Fondements sur la modélisation des réseaux

Stochastic shortest path and Markov decision processes

Ana Busic

Inria Paris - DI ENS

http://www.di.ens.fr/~busic/

Paris, Décembre 2017

Controlled dynamics

Discrete-time controlled dynamic system

$$x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \ldots$$

- State space $\mathcal{S}$ (assumed countable). For all k, $x_k \in \mathcal{S}$.
- Control space $\mathcal{C}$. For $x \in \mathcal{S}$, $U(x) \subset \mathcal{C}$ denotes a non-empty set of admissible controls in state x.
  For all k, $u_k \in U(x_k)$.
- Random disturbance space $\mathcal{D}$ (assumed countable): $w_k \in \mathcal{D}$, $\forall k$.
  For all k, $P(w_k \mid x_k, u_k)$ is the probability of occurence of w_k when the current state and control are x_k and u_k.

  Assumption (time-homogeneous disturbances): the probability distributions $P(\cdot \mid x, u)$, $x \in \mathcal{S}$, $u \in U(x)$ are assumed to be independent of k.

Cost function

Assumption: cost accumulates additively over time.

Cost per-stage function: $g : \mathcal{S} \times \mathcal{C} \times \mathcal{D} \to \mathbb{R}$

Terminal cost: $G : \mathcal{S} \to \mathbb{R}$.

Discount factor $0 < \alpha \leq 1$.

Meaning of $\alpha < 1$: 1 EUR in the future has less value than 1 EUR today. If the interest rate is r per period of time, then the value today of 1 EUR received k periods from now is $(1 + r)^{-k}$. Discount factor: $\alpha = (1 + r)^{-1}$.

Finite horizon problems: minimizing the expected N-stage costs,

$$E\left[\alpha^N G(X_N) + \sum_{k=0}^{N-1} \alpha^k g(X_k, U_k, W_k) \mid X_0 = x\right],$$

where $\alpha^N G(X_N)$ is a terminal cost for ending up with final state X_N.

Decision policies

Definition. An admissible decision policy is a sequence $\pi = \{\mu_0, \mu_1, \ldots\}$ where each μ_k is a function mapping the states into controls with $\mu_k(x) \in U(x)$ for all $x \in \mathcal{S}$.

Decision policies

Definition. An admissible decision policy is a sequence $\pi = \{\mu_0, \mu_1, \ldots\}$ where each μ_k is a function mapping the states into controls with $\mu_k(x) \in U(x)$ for all $x \in \mathcal{S}$.

Once a policy is fixed, the sequence of states X_k becomes a discrete time, countable state-space Markov chain with transition probabilities

$$P(X_{k+1} = y \mid X_k = x) = \sum_{w \,:\, f(x,\mu_k(x),w)=y} P(w \mid x, \mu_k(x)).$$

Decision policies

Definition. An admissible decision policy is a sequence $\pi = \{\mu_0, \mu_1, \ldots\}$ where each μ_k is a function mapping the states into controls with $\mu_k(x) \in U(x)$ for all $x \in \mathcal{S}$.

Once a policy is fixed, the sequence of states X_k becomes a discrete time, countable state-space Markov chain with transition probabilities

$$P(X_{k+1} = y \mid X_k = x) = \sum_{w \,:\, f(x,\mu_k(x),w)=y} P(w \mid x, \mu_k(x)).$$

Definition. A decision policy is called stationary if $\mu_k = \mu, \forall k$.

Decision policies

Definition. An admissible decision policy is a sequence $\pi = \{\mu_0, \mu_1, \ldots\}$ where each μ_k is a function mapping the states into controls with $\mu_k(x) \in U(x)$ for all $x \in \mathcal{S}$.

Once a policy is fixed, the sequence of states X_k becomes a discrete time, countable state-space Markov chain with transition probabilities

$$P(X_{k+1} = y \mid X_k = x) = \sum_{w\,:\,f(x,\mu_k(x),w)=y} P(w \mid x, \mu_k(x)).$$

Definition. A decision policy is called stationary if $\mu_k = \mu, \forall k$.

A stationary policy yields a time-homogeneous Markov chain.

Finite horizon dynamic programming

Expected N-stage cost under $\pi = \{\mu_0, \mu_1, \ldots\}$, starting from $X_0 = x$:

$$V_N^\pi(x) = E\left[\alpha^N G(X_N) + \sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

The optimal cost function: $V_N(x) = \min_\pi V_N^\pi(x)$.

