Columbia Business School

Lecture 2: Infinite Horizon and Indefinite Horizon MDPs

Last time:

- RL overview and motivation
- Finite Horizon MDPs: formulation and the DP algorithm

Today:

- Infinite horizon discounted MDPs
- Basic theory of Bellman operators; contraction mappings; existence of optimal policies;
- Analogous theory for indefinite horizon (episodic) MDPs.

Warmup: Finite Horizon Discounted MDPs

A special case of last time

- Finite state and control spaces.
- Periods $0, 1, \ldots N$ with controls $u_0, \ldots, u_{N-1}$.
- Stationary transition probabilities $f_k(x, u, w) = f(x, u, w)$ for all $k \in \{0, \ldots, N-1\}$.
- Stationary control spaces: $U_k(x) = U(x)$ for all $k \in \{0, \ldots, N-1\}$.
- Discounted costs: $g_k(x, u, w) = \gamma^k g(x, u, w)$ for $k \in \{0, \ldots, N-1\}$
- Special terminal costs: $g_N(x) = \gamma^N c(x)$.

Warmup: Finite Horizon Discounted MDPs

A policy $\pi = (\mu_0, \ldots, \mu_{N-1})$ is a sequence of mappings where $\mu_k(x) \in U(x)$ for all $x \in \mathcal{X}$..

The expected cumulative "cost-to-go" of a policy π from starting state x is

$$J_\pi(x) = \mathbb{E}\left[\sum_{k=0}^{N-1} \gamma^k g(x_k, \mu_k(x_k), w_k) + \gamma^N c(x_N)\right]$$

where the expectation is over the i.i.d disturbances $w_0, \ldots, w_{N-1}$.

The optimal expected cost to go is

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x) \qquad \forall x \in \mathcal{X}$$

The Dynamic Programming Algorithm

Set

$$J_N^*(x) = c(x) \qquad \forall x \in \mathcal{X}$$

For $k = N-1, N-2, \ldots 0$, set

$$J_k^*(x) = \min_{u \in U(x)} \mathbb{E}[g(x,u,w) + \gamma J_{k+1}^*(f(x,u,w))] \quad \forall x \in \mathcal{X}.$$

Main Proposition from last time

For all initial states $x \in \mathcal{X}$, the optimal cost to go is $J^*(x) = J_0^*(x)$. This is attained by a policy $\pi^* = (\mu_0^*, ..., \mu_{N-1}^*)$ where for all $k \in \{0, \ldots, N-1\}$, $x \in \mathcal{X}$

$$\mu_k^*(x) \in \arg\min_{u \in U(x)} \mathbb{E}[g(x,u,w) + \gamma J_{k+1}^*(f(x,u,w))].$$

The DP Algorithm for policy evaluation

How to find the cost-to-go for any policy $\pi = (\mu_0, \ldots, \mu_{N-1})$?

- $J_\pi(x) = J_0(x)$ where J_0 is output by the following iterative algorithm.

$$J_N(x) = c(x) \qquad \forall x \in \mathcal{X}$$

For $k = N-1, N-2, \ldots 0$, set

$$J_k(x) = \mathbb{E}[g(x, \mu_k(x), w) + \gamma J_{k+1}(f(x, \mu_k(x), w))] \quad \forall x \in \mathcal{X}.$$

Bellman Operators

For any stationary policy μ mapping $x \in \mathcal{X}$ to $\mu(x) \in U(x)$, define T_μ, which maps a cost to go function $J \in \mathbb{R}^{|\mathcal{X}|}$ to another cost to go function $T_\mu J \in \mathbb{R}^{|\mathcal{X}|}$, by

$$(T_\mu J)(x) = \mathbb{E}[g(x, \mu(x), w) + \gamma J(f(x, \mu(x), w))]$$

where (as usual) the expectation is take over the disturbance w.

- We call T_μ the Bellman operator corresponding to a policy μ.
- It is a map from the space of cost-to-go functions to the space of cost-to-go functions.

