c(x_k,u_k)\middle| x_0 = i\right]$$

be the optimal finite horizon cost, when the terminal cost is zero. Then:

$$\begin{aligned}
J_0^N(i) &= \min_{\mu_0,\mu_1,\ldots,\mu_N} E\left[c(x_0,u_0) + \sum_{k=1}^{N} \alpha^k c(x_k,u_k)\middle| x_0 = i\right] \\
&= \min_{\mu_0} E[c(i,\mu_0(i))|x_0 = i] + \min_{\mu_1,\ldots,\mu_N} E\left[\sum_{k=1}^{N} \alpha^k c(x_k,u_k)\middle| x_0 = i\right] \\
&= \min_{\mu_0} E[c(i,\mu_0(i))|x_0 = i] + \min_{\mu_1,\ldots,\mu_N} \alpha \sum_j P_{ij}(\mu_0(i)) E\left[\sum_{k=1}^{N} \alpha^{k-1} c(x_k,u_k)\middle| x_1 = j\right] \\
&= \min_{\mu_0} E[c(i,\mu_0(i))|x_0 = i] + \alpha \sum_j P_{ij}(\mu_0(i)) J_1^N(j).
\end{aligned}$$

Equivalently:

$$J_0^N(i) = \min_{u_0} E[c(x_0,u_0)|x_0 = i] + \alpha \sum_j P_{ij}(u_0) J_1^N(j)$$

Continuing recursively, we obtain:

$$J_k^N(i) = \min_{u_k} E[c(x_k,u_k)|x_k = i] + \alpha \sum_j P_{ij}(u_k) J_{k+1}^N(j). \tag{9.13}$$

This is called **Dynamic Programming (DP) equation** or **Bellman's equation**.

Interpretation: The optimal action at any stage k should be chosen to minimize the sum of the current cost and its impact on the future cost.

Infinite horizon case: We would like to generalize (9.13) to the infinite horizon case ($N \to \infty$). When $N \to \infty$, we hope that $J_k^*(i) = J_{k+1}^*(i)$ for any i, i.e., the optimal remaining cost, also called *cost-to-go*, does not depend on k, since both $J_k^*(i)$ and $J_{k+1}^*(i)$ correspond to infinite steps.

Theorem 9.3.1. *Let*

$$J^*(i) = \inf_{\mu_0,\mu_1,\ldots} \lim_{N\to\infty} E\left[\sum_{k=0}^{N} \alpha^k c(x_k,u_k)\middle| x_0 = i\right].$$

Then, J^* satisfies:

$$\begin{aligned}
J^*(i) &= \min_u E\left[c(x_0,u_0) + \alpha J^*(x_1)\middle| x_0 = i, u_0 = u\right] \\
&= \min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) J^*(j).
\end{aligned} \tag{9.14}$$

*This is called **infinite-horizon DP** or **Bellman's equation** for discounted cost MDPs.*

Moreover, (9.14) can be given in the form of a fixed point equation as the following theorem shows:

Theorem 9.3.2. *Let* $T(J)(i) = \min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u)J(j)$. *With this definition, (9.14) can be written as:*

$$J^* = T(J^*).$$

Then, T *is a contraction mapping with parameter* $\alpha \in [0,1)$.

Proof. The proof of this theorem relies on the following two facts, which can be easily proved:

1. If $J_1 \leq J_2$ elementwise, then $T(J_1) \leq T(J_2)$, i.e., T is a monotone operator.
2. $T(J + r1) = T(J) + \alpha r1$ for any scalar $r \in \mathbb{R}$, where 1 is the all-ones vector.

Let $J_1 \neq J_2$ be two arbitrary vectors and define $r = \|J_1 - J_2\|_\infty$. Then,

$$J_2 - r1 \leq J_1 \leq J_2 + r1$$

and by using the aforementioned properties:

$$T(J_2 - r1) \leq T(J_1) \leq T(J_2 + r1)$$

or

$$T(J_2) - \alpha r1 \leq T(J_1) \leq T(J_2) + \alpha r1,$$

which leads to

$$\|T(J_1) - T(J_2)\|_\infty \leq \alpha \|J_1 - J_2\|_\infty.$$

Therefore, T is a contraction mapping with parameter α with respect to the ℓ_∞-norm. □

Conclusion: $J^* = T(J^*)$ has a unique solution by the Contraction Mapping Theorem.

Finally, the following straightforward theorem completes our discussion on the optimal policy of infinite horizon discounted cost MDPs:

Theorem 9.3.3. *Let* $\mu^*(i)$ *be the solution to the problem*

$$\min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u)J^*(j)$$

for any i. *Then,* μ^* *is an optimal policy for the infinite horizon discounted cost MDP. In addition,* μ^* *is deterministic by construction (and clearly stationary).*

Proof. Note that $J_{\mu^*} = T_{\mu^*}(J_{\mu^*})$, i.e., $J_{\mu*}$ is the unique fixed point of T_{μ^*}. Furthermore, for any $i \in \mathcal{X}$

$$T_{\mu^*}(J^*)(i) = \bar{c}(i, \mu^*(i)) + \alpha \sum_j P_{ij}(\mu^*(i))J^*(j) = J^*(i).$$

Hence, J^* is a fixed point of T_{μ^*} and since T_{μ^*} has a unique fixed point we conclude that

$$J_{\mu^*} = J^*.$$

□

9.4 Numerical Methods

In this section, we outline two classical methods for solving infinite horizon discounted cost MDPs.