Finite horizon dynamic programming

Expected N-stage cost under $\pi = \{\mu_0, \mu_1, \ldots\}$, starting from $X_0 = x$:

$$V_N^\pi(x) = E\left[\alpha^N G(X_N) + \sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

The optimal cost function: $V_N(x) = \min_\pi V_N^\pi(x)$.

Principle of optimality. Let $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$ be an optimal policy for the initial N-stage problem. Assume that when using π^*, a given state x_i occurs at time i with positive probability. Consider a subproblem where we start at time i in state x_i and minimize the cost-to-go from time i to N

$$E\left[\alpha^N G(X_N) + \sum_{k=i}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_i = x_i\right]$$

Then the truncated policy $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ is optimal for this subproblem.

Finite horizon dynamic programming algorithm

Expected N-stage cost under $\pi = \{\mu_0, \mu_1, \ldots\}$, starting from $X_0 = x$:

$$V_N^\pi(x) = E\left[\alpha^N G(X_N) + \sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

The optimal cost function: $V_N(x) = \min_\pi V_N^\pi(x)$.

Dynamic programming: (recursive computation)

- Initialization: $J_N(x) = \alpha^N G(x), \forall x \in \mathcal{S}$.
- For $k = 1 \ldots, N$

$$J_{N-k}(x) = \min_{u \in U(x)} E\left[\alpha^{N-k} g(x, u, W) + J_{N-k+1}(f(x, u, W))\right], \forall x \in \mathcal{S}.$$

- $V_N = J_0$.

Finite horizon dynamic programming algorithm

Remark that $J_k = \alpha^k V_{N-k}$. Alternative formulation of DP algorithm:

Dynamic programming: (recursive computation)

- Initialization: $V_0(x) = G(x), \forall x \in \mathcal{S}$.
- For $k = 1 \ldots, N$

$$\begin{aligned} V_k(x) &= \min_{u \in U(x)} E\left[g(x,u,W) + \alpha V_{k-1}(f(x,u,W))\right] \\ &= \min_{u \in U(x)} \sum_{w \in \mathcal{D}} P(w \mid x,u)\left(g(x,u,w) + \alpha V_{k-1}(f(x,u,w))\right), \forall x \in \mathcal{S} \end{aligned}$$

Finite horizon dynamic programming algorithm

Notation:

- $p_{xy}(u) = \sum_{w \,:\, f(x,u,w)=y} P(w \mid x, u)$
- $\hat{g}(x, u) = \sum_{w \in \mathcal{D}} P(w \mid x, u) g(x, u, w)$

For $k = 1 \ldots, N$

$$\begin{aligned}
V_k(x) &= \min_{u \in U(x)} \sum_{w \in \mathcal{D}} P(w \mid x, u) \left(g(x, u, w) + \alpha V_{k-1}(f(x, u, w))\right) \\
&= \min_{u \in U(x)} \left(\hat{g}(x, u) + \alpha \sum_{w \in \mathcal{D}} P(w \mid x, u) V_{k-1}(f(x, u, w)) \right) \\
&= \min_{u \in U(x)} \left(\hat{g}(x, u) + \alpha \sum_{y \in \mathcal{S}} p_{xy}(u) V_{k-1}(y) \right).
\end{aligned}$$

Notation

- For any function $J : \mathcal{S} \to \mathbb{R}$, we consider the function obtained by DP iteration to J
(an optimal cost-to-go function for $N = 1$ and terminal cost J):

$$(TJ)(x) = \min_{u \in U(x)} E\left[g(x, u, W) + \alpha J(f(x, u, W))\right], \; x \in \mathcal{S}.$$

- For any function $J : \mathcal{S} \to \mathbb{R}$ and any admissible control function $\mu : \mathcal{S} \to \mathcal{C}$,

$$(T_\mu J)(x) = E\left[g(x, \mu(x), W) + \alpha J(f(x, \mu(x), W))\right], \; x \in \mathcal{S}.$$

Properties of operators T and T_μ

For any two functions J, J', we write $J \leq J'$ if $J(x) \leq J'(x)$, $\forall x \in \mathcal{S}$.

Lemma (Monotonicity)

For any two vectors $J \leq J'$, and for any stationary policy μ,

$$T^k J \leq T^k J', \quad k \geq 1,$$
$$T_\mu^k J \leq T_\mu^k J', \quad k \geq 1,$$

where T^k denotes the composition of the mapping T with itself k times (for $k = 0$, it is the identity mapping, $T^0 J := J$).

