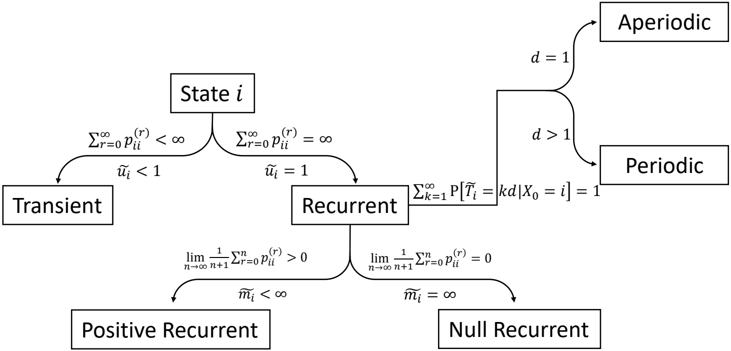
CE 273
Markov Decision Processes

Lecture 11

Infinite Horizon Total Cost MDPs

Previously on Markov Decision Processes

- Accessibility ($i \to j$)
- Communicating ($i \leftrightarrow j$)
- Communicating Class (All states communicate and the set is maximal)
- Closed Communicating Class ('Blackhole')
- Irreducibility (The entire DTMC is a 'blackhole')


State i
$\sum_{r=0}^{\infty} p_{ii}^{(r)} < \infty$
$\widetilde{u}_i < 1$
Transient
$\sum_{r=0}^{\infty} p_{ii}^{(r)} = \infty$
$\widetilde{u}_i = 1$
Recurrent
$\sum_{k=1}^{\infty} \mathrm{P}\left[\widetilde{T}_i = kd \mid X_0 = i\right] = 1$
$d = 1$
Aperiodic
$d > 1$
Periodic
$\lim_{n\to\infty} \frac{1}{n+1} \sum_{r=0}^{n} p_{ii}^{(r)} > 0$
$\widetilde{m}_i < \infty$
Positive Recurrent
$\lim_{n\to\infty} \frac{1}{n+1} \sum_{r=0}^{n} p_{ii}^{(r)} = 0$
$\widetilde{m}_i = \infty$
Null Recurrent

Previously on Markov Decision Processes

Recall that state spaces of reducible DTMCs can be partitioned as

$$S = C_1 \cup C_2 \cup \ldots \cup C_k \cup \mathcal{C}$$

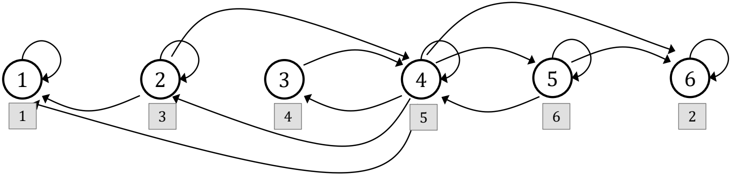
Let us first renumber states such that $i \in C_r$ and $j \in C_s$ with $r < s$ implies $i < j$. Further, $i \in C_r$ and $j \in \mathcal{C}$ implies $i < j$. Then, the transition matrix can be written in the following format

$$\begin{bmatrix} P(1) & 0 & \ldots & 0 & 0 \\ 0 & P(2) & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & P(k) & 0 \\ & D & & & Q \end{bmatrix}$$

where $P(1), \ldots, P(k)$ are the transition matrices of the k irreducible classes. Q is a $|\mathcal{C}| \times |\mathcal{C}|$ sub-stochastic matrix (row sums are ≤ 1, why?) and D is a $|\mathcal{C}| \times |S \backslash \{\mathcal{C}\}|$ matrix. We know from earlier analysis limiting distribution of $P(r)^{(n)}$. Since states in $\mathcal{C}$ are transient, one can show that $Q^{(n)} \to 0$.

Previously on Markov Decision Processes

States 1 and 6 form closed communicating classes.



The D and Q matrices are shown in blue and green respectively.

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1/16 & 1/16 & 1/4 & 1/8 & 1/4 & 1/4 \\ 0 & 1/4 & 0 & 0 & 1/4 & 1/2 \end{bmatrix} \end{array}$$