9.4.1 Value Iteration

The value iteration algorithm is a successive approximation algorithm to compute the value function J^*. Since T is a contraction mapping, one can in principle compute J^* by guessing $\hat{J}_0$ and by implementing the recursion

$$\hat{J}_{k+1} = T(\hat{J}_k)$$

up to convergence. Once J^* has been computed, the optimal policy can be found by solving

$$\min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) J^*(j)$$

for any i.

Practical Implementation: In practice, we will stop the iteration at some k, obtaining $\hat{J}_k$ as our final approximation to J^*.

1. To evaluate how far $\hat{J}_k$ is from J^*, we note that

   $$\|\hat{J}_k - J^*\|_\infty = \|T(\hat{J}_{k-1}) - T(J^*)\| \leq \alpha \|\hat{J}_{k-1} - J^*\|_\infty \leq \cdots \leq \alpha^k \|\hat{J}_0 - J^*\|_\infty.$$

   This bound is only useful if we can estimate $\|\hat{J}_0 - J^*\|_\infty$.

   Alternatively, using similar steps as in the proof of the Contraction Mapping Theorem, we have:

   $$\|\hat{J}_k - \hat{J}_{k+1}\|_\infty = \|T(\hat{J}_{k-1}) - T(\hat{J}_k)\|_\infty \leq \alpha \|\hat{J}_{k-1} - \hat{J}_k\|_\infty \leq \cdots \leq \alpha^k \|\hat{J}_1 - \hat{J}_0\|_\infty. \tag{9.15}$$

   Moreover,

   $$\begin{aligned}
   \|\hat{J}_k - J^*\|_\infty &\leq \|\hat{J}_k - \hat{J}_{k+1}\|_\infty + \|\hat{J}_{k+1} - \hat{J}_{k+2}\|_\infty + \cdots + \|\hat{J}_{k+l-1} - \hat{J}_{k+l}\|_\infty + \|\hat{J}_{k+l} - J^*\|_\infty \\
   &\leq \alpha^k \|\hat{J}_1 - \hat{J}_0\|_\infty \left(1 + \alpha + \cdots + \alpha^{l-1}\right) + \|\hat{J}_{k+l} - J^*\|_\infty \\
   &\leq \alpha^k \|\hat{J}_1 - \hat{J}_0\|_\infty \sum_{m=0}^{\infty} \alpha^m + \|\hat{J}_{k+l} - J^*\|_\infty \\
   &\leq \frac{\alpha^k}{1-\alpha} \|\hat{J}_1 - \hat{J}_0\|_\infty + \|\hat{J}_{k+l} - J^*\|_\infty,
   \end{aligned}$$

   which holds for any l. Letting $l \to \infty$ and using the fact that $\hat{J}_{k+l} \to J^*$, we finally obtain the estimate:

   $$\|\hat{J}_k - J^*\| \leq \frac{\alpha^k}{1-\alpha} \|\hat{J}_1 - \hat{J}_0\|_\infty. \tag{9.16}$$

   This bound is more useful because $\|\hat{J}_1 - \hat{J}_0\|_\infty$ can be easily computed.

2. Having $\hat{J}_k$, we will use it to solve for the "optimal" policy. Let the obtained policy be μ_k, which is calculated by solving:

   $$\underbrace{\min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) \hat{J}_k(j)}_{=T(\hat{J}_k)(i) \text{ and also } =T_{\mu_k}(\hat{J}_k)(i)}. \tag{9.17}$$

   We now want to bound the performance of μ_k, which corresponds to J_{μ_k}. Recall that J_{μ_k} is the unique solution of the fixed point equation $J = T_{\mu_k}(J)$, since T_{μ_k} is a contraction mapping. By the triangle inequality we have:

$$\|J_{\mu_k} - J^*\|_\infty \leq \|J_{\mu_k} - \hat{J}_k\|_\infty + \|\hat{J}_k - J^*\|_\infty. \tag{9.18}$$