Proof. Follows from the interpretations of T^k and T_μ^k as k-stage cost-to-go: an increase of the terminal cost can only increase the k-stage cost-to-go.

Properties of operators T and T_μ

Notation: $e : \mathcal{S} \to \mathbb{R}$ is the unit function, $e(x) = 1$, $\forall x \in \mathcal{S}$.

Lemma

For any $k \geq 0$, any function $J : \mathcal{S} \to \mathbb{R}$, any stationary policy μ and any $r > 0$,

$$(T^k(J + re))(x) = (T^k J)(x) + \alpha^k r, \quad \forall x \in \mathcal{S},$$
$$(T_\mu^k(J + re))(x) = (T_\mu^k J)(x) + \alpha^k r, \quad \forall x \in \mathcal{S}.$$

Proof. By induction on k.

Infinite horizon problems

A reasonable approximation of a finite horizon problems with very large number of stages.

Infinite horizon problems

A reasonable approximation of a finite horizon problems with very large number of stages.

Optimal policy is often stationary - easier to implement.

Infinite horizon problems

A reasonable approximation of a finite horizon problems with very large number of stages.

Optimal policy is often stationary - easier to implement.

Typical objective is to minimize

$$V^{\pi}(x) = \limsup_{N\to\infty} E\left[\sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

Infinite horizon problems

A reasonable approximation of a finite horizon problems with very large number of stages.

Optimal policy is often stationary - easier to implement.

Typical objective is to minimize

$$V^{\pi}(x) = \limsup_{N\to\infty} E\left[\sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

- Stochastic shortest path problems: $\alpha = 1$ and the state space contains a special state t that is cost-free termination state. The objective is to reach the termination state with minimal expected cost.

Infinite horizon problems

A reasonable approximation of a finite horizon problems with very large number of stages.

Optimal policy is often stationary - easier to implement.

Typical objective is to minimize

$$V^{\pi}(x) = \limsup_{N\to\infty} E\left[\sum_{k=0}^{N-1} \alpha^k g(X_k, \mu_k(X_k), W_k) \mid X_0 = x\right].$$

- Stochastic shortest path problems: $\alpha = 1$ and the state space contains a special state t that is cost-free termination state. The objective is to reach the termination state with minimal expected cost.
- Discounted problems: $\alpha < 1$.

Infinite horizon problems

In some problems (e.g. $\alpha = 1$; $g(x, u, w) > 0$, $\forall x, u, w$), $V^\pi(x) = \infty$ for all π and all initial states x.

In this case, we will be interested in minimizing the average cost per stage,

$$\lim_{N\to\infty} \frac{1}{N} V_N^\pi(x),$$

when this limit is well defined and finite.

Main questions

- Under which conditions $V^*(x) = \lim_{N\to\infty} V_N^*(x), \forall x$?

Main questions

- Under which conditions $V^*(x) = \lim_{N\to\infty} V_N^*(x), \forall x$?
- Under which conditions,

$$V^*(x) = \min_{u \in U(x)} E[g(x,u,W) + \alpha V^*(f(x,u,W))], \ \forall x.$$

This is called Bellman's equation.

Main questions

- Under which conditions $V^*(x) = \lim_{N\to\infty} V_N^*(x), \forall x$?
- Under which conditions,

$$V^*(x) = \min_{u \in U(x)} E[g(x,u,W) + \alpha V^*(f(x,u,W))], \ \forall x.$$

  This is called Bellman's equation.
- Is there an optimal policy that is stationary?
  If in the Bellman equation the minimum is attained for some μ, does his imply that the stationary policy $\pi = (\mu, \mu, \dots)$ is optimal?

Main questions

- Under which conditions $V^*(x) = \lim_{N\to\infty} V_N^*(x), \forall x$?
- Under which conditions,

$$V^*(x) = \min_{u \in U(x)} E[g(x,u,W) + \alpha V^*(f(x,u,W))], \ \forall x.$$

  This is called Bellman's equation.
- Is there an optimal policy that is stationary?
  If in the Bellman equation the minimum is attained for some μ, does his imply that the stationary policy $\pi = (\mu, \mu, \ldots)$ is optimal?
- How to compute or approximate V^* and how to find an optimal stationary policy?