Bellman Operators

Define T, which maps a cost-to-go function $J \in \mathbb{R}^{|\mathcal{X}|}$ to another cost-to-go function $TJ \in \mathbb{R}^{|\mathcal{X}|}$ by

$$(TJ)(x) = \min_{u \in U(x)} \mathbb{E}[g(x,u,w) + \gamma J(f(x,u,w))]$$

where (as usual) the expection is take over the disturbance w.

- We call T the Bellman operator.
- It is a map from the space of cost-to-go functions to the space of cost-to-go functions.

Alternate notation: transition probabilities

Write the expected cost function as

$$g(x, u) = \mathbb{E}[g(x, u, w)]$$

and transition probabilities as

$$p(x'|x, u) = \mathbb{P}(f(x, u, w) = x')$$

where both integrate over the distribution of the disturbance w.

In this notation

$$T_\mu J(x) = g(x, \mu(x)) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)) J(x')$$

and

$$TJ(x) = \min_{u \in U(X)} g(x, u) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, u) J(x').$$

The Dynamic Programming Algorithm

Old notation: Set

$$J_N^*(x) = c(x) \qquad \forall x \in \mathcal{X}$$

For $k = N-1, N-2, \ldots 0$, set

$$J_k^*(x) = \min_{u \in U(x)} \mathbb{E}[g(x,u,w) + \gamma J_{k+1}^*(f(x,u,w))] \quad \forall x \in \mathcal{X}.$$

Operator notation

$$J_N^* = c \in \mathbb{R}^{|\mathcal{X}|}$$

For $k = N-1, N-2, \ldots, 0$, set

$$J_k^* = T J_{k+1}^*.$$

The Dynamic Programming Algorithm

Main Proposition from last time: old notation

For all initial states $x \in \mathcal{X}$, the optimal cost to go is $J^*(x) = J_0^*(x)$. This is attained by a policy $\pi^* = (\mu_0^*, ..., \mu_{N-1}^*)$ where for all $k \in \{0, \ldots, N-1\}$, $x \in \mathcal{X}$

$$\mu_k^*(x) \in \arg \min_{u \in U(x)} \mathbb{E}[g(x, u, w) + \gamma J_{k+1}^*(f(x, u, w))].$$

Main Proposition from last time: operator notation

For all initial states $x \in \mathcal{X}$, the optimal cost to go is $J^*(x) = J_0^*(x)$. This is attained by a policy $\pi^* = (\mu_0^*, ..., \mu_{N-1}^*)$ satisfying

$$T_{\mu_k^*} J_{k+1}^* = T J_{k+1}^* \qquad \forall k \in \{0, 1, \ldots, N-1\}.$$

The DP Algorithm for policy evaluation

How to find the cost-to-go for any policy $\pi = (\mu_0, \ldots, \mu_{N-1})$?

- $J_\pi(x) = J_0(x)$ where J_0 is output by the following iterative algorithm.

Old notation

$$J_N(x) = c(x) \qquad \forall x \in \mathcal{X}$$

For $k = N-1, N-2, \ldots 0$, set

$$J_k(x) = \mathbb{E}[g(x, \mu_k(x), w) + \gamma J_{k+1}(f(x, \mu_k(x), w))] \quad \forall x \in \mathcal{X}.$$

Operator notation

$$J_N = c \in \mathbb{R}^{|\mathcal{X}|}$$

For $k = N-1, N-2, \ldots, 0$, set

$$J_k = T_{\mu_k} J_{k+1}.$$

Composition of Bellman Operators

In the DP algorithm

$$J^* = TJ_1^* = T(TJ_2^*) = \cdots = T^N c.$$

Analogously, for any policy $\pi = (\mu_0, \mu_1, \ldots \mu_{N-1})$,

$$J_\pi = T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} c.$$

- *Applying the Bellman operator to c iteratively N times gives the optimal cost-to-go in an N period problem with terminal costs c.*
- *Applying the Bellman operators associated with a policy to c iteratively N times gives its cost-to-go in an N period problem with terminal costs c.*

Infinite Horizon Discounted MDPs

The same problem as before, but take $N \to \infty$.