For the first term $||J_{\mu_k} - \hat{J}_k||_\infty$, we apply again the triangle inequality:

$$\begin{aligned}\|J_{\mu_k} - \hat{J}_k\|_\infty &\leq \|J_{\mu_k} - T_{\mu_k}(\hat{J}_k)\|_\infty + \|T_{\mu_k}(\hat{J}_k) - \hat{J}_k\|_\infty \\ &= \|T_{\mu_k}(J_{\mu_k}) - T_{\mu_k}(\hat{J}_k)\|_\infty + \|T_{\mu_k}(\hat{J}_k) - \hat{J}_k\|_\infty \\ &\leq \alpha\|J_{\mu_k} - \hat{J}_k\|_\infty + \|T_{\mu_k}(\hat{J}_k) - \hat{J}_k\|_\infty,\end{aligned}$$

or

$$\|J_{\mu_k} - \hat{J}_k\|_\infty \leq \frac{1}{1-\alpha}\|T_{\mu_k}(\hat{J}_k) - \hat{J}_k\|_\infty = \frac{1}{1-\alpha}\|T(\hat{J}_k) - \hat{J}_k\|_\infty. \tag{9.19}$$

For the second term $\|\hat{J}_k - J^*\|_\infty$ we have:

$$\begin{aligned}\|\hat{J}_k - J^*\|_\infty &\leq \|\hat{J}_k - T(\hat{J}_k)\|_\infty + \|T(\hat{J}_k) - J^*\|_\infty \\ &= \|\hat{J}_k - T(\hat{J}_k)\|_\infty + \|T(\hat{J}_k) - T(J^*)\|_\infty \\ &\leq \|\hat{J}_k - T(\hat{J}_k)\|_\infty + \alpha\|\hat{J}_k - J^*\|_\infty,\end{aligned}$$

or

$$\|\hat{J}_k - J^*\|_\infty \leq \frac{1}{1-\alpha}\|\hat{J}_k - T(\hat{J}_k)\|_\infty = \frac{1}{1-\alpha}\|\hat{J}_k - T_{\mu_k}(\hat{J}_k)\|_\infty. \tag{9.20}$$

By (9.18)-(9.20) and by (9.15) we obtain the estimate:

$$||J_{\mu_k} - J^*||_\infty \leq \frac{2}{1-\alpha}||\hat{J}_k - T(\hat{J}_k)||_\infty = \frac{2}{1-\alpha}||\hat{J}_k - \hat{J}_{k+1}||_\infty \leq \frac{2\alpha^k}{1-\alpha}||\hat{J}_1 - \hat{J}_0||_\infty. \tag{9.21}$$

9.4.2 Policy Iteration

In Policy Iteration we take a different approach by starting with any stationary policy μ_0 at time $k = 0$. We obtain J_{μ_0} by solving $J_{\mu_0} = T_{\mu_0}(J_{\mu_0})$. Then, we perform the following algorithm:

Algorithm 1: Policy Iteration

Initialization: μ_0, J_{μ_0};
for $k = 0, 1, 2, ...$ **do**
  for $i \in \mathcal{X}$ **do**
    $\mu_{k+1}(i) \leftarrow \arg\min_u \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) J_{\mu_k}(j)$;
  end
  get $J_{\mu_{k+1}}$ by solving $J_{\mu_{k+1}} = T_{\mu_{k+1}}(J_{\mu_{k+1}})$;
  if $J_{\mu_{k+1}} = J_{\mu_k}$ **then**
    stop;
  end
end

Notice that at each stage, μ_{k+1} is set to be the "optimal" policy using J_{μ_k}. Therefore, the updated cost function $J_{\mu_{k+1}}$ must satisfy $J_{\mu_{k+1}} \leq J_{\mu_k}$. When these two vectors are equal, we claim that $J_{\mu_{k+1}} = J^*$ and thus $\mu_{k+1} = \mu^*$.

Lemma 9.4.1. *$J_{\mu_{k+1}} \leq J_{\mu_k}$, i.e., the cost improves at every iteration.*

Proof. We note that:

$$\begin{aligned} T_{\mu_{k+1}}(J_{\mu_k})(i) &= \bar{c}(i, \mu_{k+1}(i)) + \alpha \sum_j P_{ij}(\mu_{k+1}(i)) J_{\mu_k}(j) \\ &\leq \bar{c}(i, \mu_k(i)) + \alpha \sum_j P_{ij}(\mu_k(i)) J_{\mu_k}(j) \\ &= T_{\mu_k}(J_{\mu_k}) = J_{\mu_k}, \end{aligned}$$

where the last equality is due to the fact that J_{μ_k} is the unique fixed point of T_{μ_k}. Relying on the monotonicity of the operator T_μ for any μ, we obtain[3]:

$$T^{(n)}_{\mu_{k+1}}(J_{\mu_k}) = \underbrace{T_{\mu_{k+1}}(T_{\mu_{k+1}}(\cdots T_{\mu_{k+1}}(J_{\mu_k})\cdots))}_{n \text{ times}} \leq J_{\mu_k}.$$

Since this inequality is valid for any $n \geq 1$, letting $n \to \infty$ we obtain:

$$T^{(n)}_{\mu_{k+1}}(J_{\mu_k}) \xrightarrow[n\to\infty]{} J_{\mu_{k+1}} \leq J_{\mu_k}.$$