Stochastic shortest path problems

Deterministic shortest path problem:

- Input: a graph with nodes $1, 2, ..., n, t$, where t is a special state called the *destination* or *termination state*.
- Problem: for each node $i \neq t$, choose a successor node $\mu(i)$ so that $(i, \mu(i))$ is an arc, and the path formed by a sequence of successor nodes starting at any node j terminates at t and has minimum sum of arc lengths over all paths that start at j and terminate at t.

Stochastic shortest path problems

Deterministic shortest path problem:

- Input: a graph with nodes $1, 2, \ldots, n, t$, where t is a special state called the *destination* or *termination state*.
- Problem: for each node $i \neq t$, choose a successor node $\mu(i)$ so that $(i, \mu(i))$ is an arc, and the path formed by a sequence of successor nodes starting at any node j terminates at t and has minimum sum of arc lengths over all paths that start at j and terminate at t.

Stochastic shortest path problem (SSP):

- At each node i, we must select a probability distribution over all possible successor nodes j out of a given set of probability distributions $p_{ij}(u)$ parametrized by a control $u \in U(i)$.
- For a given selection of distributions and for a given origin node, the path traversed as well as its length are now random, but we wish that the path leads to the destination t with probability 1 and has minimum expected length.

Stochastic shortest path problems

A special case of the total cost infinite horizon problem where:

1. No discounting ($\alpha = 1$).
2. State space $\mathcal{S} = \{1, \ldots, n, t\}$ with transition probabilities

$$p_{ij}(u) = P(X_{k+1} = j | X_k = i, U_k = u), \qquad i, j \in S, u \in U(i).$$

   The destination t is absorbing, i.e., for all $u \in U(t)$,

$$p_{tt}(u) = 1.$$

3. The control constraint set $U(i)$ is a finite set for all i.
4. A cost $g(i, u)$ is incurred when control $u \in U(i)$ is selected. The destination is *cost-free*. i.e. $g(t, u) = 0$ for all $u \in U(t)$.

Note: If the cost of the applying control u at state i and moving to state j is $\tilde{g}(i, u, j)$, we use as cost per stage the expected cost

$$g(i, u) = \sum_{j=1,\ldots,n,t} p_{ij}(u)\tilde{g}(i, u, j).$$

Stochastic shortest path problems

Objective: to reach the termination state with minimal expected cost.

Two special cases:

- Deterministic shortest path problem. States: nodes (state t is the destination), controls: arcs, costs: values of arcs.
- Finite horizon problem. Transitions from state-time pairs (i, k) to $(j, k+1)$ according to $p_{ij}(u)$ of the finite horizon problem. The termination state corresponds to the end of horizon and it is reached with probability 1 in one step from any (j, N) at a cost $G(j)$.

DP operators

Since the destination t is cost-free and absorbing, the cost starting from t is zero for every policy.

Define the mappings T and T_μ on functions J with components $J(1), \dots, J(n)$ by

$$(TJ)(i) = \min_{u \in U(i)} [g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j)], \qquad i = 1, \dots, n,$$

$$(T_\mu J)(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j), \qquad i = 1, \dots, n,$$

For the states i and controls u for which $p_{it}(u) > 0$, we have

$$\sum_{j=1}^{n} p_{ij}(u) = 1 - p_{it}(u) < 1.$$

Vector notation

For any stationary policy μ,

$$P_\mu = \begin{bmatrix} p_{11}(\mu(1)) & \cdots & p_{1n}(\mu(1)) \\ \vdots & \vdots & \vdots \\ p_{n1}(\mu(n)) & \cdots & p_{nn}(\mu(n)) \end{bmatrix}, \quad g_\mu = \begin{bmatrix} g(1,\mu(1)) \\ \vdots \\ g(n,\mu(n)) \end{bmatrix}.$$

Then

$$T_\mu J = g_\mu + P_\mu J.$$

The cost function of a policy $\pi = \mu_0, \mu_1, \ldots$

$$J_\pi = \limsup_{N\to\infty} T_{\mu_0} \cdots T_{\mu_{N-1}} J_0 = \limsup_{N\to\infty} (g_{\mu_0} + \sum_{k=1}^{N-1} P_{\mu_0} \cdots P_{\mu_{k-1}} g_{\mu_k}),$$

where J_0 denotes the zero vector.

The cost function of a stationary policy μ

$$J_\mu = \limsup_{N\to\infty} T_\mu^N J_0 = \limsup_{N\to\infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu.$$

Assumptions

Definition. A stationary policy μ is said to be proper if,

$$\rho_\mu = \max_{i=1,\ldots,n} P\{x_n \neq t | x_0 = i, \mu\} < 1.$$

A stationary policy that is not proper is said to be improper.