- Finite state and control spaces.
- Periods $0, 1, \ldots$ with controls $u_0, u_1 \ldots,$.
- Stationary transition probabilities $f_k(x, u, w) = f(x, u, w)$ for all $k \in \mathbb{N}$.
- Stationary control spaces: $U_k(x) = U(x)$ for all $k \in \mathbb{N}$.
- Discounted costs: $g_k(x, u, w) = \gamma^k g(x, u, w)$ for $k \in \mathbb{N}$

The objective is to minimize

$$\lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N} \gamma^k g(x_k, u_k, w_k)\right]$$

Infinite Horizon Discounted MDPs

- A policy $\pi = (\mu_0, \mu_1, \mu_2, \ldots)$ is a sequence of mappings where $\mu_k : x \mapsto U(x)$.

- The expected cumulative "cost-to-go" of a policy π from starting state x is

$$J_\pi(x) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N} \gamma^k g(x_k, \mu_k(x_k), w_k)\right]$$

  where $x_{k+1} = f(x_k, \mu_k(x_k), w_k)$ and the expectation is over the i.i.d disturbances $w_0, w_1, w_2 \ldots$

- The optimal expected cost-to-go is

$$J^*(x) = \inf_{\pi\in\Pi} J_\pi(x) \qquad \forall x \in \mathcal{X}.$$

- We say a policy π is optimal if $J_\pi = J^*$.

- For a **stationary policy** $\pi = (\mu, \mu, \mu, \ldots)$ we write J_μ instead of J_π.

Infinite Horizon Discounted MDPs: Main Results

Cost-to go functions

J_μ is the unique solution to the equation $T_\mu J = J$ and iterates of the relation $J_{k+1} = T_\mu J_k$ converge to J_μ at a geometric rate.

Optimal cost-to go functions

J^* is the unique solution to the Bellman equation $TJ = J$ and iterates of the relation $J_{k+1} = TJ_k$ converge to J^* at a geometric rate.

Optimal policies

There exists an optimal stationary policy. A stationary policy $(\mu, \mu, \ldots)$ is optimal if and only if $T_\mu J^* = TJ^*$.

By computing the optimal cost-to-go function we are solving a fixed point equation, and one way to solve this equation is by iterating the Bellman operator. Once we calculate the optimal cost-to-go function we can find the optimal policy by solving the one period problem

$$\min_{u \in U(x)} \mathbb{E}\left[g(x,u,w) + \gamma J^*(f(x,u,w))\right].$$

Example: selling an asset

An instance of optimal stopping.

- *No deadline to sell.*
- Potential buyers make offers in sequence.
- The agent chooses to accept or reject each offer
  - The asset is sold once an offer is accepted.
  - Offers are no longer available once declined.
- Offers are iid.
- Profits can be invested with interest rate $r > 0$ per period.
  - We discounting with rate $\gamma = 1/(1 + r)$.

Example: selling an asset

- Special terminal state t (costless and absorbing)
- $x_k \neq t$ is the offer considered at time k.
- $x_0 = 0$ is fictitious null offer.
- $g(x, \text{sell}) = x$.
- $x_k = w_{k-1}$ for independent $w_0, w_1, \ldots$

Bellman equation $J^* = TJ^*$ becomes

$$J^*(x) = \max\{x, \gamma \mathbb{E}[J^*(w)]\}$$

The optimal policy is a threshold

$$\text{Sell} \iff x_k \geq \alpha \quad \text{where} \quad \alpha = \gamma \mathbb{E}[J^*(w)].$$

This stationary policy is much simpler than what we saw last time.

Properties of the Bellman operator

Monotonicity: T and T_μ are *monotone*.
For any $J \leq J'$

$$T_\mu J \;\leq\; T_\mu J'$$

$$TJ \;\leq\; TJ'$$

Contraction: T and T_μ are maximum-norm *contractions* with modulus γ.
For any J, J'

$$\|T_\mu J - T_\mu J'\|_\infty \;\leq\; \gamma \|J - J'\|_\infty$$

$$\|TJ - TJ'\|_\infty \;\leq\; \gamma \|J - J'\|_\infty$$

where $\|J\|_\infty = \max_{x \in \mathcal{X}} |J(x)|$ is called the "maximum-norm" or "supremum norm".