□

Lemma 9.4.2. *$J_{\mu_{k+1}} = J_{\mu_k}$ implies that $\mu_{k+1} = \mu^*$.*

Proof. By the definition of μ_{k+1}:

$$T_{\mu_{k+1}}(J_{\mu_k}) = T(J_{\mu_k})$$

or

$$T_{\mu_{k+1}}(J_{\mu_{k+1}}) = T(J_{\mu_{k+1}})$$

since $J_{\mu_{k+1}} = J_{\mu_k}$ and therefore:

$$J_{\mu_{k+1}} = T(J_{\mu_{k+1}}). \tag{9.22}$$

Here, we use the fact that $J_{\mu_{k+1}}$ is the unique fixed point of $T_{\mu_{k+1}}$. (9.22) shows that when $J_{\mu_{k+1}} = J_{\mu_k}$, then $J_{\mu_{k+1}}$ is the unique fixed point of T. Therefore, $J_{\mu_{k+1}} = J_{\mu_k} = J^*$ and hence $\mu_{k+1} = \mu^*$.

□

Final Remark: Although Policy Iteration converges in finite steps because the number of possible policies is finite for finite state and action spaces, solving $J_{\mu_k} = T_{\mu_k}(J_{\mu_k})$ at every iteration can be costly.

9.5 Monotone Policies for MDPs

For large state spaces, computing the optimal policy by solving the DP equation can be very expensive.Thus, we seek for sufficient conditions which, if satisfied by an MDP model, can greatly simplify computation. As an example, there

[3]The monotonicity of T_μ can be shown similarly to the monotonicity of T.

are conditions which, if satisfied, lead to the existence of an optimal policy μ^* which is increasing or decreasing in the state value. Such policies are called *monotone*.

Example: Let $\mathcal{X} = \{1, 2, \ldots, X\}$ and $\mathcal{U} = \{1, 2\}$. If $\mu^*(x)$ is nondecreasing in x, then it is a step function of the form:

$$\mu^*(x) = \begin{cases} 1, x < x^* \\ 2, x \geq x^* \end{cases} \quad \text{for some state } x^* \in \mathcal{X}. \tag{9.23}$$

Such policies are sometimes called **threshold policies** (x^* is the *threshold state*) or **control limit policies** (x^* is a *control limit*).

Compared to solving the DP equation, computing x^* can be very cheap. Also, when implementing the controller we only need to store x^* instead of a lookup table.

9.5.1 Bellman's Equation in terms of the Q function

For infinite horizon discounted cost MDPs, we rewrite Bellman's equation as follows:

$$\begin{aligned} J^*(i) &= \min_u Q(i,u) \quad \text{and} \quad \mu^*(i) = \arg\min_u Q(i,u), \quad \text{where} \\ Q(i,u) &= \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) J^*(j) = \bar{c}(i,u) + \alpha \sum_j P_{ij}(u) \min_v Q(j,v). \end{aligned} \tag{9.24}$$

We want to find some sufficient conditions on the model to ensure that the optimal policy $\mu^*(x)$ is monotone.

9.5.2 Submodularity and Supermodularity

Definition 9.5.1. *A real-valued function $f(i,u)$ is submodular in (i,u) if:*

$$f(i,u+1) - f(i,u) \geq f(i+1,u+1) - f(i+1,u),$$

that is, if $f(i,u+1) - f(i,u)$ is decreasing in i.

Similarly, f is supermodular if $-f(i,u)$ is submodular, that is, if:

$$f(i,u+1) - f(i,u) \leq f(i+1,u+1) - f(i+1,u).$$

Note: Submodularity and supermodularity treat i and u symmetrically. For example, an equivalent definition of submodularity is that $f(i+1,u) - f(i,u)$ is decreasing with respect to u.

Proof. Let $f(i+1,u) - f(i,u)$ be decreasing in u:

$$\begin{aligned} f(i+1,u+1) - f(i,u+1) &\leq f(i+1,u) - f(i,u) \\ \text{or} \quad f(i+1,u+1) - f(i+1,u) &\leq f(i,u+1) - f(i,u). \end{aligned}$$

That is, $f(i,u+1) - f(i,u)$ is decreasing in i, which means that f is submodular in (i,u) according to the previous definition. □

Remark: Submodularity is often called **subadditivity** and supermodularity is called **superadditivity**. A corresponding slightly more general definition (in terms of notation) is:

Definition 9.5.2. *Let $\mathcal{X}, \mathcal{Y}$ be partially-ordered sets. A real-valued function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is said to be* subadditive *(or submodular) if for all $x^- \leq x^+$ in $\mathcal{X}$ and all $y^- \leq y^+$ in $\mathcal{Y}$*

$$f(x^-, y^-) + f(x^+, y^+) \leq f(x^-, y^+) + f(x^+, y^-). \tag{9.25}$$

f is said to be superadditive *(or supermodular) if the reverse inequality holds.*

The defining inequality (9.25) is called **quadrangle inequality**.