- μ is proper iff in the Markov chain corresponding to μ for any state i there is a path of positive probability to the termination state.
- Under a proper policy,

$$\begin{aligned} P(X_{2n} \neq t \mid X_0 = i, \mu) &= P(X_{2n} \neq t \mid X_n \neq t,\ X_0 = i, \mu) \\ &\quad \times P(X_n \neq t \mid X_0 = i, \mu) \\ &\leq \rho_\mu^2 \end{aligned}$$

  and for any k, $P(X_k \neq t \mid X_0 = i, \mu) \leq \rho_\mu^{\lfloor k/n \rfloor}$.
  $\Rightarrow$ the termination state will eventually be reached with probability 1 under a proper policy.

Assumptions

The associated total cost-to-go vector J_μ exists and is finite as the expected cost at the kth period is bounded in absolute value by

$$\rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i, \mu(i))|,$$

so that

$$|J_\mu(i)| \leq \lim_{N\to\infty} \sum_{k=0}^{N-1} \rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i, \mu(i))| < \infty.$$

Assumptions:

- A1 There exists at least one proper policy.
- A2 For every improper policy μ, $J_\mu(i) = \infty$ for at least one state i.

Assumptions

The associated total cost-to-go vector J_μ exists and is finite as the expected cost at the kth period is bounded in absolute value by

$$\rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i,\mu(i))|,$$

so that

$$|J_\mu(i)| \le \lim_{N\to\infty} \sum_{k=0}^{N-1} \rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i,\mu(i))| < \infty.$$

Assumptions:

A1 There exists at least one proper policy.

A2 For every improper policy μ, $J_\mu(i) = \infty$ for at least one state i.

Remarks:

- Sufficient conditions for A2: $g(i,u) > 0$ for all $i \neq t$ and $u \in U(i)$.
- Special case: A1 and A2 are satisfied is when *all* policies are proper.
- In the deterministic shortest path problem, A1 corresponds to the existence of a path from each node to the destination and A2 to assuming all cycles have strictly positive cost.

Main results

1. The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.

Main results

1. The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.
2. DP algorithm converges to the optimal cost vector J^* for an arbitrary starting vector.

Main results

1. The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.
2. DP algorithm converges to the optimal cost vector J^* for an arbitrary starting vector.
3. A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

Main results

1. The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.
2. DP algorithm converges to the optimal cost vector J^* for an arbitrary starting vector.
3. A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.
4. Computation of an optimal proper policy.

Properties of proper policies

Proposition 1.

(a) For a proper policy μ, the associated cost vector J_μ satisfies

$$\lim_{k\to\infty} (T_\mu^k J)(i) = J_\mu(i), \qquad i = 1, \ldots, n,$$

for every vector J. Furthermore,

$$J_\mu = T_\mu J_\mu,$$

and J_μ is the unique solution of this equation.

(b) A stationary policy μ satisfying for some vector J,

$$J(i) \geq (T_\mu J)(i), \qquad i = 1, \ldots, n,$$

is proper.

Proof

Property (a). We have:

$$T_\mu J = g_\mu + P_\mu J.$$

By induction, for all $J \in \mathbb{R}^n$ and $k \geq 1$

$$T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu.$$

As μ is proper, for all $J \in \mathbb{R}^n$, we have $\lim\limits_{k\to\infty} P_\mu^k J = 0$, so that

$$\lim_{k\to\infty} T_\mu^k J = \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu = J_\mu.$$

Also, by definition

$$T_\mu^{k+1} J = g_\mu + P_\mu T_\mu^k J,$$

as $k \to \infty$, we obtain $J_\mu = g_\mu + P_\mu J_\mu$, which is equivalent to $J_\mu = T_\mu J_\mu$.

Uniqueness: if $J = T_\mu J$, then we have $J = T_\mu^k J$ for all k, so that $J = \lim_{k\to\infty} T_\mu^k J = J_\mu$.

Proof

Property (b). By the hypothesis $J \geq T_\mu J$, and the monotonicity of T_μ,

$$J \geq T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu, \qquad k = 1, 2, \ldots$$

If μ were not proper, by A2, some component of the sum in the right hand side of the above relation would diverge to ∞ as $k \to \infty$, which is a contradiction.

Q.E.D.