Relating T and T_μ: $TJ \leq T_\mu J$ but equality always holds for some μ.

- For all J and μ, $TJ \leq T_\mu J$.
- For any J, there is a μ such that $TJ = T_\mu J$

Properties of the Bellman operator: proofs

Relating T and T_μ:

$$\begin{aligned} T_\mu J(x) &= g(x, \mu(x)) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)) J(x') \\ &\geq \min_{u \in U(x)} g(x, u) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, u) J(x') = TJ(x). \end{aligned}$$

The inequality is an equality for all x if

$$\mu(x) \in \operatorname*{argmin}_{u \in U(x)} g(x, u) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, u) J(x') \qquad \forall x \in \mathcal{X}.$$

Properties of the Bellman operator: proofs

Monotonicity: For any $J \leq J'$

$$\begin{aligned} T_\mu J(x) &= g(x, \mu(x)) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)) J(x') \\ &\leq g(x, \mu(x)) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)) J'(x') = T_\mu J'(x). \end{aligned}$$

For any J, $TJ(x) = \min_\mu T_\mu J(x)$.

Therefore

$$J \leq J' \implies TJ(x) = \min_\mu T_\mu J(x) \leq \min_\mu T_\mu J'(x) = TJ'(x)$$

Properties of the Bellman operator: proofs

Basic fact: For any functions f and g, $|\min_z f(z) - \min_z g(z)| \leq \max_z |f(z) - g(z)|$.

Contraction: Fix any J, J' and $x \in \mathcal{X}$

$$|T_\mu J'(x) - T_\mu J'(x)| = \left| \gamma \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)))(J(x') - J'(x')) \right| \leq \gamma \max_{x' \in \mathcal{X}} |J(x') - J'(x') = \gamma \|J - J'\|_\infty.$$

Maximizing over $x \in \mathcal{X}$ gives

$$\|T_\mu J' - T_\mu J'\|_\infty \leq \gamma \|J - J'\|_\infty.$$

Now, we use this to prove T is a contraction.

$$\begin{aligned} |TJ(x) - TJ'(x)| &= |\min_\mu T_\mu J(x) - \min_\mu T_\mu J(x)| \\ &\leq \max_\mu |T_\mu J(x) - T_\mu J'(x)| \qquad \text{(fact above)} \\ &\leq \gamma \|J - J'\|_\infty \qquad \text{(contraction)}. \end{aligned}$$

Maximizing over x implies the result.

Basic fact on previous slide (You can skip this)

We show, for any functions f and g with the same domain,

$$|\min_{z_1} f(z_1) - \min_{z_2} g(z_2)| \leq \max_{z} |f(z) - g(z)|.$$

Proof:

First,

$$\min_{z_1} f(z_1) - \min_{z_2} g(z_2) = \min_{z_1} \max_{z_2} \left(f(z_1) - g(z_2)\right) \leq \max_{z} \left(f(z) - g(z)\right)$$

Analogously

$$\min_{z_1} f(z_1) - \min_{z_2} g(z_2) = \min_{z_1} \max_{z_2} \left(f(z_1) - g(z_2)\right) \geq \min_{z} \left(f(z) - g(z)\right)$$

If $C \equiv \min_{z_1} f(z_1) - \min_{z_2} g(z_2)$ is positive, one can choose z such that $f(z) - g(z)$ is also positive and is larger than C. If C is negative, we can choose z such that $f(z) - g(z)$ is negative and smaller than C. Therefore

$$|\min_{z_1} f(z_1) - \min_{z_2} g(z_2)| \leq \max_{z} |f(z) - g(z)|.$$

Banach Fixed Point Theorem

Definition: $F : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction with respect to $\|\cdot\|$ with modulus $\rho \in (0, 1)$ if

$$\|FJ - FJ'\| \leq \rho \|J - J'\| \qquad \forall J, J' \in \mathbb{R}^n.$$

Theorem If $F : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction with respect to $\|\cdot\|$ with modulus ρ then

- There exists a unique $J^* \in \mathbb{R}^n$ satisfying $FJ^* = J^*$
- For any $J \in \mathbb{R}^n$, $\|F^k J - J^*\| \leq \rho^k \|J - J^*\|$.