Equivalence of definitions: Apply the last definition for $x^- = i, x^+ = i + 1, y^- = u, y^+ = u + 1$ to obtain the first definition of submodularity and supermodularity.

Properties:

- Sums of submodular or supermodular functions are also submodular or supermodular, respectively.
- Pointwise limits of submodular or supermodular functions are also submodular or supermodular, respectively.

Note: If $(i, u) \in \mathbb{R} \times \mathbb{R}$ and f is twice continuously differentiable, then $f(i, u)$ is supermodular if $\dfrac{\partial^2 f}{\partial i \partial u} \geq 0$. In other words, $f(i, u)$ is supermodular if f has nonnegative mixed derivatives.

Examples:

- $f(i, u) = -iu$ is submodular in (i, u).
- $f(i, u) = \max\{i, u\}$ is submodular in (i, u).
- Any separable function $h(i, u) = f(i) + g(u)$ is both submodular and supermodular. Functions that are both submodular and supermodular are called **modular** or **additive**.

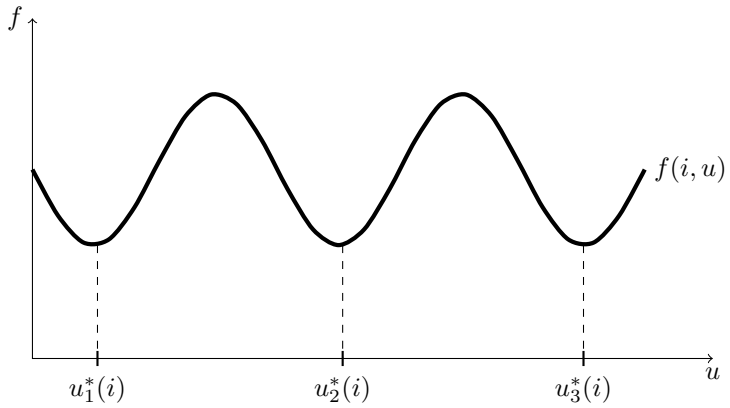
9.5.3 Topkis' Monotonocity Theorem

The following result introduced by Donald M. Topkis in 1978 reveals the connection of submodular and supermodular functions with monotone policies:

Theorem 9.5.3. *Let $\mathcal{U}^*(i)$ be the set of possible minimizers of $f(i, u) : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ with respect to u, i.e., $\mathcal{U}^*(i) = \arg\min_u f(i, u)$. Let also $u^*_{\max}(i), u^*_{\min}(i)$ be the maximum and minimum elements in $\mathcal{U}^*(i)$. Then:*

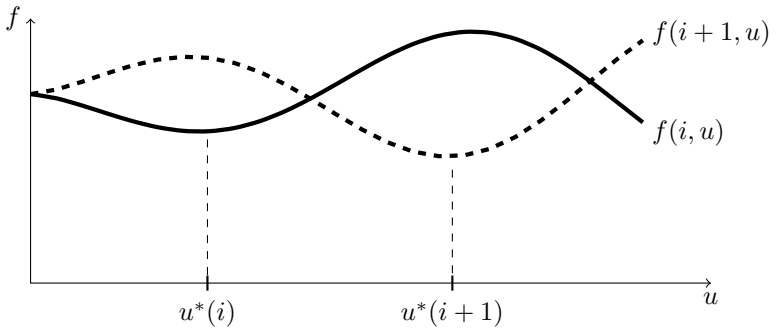
1. *If f is submodular in (i, u), then $u^*_{\max}(i), u^*_{\min}(i)$ are nondecreasing in i.*
2. *If f is supermodular in (i, u), then $u^*_{\max}(i), u^*_{\min}(i)$ are nonincreasing in i.*

The following picture illustrates $\mathcal{U}^*(i), u^*_{\max}(i), u^*_{\min}(i)$.



$\mathcal{U}^*(i) = \{u_1^*(i), u_2^*(i), u_3^*(i)\}$, $u_{\max}^* = u_3^*(i)$, $u_{\min}^* = u_1^*(i)$.

The following picture illustrates the theorem when $f(i,u)$ is submodular.



Proof. Here we only give the proof for submodular functions in the case where $\mathcal{U}^*(i)$ is a singleton for all $i \in \mathcal{X}$, i.e. $u_{\max}^*(i) = u_{\min}^*(i)$. By the definition of submodularity (or subadditivity), we have

$$f(i,u) - f(i,\bar{u}) \geq f(i+1,u) - f(i+1,\bar{u}) \quad \forall u \geq \bar{u}.$$

Letting $\bar{u} = u^*(i+1)$ we have:

$$f(i,u) - f(i,u^*(i+1)) \geq f(i+1,u) - f(i+1,u^*(i+1)) \quad \forall u \geq u^*(i+1).$$

By the definition of $u^*(i+1)$, we know that $f(i+1,u) - f(i+1,u^*(i+1)) \geq 0,\ \forall u \in \mathcal{U}$, thus we have

$$f(i,u) - f(i,u^*(i+1)) \geq 0 \quad \forall u \geq u^*(i+1).$$

This already implies that $u^*(i) \leq u^*(i+1)$ and concludes the proof. □

9.5.4 Stochastic Dominance

In order to apply the monotonicity theorem to MDPs, we need the notion of stochastic dominance.