Bellman's equation

Theorem

1. The optimal cost vector J^* satisfies Bellman's equation
$$J^* = TJ^*.$$
Furthermore, J^* is the unique solution of this equation.
2. We have
$$\lim_{k\to\infty} (T^k J)(i) = J^*(i), \quad i = 1, ..., n,$$
for every vector J.
3. A stationary policy μ is optimal if and only if
$$T_\mu J^* = TJ^*.$$

Proof

Step I. T has at most one fixed point.

If J and J' are two fixed points, then we select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'} J'$;
(possible because the control constraint set is finite)

By Prop. 1(b), we have that μ and μ' are proper, and Prop. 1(a) implies that $J = J_\mu$ and $J' = J_{\mu'}$. We have $J = T^k J \leq T^k_{\mu'} J$ for all $k \geq 1$, and by Prop. 1(a), we obtain $J \leq \lim_{k\to\infty} T^k_{\mu'} J = J_{\mu'} = J'$. Similarly, $J' \leq J$, showing that $J = J'$ and that T has at most one fixed point.

Proof

Step II. T has at least one fixed point.
Let μ be a proper policy (there exists one by A1). Choose μ' such that

$$T_{\mu'} J_\mu = T J_\mu.$$

Then we have $J_\mu = T_\mu J_\mu \geq T_{\mu'} J_\mu$. By Prop. 1(b), μ' is proper, and using the monotonicity of $T_{\mu'}$ and Prop. 1(a), we obtain

$$J_\mu \geq \lim_{k \to \infty} T_{\mu'}^k J_\mu = J_{\mu'}.$$

Continuing in the same manner, we construct a sequence $\{\mu^k\}$ such that each μ^k is proper and

$$J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}}, \quad k = 0, 1, \ldots$$

Since the set of proper policies is finite, some policy μ must be repeated within the sequence $\{\mu^k\}$, and for this policy

$$J_\mu = T J_\mu.$$

Thus J_μ is a fixed point of T.
Step I $\Rightarrow$ J_μ is the unique fixed point of T.

Proof

Step III. The unique fixed point of T is equal to the optimal cost vector J^*, and $T^k J \to J^*$ for all J.

The construction in Step II provides a proper μ such that $TJ_\mu = J_\mu$.

We will show that $T^k J \to J_\mu$ for all J and that $J_\mu = J^*$.

Let $e = (1, 1, ..., 1)$, let $\delta > 0$ be some scalar, and let $\hat{J}$ be the vector satisfying

$$T_\mu \hat{J} = \hat{J} - \delta e.$$

There is a unique such vector because the equation $\hat{J} = T_\mu \hat{J} + \delta e$ can be written $\hat{J} = g_\mu + \delta e + P_\mu \hat{J}$, so $\hat{J}$ is the cost vector corresponding to μ for g_μ replaced by $g_\mu + \delta e$. Since μ is proper, by Prop. 1(a), $\hat{J}$ is unique.

Furthermore, we have $J_\mu \le \hat{J}$, which implies that

$$J_\mu = TJ_\mu \le T\hat{J} \le T_\mu \hat{J} = \hat{J} - \delta e \le \hat{J}.$$

Proof

Using the monotonicity of T, we obtain

$$J_\mu = T^k J_\mu \leq T^k \hat{J} \leq T^{k-1} \hat{J} \leq \hat{J}, \qquad k \geq 1.$$

Hence, $T^k \hat{J}$ converges to some vector $\tilde{J}$, and we have

$$T\tilde{J} = T(\lim_{k\to\infty} T^k \hat{J}).$$

The mapping T can be seen to be continuous, so we can interchange T with the limit in the preceding relation, obtaining $\tilde{J} = T\tilde{J}$.

By the uniqueness of the fixed point of T, we must have $\tilde{J} = J_\mu$.
Also,

$$J_\mu - \delta e = TJ_\mu - \delta e \leq T(J_\mu - \delta e) \leq TJ_\mu = J_\mu.$$

Thus, $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above. As earlier, it follows that $\lim_{k\to\infty} T^k(J_\mu - \delta e) = J_\mu$. For any J, we can find $\delta > 0$ such that

$$J_\mu - \delta e \leq J \leq \hat{J}.$$

Proof

By the monotonicity of T, we then have

$$T^k(J_\mu - \delta e) \le T^k J \le T^k \hat{J}, \qquad k \ge 1,$$

and since $\lim_{k\to\infty} T^k(J_\mu - \delta e) = \lim_{k\to\infty} T^k \hat{J} = J_\mu$, it follows that

$$\lim_{k\to\infty} T^k J = J_\mu.$$

To show that $J_\mu = J^*$, take any policy $\pi = \{\mu_0, \mu_1, ...\}$. We have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} J_0 \ge T^k J_0,$$

where J_0 is the zero vector. Taking the limsup of both sides as $k \to \infty$,

$$J_\pi \ge J_\mu,$$

so μ is an optimal stationary policy and $J_\mu = J^*$.