(*The theorem actually holds for any complete metric space.*)

Proof of Banach's Fixed Point Theorem

We'll first show $J_\infty \equiv \lim_{N\to\infty} F^N J$ exits, then that J_∞ is a fixed point of F and $F^k V$ converges at a geometric rate to J_∞. Finally, we'll conclude the fixed point must be unique.

For some $J \in \mathbb{R}^n$, set $J_0 = J$ and $J_{k+1} = TJ_k$. Then

$$\|J_2 - J_1\| \le \rho\|J_1 - J_0\| \implies \|J_{k+1} - J_k\| \le \rho^k\|J_1 - J_0\|.$$

Then for all $m \ge 1$

$$\|J_{k+m} - J_k\| \le \sum_{\ell=1}^{m} \|J_{k+\ell} - J_k\| \le \sum_{\ell=1}^{\infty} \|J_{k+\ell} - J_k\| \le \sum_{\ell=0}^{\infty} \rho^k\rho^\ell\|J_1 - J_0\| = \frac{\rho^k}{1-\rho}\|J_1 - J_0\|.$$

This shows the sequence is *Cauchy* and hence $J_\infty \equiv \lim_{N\to\infty} F^N J$ exits.

Existence of a fixed point: We'll show $FJ_\infty = J_\infty$.

$$\begin{aligned} 0 \le \|FJ_\infty - J_\infty\| &\le \|FJ_\infty - J_k\| + \|J_k - J_\infty\| \qquad \forall k \\ &\le \rho\|J_\infty - J_{k-1}\| + \|J_k - J_\infty\| \\ &\to 0 \text{ as } k \to \infty. \end{aligned}$$

Convergence Rate: Since J_∞ is a fixed point

$$\|J_k - J_\infty\| = \|F^k J_0 - F^k J_\infty\| \leq \rho^k \|J_0 - J_\infty\|$$

Uniqueness: If $J = FJ$ and $J' = FJ'$ then

$$\|J - J'\| = \|FJ - FJ'\| \leq \rho \|J - J'\|$$

which implies $\|J - J'\| = 0$.

■

Bellman's equation and optimal policies

Since T is a contraction:

1. There exists a unique solution to the "Belman equation" $TJ = J$.
2. The solution can be found by iterating the relation $J_{k+1} = TJ_k$.

We have defined

$$J^*(x) = \inf_\pi J_\pi(x) \qquad \text{where} \qquad J_\pi(x) = \lim_{N\to\infty} \mathbb{E}_\pi \left[\sum_{k=0}^{N} \gamma^k g(x_k, \mu_k(x_k), w_k) \right]$$

We simplify notation by writing J_μ when $\pi = (\mu, \mu, \mu, \ldots)$

Proposition:

- J^* is the unique solution to the Bellman equation $J = TJ$.
- The greedy policy μ w.r.t J^*, defined by $T_\mu J^* = TJ^*$, satisfies $J_\mu = J^*$

Bellman's equation and optimal policies

Proposition:

- J^* is the unique solution to the Bellman equation $J = TJ$.
- The greedy policy μ w.r.t J^*, defined by $T_\mu J^* = TJ^*$, satisfies $J_\mu = J^*$

Proof: Let $0 \in \mathbb{R}^{|\mathcal{X}|}$ denote a vector of zeros.
For any $\pi = (\mu_0, \mu_1, \dots,)$,

$$J_\pi = \lim_{N\to\infty} T_{\mu_0} T_{\mu_1} \cdots T_{\mu_N} 0.$$

- Fix $\bar{J}$ solving $T\bar{J} = \bar{J}$
- Fix μ solving $T_\mu \bar{J} = T\bar{J}$
- Then

$$\begin{aligned} T_\mu \bar{J} = \bar{J} &\implies T_\mu^k \bar{J} = \bar{J} \\ &\implies J_\mu \equiv \lim_{N\to\infty} T_\mu^N \bar{J} = \bar{J}. \end{aligned}$$

It remains to show $\bar{J} = J^*$.