Definition 9.5.4. *Let p_1, p_2 be two pdfs or alternatively P_1, P_2 be two pmfs with associated distribution functions F_1, F_2, respectively. We say that* p_1 first-order stochastically dominates p_2 *or* P_1 first-order stochastically dominates P_2 *and we write $p_2 \leq_{st} p_1$ or $P_2 \leq_{st} P_1$ if*

$$F_2^c(x) \leq F_1^c(x), \quad \forall x \in \mathbb{R} \quad \textit{or} \quad 1 - F_2(x) \leq 1 - F_1(x), \quad \forall x \in \mathbb{R} \quad \textit{or} \quad F_2(x) \geq F_1(x), \quad \forall x \in \mathbb{R}.$$

Equivalently, let $X \sim F_1$ and $Y \sim F_2$. Then, $Y \leq_{st} X$ if

$$P(Y > x) \leq P(X > x), \quad \forall x \in \mathbb{R}.$$

Intuition: Stochastically, X tends to be larger than Y.

In the case of pmfs, we have that $P_2 \leq_{st} P_1$ is equivalent to $\sum_{i \geq j} P_2(i) \leq \sum_{i \geq j} P_1(i),\ \forall j$.

Lemma 9.5.5. $Y \leq_{st} X \Rightarrow E[Y] \leq E[X]$.

The next theorem gives an equivalent characterization of stochastic dominance:

Theorem 9.5.6. *Let $\bar{D}$ be the set of X-dimensional vectors d with increasing elements, i.e.,*

$$\bar{D} = \left\{ d = (d_1, d_2, \ldots, d_X) \in \mathbb{R}^X : d_1 \leq d_2 \cdots \leq d_X \right\}.$$

Then, $P_2 \leq_{st} P_1$ if and only if

$$d^T P_2 \leq d^T P_1, \quad \forall d \in \bar{D}.$$

Alternatively: Let $\mathcal{X} = \{1, 2, \ldots, X\}$, $P_1(i) = P(X = i)$ and $P_2(i) = P(Y = i)$ for every $i \in \mathcal{X}$. Then, $P_2 \leq_{st} P_1$ or $Y \leq_{st} X$ if and only if $E[f(Y)] \leq E[f(X)]$ for any nondecreasing function $f : \mathcal{X} \to \mathbb{R}$.

9.5.5 Monotone Optimal Policies for MDPs

With the previous definitions we are now ready to state the following theorem about monotone policies for MDPs:

Theorem 9.5.7. *Assume that the following conditions are satisfied by an infinite horizon discounted cost MDP model:*

1. *$c(i, u)$ is nonincreasing in i, $\forall u$,*
2. *$c(i, u)$ is submodular in (i, u), that is $c(i, u + 1) - c(i, u)$ is nonincreasing in i, $\forall u$,*
3. *$P_i(u) \leq_{st} P_{i+1}(u),\ \forall (i, u)$, where $P_i(u)$ is the ith row of the transition matrix $P(u)$,*
4. *$P_{ij}(u)$ is "tail-sum supermodular" in (i, u), i.e., $\sum_{j \geq l} P_{ij}(u)$ is supermodular in (i, u) for any l or equivalently $\sum_{j \geq l} (P_{ij}(u + 1) - P_{ij}(u))$ is nondecreasing in i, $\forall l$.*

Then, there exists an optimal stationary policy $\mu^*(i)$, which is nondecreasing in i.

Clarification: If μ^* is not the unique optimal policy, then there exists an optimal stationary policy μ^* that is nondecreasing in $i \in \mathcal{X}$.

Proof Sketch: First, it can be shown that $J^*(i)$ is nonincreasing in i. Moreover, recall that

$$Q(i, u) = \bar{c}(i, u) + \alpha \sum_j P_{ij}(u) J^*(j)$$

and $J^*(i) = \min_u Q(i, u)$. Then, it can be shown that $Q(i, u)$ is submodular in (i, u). Applying Topkis' theorem, we conclude that $\mu^*(i)$ is nondecreasing in i.

Remark: Similar structural results hold for finite-horizon and average cost MDPs.

A A bit of background material: Some Other Common Fixed Point Theorems

The following fixed point theorems are often used in engineering problems:

Theorem A.1. ***Brouwer's Fixed Point Theorem****: If $\mathcal{D}$ is a nonempty, compact, convex subset of $\mathbb{R}^n$ and $g : \mathcal{D} \to \mathcal{D}$ is continuous, then g has a fixed point.*

Importance: Brouwer's fixed point theorem is the basis for general fixed point theorems in functional analysis.