Proof

(c) If μ is optimal, then $J_\mu = J^*$ and, by A1 and A2, μ is proper, so by Prop. 1(a),

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*.$$

Conversely, if $J^* = TJ^* = T_\mu J^*$, it follows from Prop. 1(b) that μ is proper, and by using Prop. 1(a), we obtain $J^* = J_\mu$. Therefore, μ is optimal.

Q.E.D.

Example: Minimizing Expected Time to Termination

Problem: Minimize the expected time to termination.

Cost: $g(i, u) = 1, \qquad i = 1, ..., n, \qquad u \in U(i),$

$J^*(i)$ uniquely solve Bellman's equation:

$$J^*(i) = \min_{u \in U(i)} \left[1 + \sum_{j=1}^{n} p_{ij}(u) J^*(j) \right], \qquad i = 1, ..., n.$$

Special case: if only one control at each state, $J^*(i)$ represents the mean first passage time m_i from i to t:

$$m_i = 1 + \sum_{j=1}^{n} p_{ij} m_j, \qquad i = 1, ..., n.$$

Example

A spider and a fly move along a line $\mathbb{Z}$ at times $k = 0, 1, \ldots$.

At each time, the following transitions:

- Fly: one unit to the left with probability p, one unit to the right with probability p, and stays where it is with probability $1 - 2p$.
- Spider: one unit towards the fly if its distance from the fly is more that one unit. If the spider is one unit away from the fly, it will either move one unit towards the fly or stay where it is.
- If the spider and the fly land in the same position at the end of a period, then the spider captures the fly and the process terminates.
- Spider's objective: to capture the fly in minimum expected time.

Example

State: distance between spider and fly.

A stochastic shortest path problem with states $0, 1, ..., n$;
n is the initial distance; 0 is the termination state.

p_{ij} the transition probabilities for $i \geq 2$

$p_{1j}(M)$ and $p_{1j}(\overline{M})$ the transition probabilities from state 1 to state j if the spider moves and does not move

Example

State: distance between spider and fly.

A stochastic shortest path problem with states $0, 1, \ldots, n$;
n is the initial distance; 0 is the termination state.

p_{ij} the transition probabilities for $i \geq 2$

$p_{1j}(M)$ and $p_{1j}(\overline{M})$ the transition probabilities from state 1 to state j if the spider moves and does not move

$$p_{ii} = p, \quad p_{i(i-1)} = 1 - 2p, \quad p_{i(i-2)} = p, \qquad i \geq 2,$$
$$p_{11}(M) = 2p, \quad p_{10}(M) = 1 - 2p,$$
$$p_{12}(\overline{M}) = p, \quad p_{11}(\overline{M}) = 1 - 2p, \quad p_{10}(\overline{M}) = p,$$

with all other transition probabilities being 0.

Example

Bellman's equation

- $J^*(0) = 0$

Example

Bellman's equation

- $J^*(0) = 0$
- $J^*(1) = 1 + \min[2pJ^*(1), pJ^*(2) + (1 - 2p)J^*(1)]$

Example

Bellman's equation

- $J^*(0) = 0$
- $J^*(1) = 1 + \min[2pJ^*(1), pJ^*(2) + (1-2p)J^*(1)]$
- $J^*(i) = 1 + pJ^*(i) + (1-2p)J^*(i-1) + pJ^*(i-2), \quad i \geq 2$

Example

Bellman's equation

- $J^*(0) = 0$
- $J^*(1) = 1 + \min[2pJ^*(1), pJ^*(2) + (1-2p)J^*(1)]$
- $J^*(i) = 1 + pJ^*(i) + (1-2p)J^*(i-1) + pJ^*(i-2), \quad i \geq 2$