- Certainly $\bar{J} \geq J^*$ since $\bar{J}(x) = J_\mu(x) \geq \inf_\pi J_\pi(x) = J^*(x)$
- But also $\bar{J} \leq J^*$ since any policy $\pi = (\mu_0, \mu_1, \dots)$

$$\bar{J}(x) = \lim_{N\to\infty} (T^N 0)(x) \leq \lim_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_N} 0)(x) = J_\pi(x).$$

■

Indefinite Horizon Problems

We consider the problem of minimizing expected costs until a special termination state t is reached.

- *The problem will end in finite time, but we're not sure when.*

Many RL problems involve learning over a sequence of episodes, each of which has indefinite horizon.

Examples

- Atari games
- Many models of customer interaction with a web service
- Problems with a regenerative structure (e.g. Queuing)

The book calls these Stochastic Shortest Path Problems.

Indefinite Horizon Problems

We consider the problem of minimizing expected costs until a special termination state t is reached.

- The state space is $\mathcal{X} \cup \{t\}$.
- $\mathcal{X}$ is a finite set
- t is costless ($g(t,u) = 0$ and absorbing ($p(t|t,u) = 1$)
- Any policy incurs zero expected cost starting from t.

Assumption: Under any policy and initial state, the terminal node is reached with probability 1.

It turns out to be more elegant to explicitly track the cost only of non terminal states $x \in \mathcal{X}$.

Define the Bellman operators

$$\begin{aligned} T_\mu J(x) &= g(x,\mu(x)) + \sum_{x' \in \mathcal{X}} p(x'|x,\mu(x)) J(x') \\ TJ(x) &= \min_{u \in U(x)} \; g(x,u) + \sum_{x' \in \mathcal{X}} p(x'|x,u) J(x') \end{aligned}$$

where $J \in \mathbb{R}^{|\mathcal{X}|}$.

Warmup: Geometrically distributed horizon

Consider a special case of the problem above with independent geometric horizon.

The probability of termination in the next period is $1 - \gamma$:

- $\sum_{x' \in \mathcal{X}} p(x'|x, u) = \gamma$ for all x, u.

> Your homework asks you to show this is equivalent
> *(in terms of expected costs incurred)*
> to an infinite horizon problem with discount factor γ.

Then T and T_μ are maximum norm contractions with modulus γ.

Proof for T_μ

$$|T_\mu J(x) - T_\mu J'(x)| = \left| \sum_{x' \in \mathcal{X}} p(x'|x, \mu(x))(J(x') - J'(x')) \right| \leq \left(\sum_{x' \in \mathcal{X}} p(x'|x, \mu(x)) \right) \|J - J'\|_\infty = \gamma |J - J'|$$

■

Properties of the Bellman operator

Monotonicity: T and T_μ are *monotone*.

Contraction: T and T_μ are **weighted maximum-norm contractions** with modulus that depends on the transition probabilities.

Relating T and T_μ: $TJ \leq T_\mu J$ but equality always holds for some μ.

Due to these properties much of the theory from infinite horizon discounted problems applies to indefinite horizon problems.

Contraction

For $w : x \mapsto w(x) > 0$, define the weighted maximum-norm

$$\|J\|_{\infty,w} = \max_{x \in \mathcal{X}} w(x)|J(x)|.$$

Goal: construct a w such that T is a contraction w.r.t. $\|\cdot\|_{\infty,w}$.

Define $\tau = \inf\{k \in \mathbb{N} : x_k = t\}$ to be the first hitting time of t.

For $x \in \mathcal{X}$, define

$$V(x) = \sup_{\pi} \mathbb{E}_{\pi}[\tau | x_0 = x]$$

This satisfies the Bellman equation

$$V(x) = 1 + \max_{u \in U(x)} \sum_{x' \in \mathcal{X}} p(x'|x,u)V(x') \qquad \forall x \in \mathcal{X}$$

for an MDP with "costs" $g(x,u) = -1$ for all $x \in \mathcal{X}$.