Kakutani's Fixed Point Theorem extends Brouwer's Theorem to set-valued functions.

Definition A.2. *Let $P(\mathcal{D})$ denote all nonempty, closed, convex subsets of $\mathcal{D}$, where $\mathcal{D}$ is a subset of a Euclidean space. If $\mathcal{S}$ is nonempty, compact, and convex, then the set-valued function $\Phi : \mathcal{S} \to P(\mathcal{S})$ is* upper semi-continuous *if for arbitrary sequences $\{x_n\}, \{y_n\} \subset \mathcal{S}$ we have that:*

$$x_0 = \lim_{n\to\infty} x_n, \;\; y_0 = \lim_{n\to\infty} y_n \;\; \textit{and} \;\; y_n \in \Phi(x_n), \forall n \in \mathbb{N} \Longrightarrow y_0 \in \Phi(x_0).$$

Definition A.3. *A fixed point of a set-valued function $\Phi : \mathcal{S} \to P(\mathcal{S})$ is a point $x^* \in \mathcal{S}$ such that $x^* \in \Phi(x^*)$.*

Theorem A.4. ***Kakutani's Fixed Point Theorem****: If $\mathcal{S}$ is a nonempty, compact, convex set in a Euclidean space and $\Phi : \mathcal{S} \to P(\mathcal{S})$ is upper semi-continuous, then Φ has a fixed point.*

Importance: Kakutani's Fixed Point Theorem is a key result in proving the existence of a *Nash equilibrium* in strategic games.

B A bit more material on Supermodularity and Submodularity

Motivation: A class of interesting games are the so called **supermodular games**. Such games are characterized by "strategic complementarities". Roughly speaking, this means that when one player takes a higher action, the others want to do the same. Supermodular games are analytically appealing. They have nice comparative statistics properties and behave well under various learning rules. The aforementioned "complementarities" are expressed both via constraints and payoff functions. Mathematically, they are captured by *lattices* and *supermodular* payoff functions.

Definition B.1. *Let $x, y \in \mathbb{R}^n$ and define their **join** as*

$$x \vee y = [\max\{x_1, y_1\}, \ldots, \max\{x_n, y_n\}]^T$$

*and their **meet** as*

$$x \wedge y = [\min\{x_1, y_1\}, \ldots, \min\{x_n, y_n\}]^T.$$

If $x \leq y$ elementwise, then $x \vee y = y$ and $x \wedge y = x$.

Definition B.2. *A set $\mathcal{L} \subseteq \mathbb{R}^n$ is called a **lattice** if for any $x, y \in \mathcal{L}$ we have $x \vee y \in \mathcal{L}$ and $x \wedge y \in \mathcal{L}$.*

Note: More generally, a partially-ordered set $(\mathcal{L}, \geq)$ is said to be *lattice* if for any $x, y \in \mathcal{L}$ we have $x \vee y \in \mathcal{L}$ and $x \wedge y \in \mathcal{L}$, where the join and meet operators are defined based on the partial order $\geq$ as follows:

$$x \vee y = \inf\{z \in \mathcal{L} : x \leq z \text{ and } y \leq z\},$$
$$x \wedge y = \sup\{z \in \mathcal{L} : x \geq z \text{ and } y \geq z\}.$$

Examples:

1. Let $\mathcal{L}$ be the powerset of a set $\mathcal{D}$ and for $\mathcal{A}, \mathcal{B} \in \mathcal{L}$ let $\mathcal{A} \geq \mathcal{B}$ be equivalent to $\mathcal{A} \supseteq \mathcal{B}$. Then, $A \vee B = A \cup B \in \mathcal{L}$, $A \wedge B = A \cap B \in \mathcal{L}$ and therefore $(\mathcal{L}, \supseteq)$ is a lattice.
2. According to the previous definitions, $(\mathbb{R}^n, \geq)$, where $\geq$ corresponds to the elementwise ordering, is a lattice.

Definition B.3. *Given any lattice $(\mathcal{L}, \geq)$ a function $f : \mathcal{L} \to \mathbb{R}$ is said to be **supermodular** if for any $x, y \in \mathcal{L}$:*

$$f(x \vee y) + f(x \wedge y) \geq f(x) + f(y).$$

f is submodular if $-f$ is supermodular.

To recover (9.25), consider the elementwise ordering and let in the above definition

$$x = \begin{bmatrix} x^- \\ y^+ \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} x^+ \\ y^- \end{bmatrix}.$$

Note: Let $\mathcal{S}$ be a finite set and consider a function $f : 2^{\mathcal{S}} \to \mathbb{R}$, where $2^{\mathcal{S}}$ is the powerset of $\mathcal{S}$. For $(\mathcal{L} = 2^{\mathcal{S}}, \supseteq)$, the above defining inequality for supermodularity takes the following form often appearing in the literature:

$$f(A \cup B) + f(A \cap B) \geq f(A) + f(B), \quad \forall A, B \subseteq \mathcal{S} \text{ (or } A, B \in \mathcal{L}).$$

We now give a few more results about supermodular functions that may be useful in various contexts.