Eq. for $i = 2$:

$$J^*(2) = \frac{1}{1-p} + \frac{(1-2p)J^*(1)}{1-p}.$$

Example

Bellman's equation

- $J^*(0) = 0$
- $J^*(1) = 1 + \min[2pJ^*(1), pJ^*(2) + (1-2p)J^*(1)]$
- $J^*(i) = 1 + pJ^*(i) + (1-2p)J^*(i-1) + pJ^*(i-2), \quad i \geq 2$

Eq. for $i = 2$:

$$J^*(2) = \frac{1}{1-p} + \frac{(1-2p)J^*(1)}{1-p}.$$

Combining with $i = 1$,

$$J^*(1) = 1 + \min\left[2pJ^*(1), \frac{p}{1-p} + \frac{p(1-2p)J^*(1)}{1-p} + (1-2p)J^*(1)\right],$$

or equivalently,

$$J^*(1) = 1 + \min\left[2pJ^*(1), \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p}\right].$$

Example

Two cases where

$$J^*(1) = 1 + 2pJ^*(1),$$

$$2pJ^*(1) \leq \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p},$$

and

$$J^*(1) = 1 + \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p},$$

$$2pJ^*(1) \geq \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p}.$$

Case 1) $J^*(1) = 1/(1-2p)$, and using the second eq., we find that this solution is valid when

$$\frac{2p}{1-2p} \leq \frac{p}{1-p} + \frac{1}{1-p},$$

or equivalently (after some calculation), $p \leq 1/3$. Thus for $p \leq 1/3$, it is optimal for the spider to move when it is one unit away from the fly.

Example

Case 2) $J^*(1) = 1/p$, when

$$2 \geq \frac{p}{1-p} + \frac{1-2p}{p(1-p)},$$

or equivalently, $p \geq 1/3$. Thus, for $p \geq 1/3$ it is optimal for the spider not to move when it is one unit way from the fly.

The minimal expected number of steps for capture when the spider is one unit away from the fly:

$$J^*(1) = \begin{cases} 1/(1-2p) & \text{if } p \leq 1/3, \\ 1/p & \text{if } p \geq 1/3. \end{cases}$$

Given the value of $J^*(1)$, we can calculate $J^*(i)$, $i = 2, ..., n$.

Example: The Blackmailer's Dilemma

There are two states, state 1 and the destination state t.

At state 1, we can choose a control u with $0 < u \leq 1$, while incurring a cost $-u$; we then move to state t with probability u^2, and stay in state 1 with probability $1 - u^2$.

Interpretation: u is a demand made by a blackmailer, state 1 the situation where the victim complies, and state t the situation where the victim refuses. The blackmailer tries to maximize his total gain by balancing his desire for increased demands with keeping his victim compliant.

Note: every stationary policy is proper.

Example: The Blackmailer's Dilemma

For any stationary policy μ with $\mu(1) = u$, we have

$$J_\mu(1) = -u + (1 - u^2) J_\mu(1)$$

from which

$$J_\mu(1) = -\frac{1}{u}.$$

Since u can be taken arbitrarily close to 0, it follows that $J^*(1) = -\infty$, but there is no stationary policy that achieves the optimal cost.

Example: The Blackmailer's Dilemma

Bellman's equation,

$$J^*(1) = (TJ^*)(1) = \min_{u \in (0,1]} [-u + (1 - u^2)J^*(1)],$$

has no (real number) solution.

The equation cannot have a solution with $J^*(1) \geq 0$, since then $u = 1$ attains the minimum leading to a contradiction, and it cannot have a solution with $J^*(1) < 0$, since then the minimizing value of u is

$$u = min\left[1, -\frac{1}{2J^*(1)}\right],$$

and by substitution, we have

$$J(1) = (TJ^*)(1) = \begin{cases} -1 & \text{if } J^*(1) \geq -1/2, \\ J^*(1) + \frac{1}{4J^*(1)} & \text{if } J^*(1) \leq -1/2, \end{cases}$$

a contradiction.

Example: The Blackmailer's Dilemma

There is an optimal *nonstationary* policy $\pi = \{\mu_0, \mu_1, ...\}$ that applies $\mu_k(1) = \gamma/(k+1)$ at time k and state 1, where $\gamma \in (0, 1/2)$.

One can show that $J_\pi(1) = -\infty$.

The blackmailer requests diminishing amounts over time, which nonetheless add to ∞.

However, the probability of the victim's refusal diminishes at a much faster rate over time, and as a result, the probability of the victim remaining compliant forever is strictly positive, leading to an infinite total expected payoff to the blackmailer.