Contraction

Proposition:

T and T_μ are contractions with respect to the weighted maximum norm $\|\cdot\|_{\infty,1/V}$ with modulus $\alpha = \max_{x\in\mathcal{X}} \frac{V(x)-1}{V(x)}$

.

Proof for T_μ:

Note that from Bellman's equation for V, for all $x \in \mathcal{X}$

$$\sum_{x'\in\mathcal{X}} p(x'|x,\mu(x))V(x') \leq V(x) - 1 \leq \alpha V(x)$$

so

$$\max_{x\in\mathcal{X}} \frac{\sum_{x'\in\mathcal{X}} p(x'|x,\mu(x))V(x')}{V(x)} \leq \alpha.$$

Then

$$\begin{aligned}
\|T_\mu J - T_\mu J'\|_{\infty,1/v} &= \max_{x\in\mathcal{X}} \frac{1}{V(x)} \left| \sum_{x'\in\mathcal{X}} p(x'|x,\mu(x))(J(x') - J'(x')) \right| \\
&= \max_{x\in\mathcal{X}} \frac{1}{V(x)} \left| \sum_{x'\in\mathcal{X}} p(x'|x,\mu(x))V(x') \left(\frac{J(x') - J'(x')}{V(x')} \right) \right| \\
&\leq \max_{x\in\mathcal{X}} \left(\frac{\sum_{x'\in\mathcal{X}} p(x'|x,\mu(x))V(x')}{V(x)} \right) \|J - J'\|_{\infty,1/V} \\
&\leq \alpha \|J - J'\|_{\infty,1/V}
\end{aligned}$$

Contraction

Proposition:

T and T_μ are contractions with respect to the weighted maximum norm $\|\cdot\|_{\infty,1/V}$ with modulus $\alpha = \max_{x\in\mathcal{X}} \frac{V(x)-1}{V(x)}$

.

Proof for T:

Since T_μ is a contraction

$$\frac{T_\mu J(x)}{V(x)} \leq \frac{T_\mu J'(x)}{V(x)} + \alpha\|J - J'\|_{\infty,1/V} \qquad \forall\mu.$$

Then

$$\frac{TJ(x)}{V(x)} = \min_\mu \frac{T_\mu J(x)}{V(x)} \leq \min_\mu \frac{T_\mu J'(x)}{V(x)} + \alpha\|J - J'\|_{\infty,1/V} = \frac{TJ'(x)}{V(x)} + \alpha\|J - J'\|_{\infty,1/V}.$$

Reversing the role of J and J' gives

$$\frac{|TJ(x) - TJ'(x)|}{V(x)} \leq \alpha\|J - J'\|_{\infty,1/V} \qquad \forall x \in \mathcal{X}.$$

■.

Understanding the weighted max-norm

Proposition:

T and T_μ are contractions with respect to the weighted maximum norm $\|\cdot\|_{\infty,1/V}$ with modulus $\alpha = \max_{x\in\mathcal{X}} \frac{V(x)-1}{V(x)}$

.

- Maximizing over x we see

$$\alpha = \frac{\|V\|_\infty - 1}{\|V\|_\infty} = 1 - \frac{1}{\|V_\infty\|}$$

  α is close to 1 when the expected termination time is large from some initial states.

- When the termination time has distribution $\mathrm{Geometric}(1-\gamma)$, $\|V\|_\infty = 1/(1-\gamma)$ so

$$\alpha = 1 - \frac{1}{1/(1-\gamma)} = \gamma$$

  and the theory here generalizes our previous result.

- A small weighted-max norm implies the max-norm is small, since

$$\|J\|_{\infty,1/V} = \max_{x\in\mathcal{X}} \frac{J(x)}{V(x)} \geq \max_{x\in\mathcal{X}} \frac{J(x)}{\|V\|_\infty} = \frac{\|J\|_\infty}{\|V\|_\infty},$$

  or $\|J\|_\infty \leq \|J\|_{\infty,1/V}\|V\|_\infty$.