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is both supermodular and submodular if and only if f is separable, i.e., there exist functions $f_i : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \sum_{i=1}^n f_i(x_i), \forall x \in \mathbb{R}^n$.
- $f(x, z) = x^T z$ is supermodular on the product space $\mathbb{R}^n \times \mathbb{R}^n$.
- $f(x, z) = -\|x - z\|_p^p$ is supermodular on $(x, z) \in \mathbb{R}^{2n}$ for any $p \geq 1$.
- If $f_i(z)$ is increasing (decreasing) on $\mathbb{R}$ for $i = 1, 2, \ldots, n$, then $f(x) = \min_{1 \leq i \leq n} f_i(x_i)$ is supermodular on $\mathbb{R}^n$.
- Any positive linear combination of supermodular functions is supermodular.
- Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and suppose that $f(\cdot, y)$ is supermodular for any $y \in \mathbb{R}^m$. Then, $F(x) = E_\xi[f(x, \xi)]$ is supermodular, provided that it is well-defined.
- Let $\mathcal{L}$ be a lattice in $\mathbb{R}^n$. For $f : \mathbb{R} \to \mathbb{R}$ define $g : \mathbb{R}^n \to \mathbb{R}$ as $g(x) = f\left(\sum_{i=1}^n a_i x_i\right)$ for some $a_i \in \mathbb{R}, i = 1, 2, \ldots, n$. Suppose that $a_i \geq 0, \forall i$ and f is convex. Then, g is supermodular on $\mathcal{L}$.

B.1 Submodularity and Discrete Optimization

Motivation: Maximize or minimize an objective function $f : 2^{\mathcal{S}} \to \mathbb{R}$, where $\mathcal{S}$ is a finite set with cardinality n. Because $\mathcal{S}$ is finite, a function $f : 2^{\mathcal{S}} \to \mathbb{R}$ can be equivalently regarded as a function on the hypercube, i.e., $f : \{0, 1\}^n \to \mathbb{R}$. We further note that submodular functions are functions assigning values to all subsets of a finite set $\mathcal{S}$.

Analogy between Concavity and Submodularity: Consider $f : \mathbb{R} \to \mathbb{R}$. Then, f is concave if $f'(x)$ is nonincreasing. Let now $f : \{0,1\}^n \to \mathbb{R}$. f is submodular if for every i, the discrete derivative $\partial_i f(x) = f(x + e_i) - f(x)$ is nonincreasing. Here, e_i is a vector with 1 at the ith location and zeros elsewhere.

Where submodular functions appear:

- **Combinatorial Optimization**: Rank functions of matroids, polymatroids, submodular flows, submodular minimization.
- **Game Theory**: Submodular functions model *valuation functions* of agents with *diminishing returns*. They often appear in problems like combinatorial auctions and cost sharing and marketing on social networks.
- **Machine Learning**: Submodular functions appear as objective functions of machine learning tasks such as sensor placement, active learning, etc.

  A comment: Although submodularity looks like a discrete analogue of concavity, submodularity is more useful for *minimization* rather than maximization. Moreover, problems involving maximization of submodular functions are typically NP-hard! An example is the problem of finding a maximum cut in a graph, also known as Max Cut problem. In this problem, one tries to find a vertex subset $\mathcal{S}$ such that the number of edges between $\mathcal{S}$ and the complement $\mathcal{S}^c$ is maximal.

  The following result implies that submodularity is a very useful structure to have in a discrete optimization problem:

  Theorem B.4. *Let $\mathcal{S}$ be a finite set with cardinality n and $f : 2^{\mathcal{S}} \to \mathbb{R}$ be a submodular function. Then, there is an algorithm that computes the minimum of f, i.e., solves the problem $\min_{A \subseteq \mathcal{S}} f(A)$ in poly(n) time using queries of the form $f(A) =?$.*

  Remark: The mentioned "algorithm" in the above theorem is based on the *ellipsoid method*. More efficient combinatorial algorihms have been found more recently.

  Combinatorial algorithms are usually very sophisticated. Nevertheless, for submodular objective functions there is a simple explanation on why it is possible to minimize them:

  Lovász Extension: Let $f : \{0,1\}^n \to \mathbb{R}$ be a submodular function. Consider the function:

  $$\tilde{f}(x) = E_{c \sim \mathrm{Unif}[0,1]} \left[f(\{i : x_i > c\}) \right].$$

  Clearly, $\tilde{f} : [0,1]^n \to \mathbb{R}$ is a continuous extension of the submodular f to the domain $[0,1]^n$. Then, $\tilde{f}$ is convex and therefore, it can be minimized efficiently. Moreover, a minimizer of $\tilde{f}$ can be converted into a minimizer of f.