Previously on Markov Decision Processes

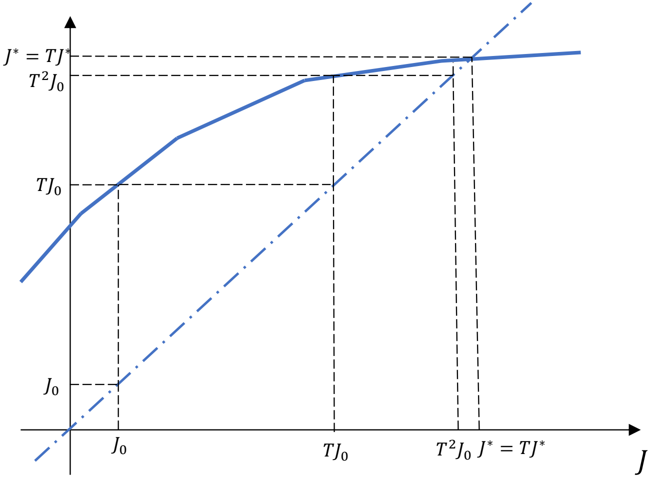


Figure: Value Iteration

Previously on Markov Decision Processes

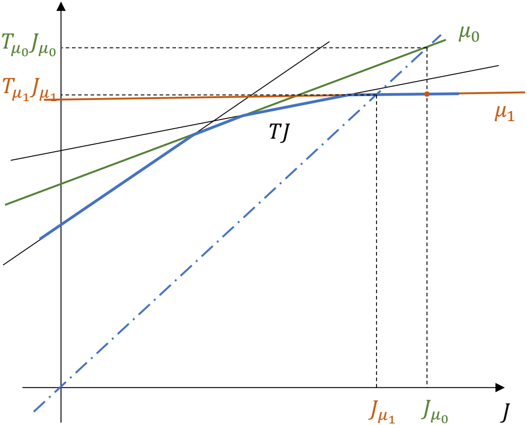
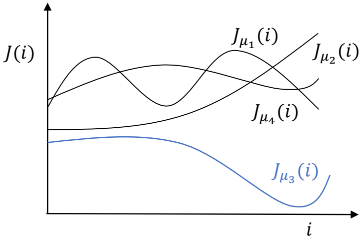
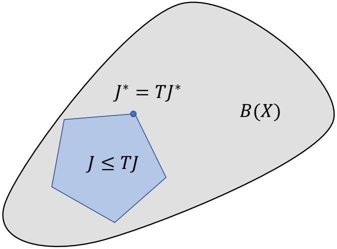


Figure: Policy Iteration

Previously on Markov Decision Processes

Note that the new optimization model has a maximization objective but we are still minimizing the total expected discounted cost.





The minimization objective we had earlier was across the set of policies (a finite set when the states and actions are finite). The LP on the other hand operates in the space of value functions.

Lecture Outline

1. Motivation
2. Stochastic Shortest Paths
3. Proper Policies

Lecture Outline

Motivation

Motivation

Discount Factors

The discount factor α ensured that the total cost was bounded and also guaranteed that the value and policy iteration algorithms converge. (How?)

In many problems such as asset management, periodic maintenance, and inventory control using a discount factor $\alpha < 1$ makes sense since we can connect it to interest rates.

However, in several other problems that do not warrant discounting (e.g., queuing systems, game playing), you will find it being used primarily because of the mathematical properties.

It is also seen as a tunable parameter and is tweaked to get good results. (One can also always set α to 0.9999999999999999999999999!)

Motivation

Discount Factors

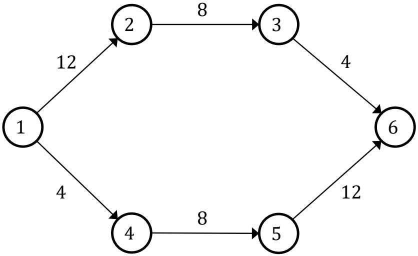
Even when costs are not involved, the notion of discounting may hold water since it is human nature to place more weight on short-term costs/rewards.

But discounting doesn't always make sense. Let's see why.

Motivation

Shortest Path Example I

Consider a deterministic shortest path problem. What is the shortest path between 1 and 6?



Suppose now travel times/costs are discounted by a factor 0.5. What is the optimal discounted path?

Motivation

Shortest Path Example I

Even if travel times were converted to costs using value of time, this doesn't represent money that you can hold.

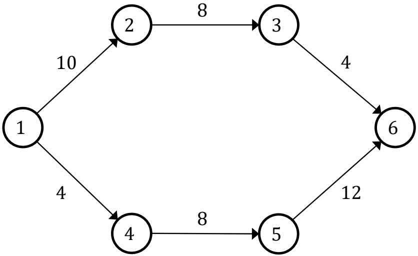
Note that stages are defined by the number of links in different paths and hence some travel times on some paths can be heavily discounted. This is however fixable. (How?)

One could still argue that the disutility of one minute of travel is different when you are closer vs. farther from the destination and the discount factor may help capture such effects.

Motivation

Shortest Path Example II

Consider a modified revision shown below.

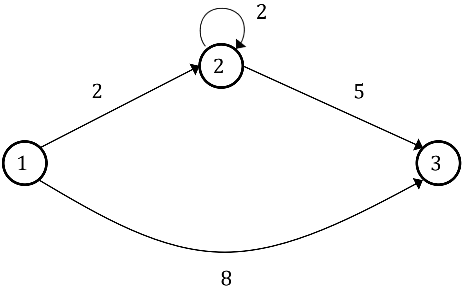


What are the undiscounted and discounted shortest paths? Assume a discount factor of 0.5.

Motivation

Shortest Path Example III

Here's a more compelling example:



With a discount factor of say $\alpha = 0.5$, it is always optimal to cycle one more time and we'd never reach the destination!

Motivation

Remedy

In cases like these, one can instead redefine the objective using the

- Total undiscounted cost ($\alpha = 1$)
- Average cost per stage

We will first look at total cost problems. However, even with bounded one-step costs $|g(i, u)| < M$, it is possible that the objective is unbounded.

So the problem makes sense when there are zero-cost termination stages. This feature is present in all optimal stopping problems (e.g., selling an asset, secretary problem) and stochastic shortest paths.

We will use a specific version of stochastic shortest paths as a working example throughout, but the results hold for any total cost MDP that satisfies the assumptions we'll make.

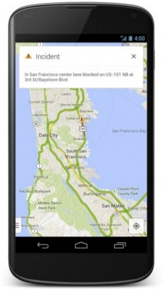
Lecture Outline

Stochastic Shortest Paths

Stochastic Shortest Paths

Introduction

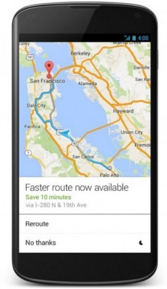
Imagine you are traveling along route A. Say your app informs you of an incident downstream and suggests an alternate route B.



4:00

Incident ×

In San Francisco center lane blocked on US-101 NB at 3rd St/Bayshore Blvd



4:00

Faster route now available

Save 10 minutes

via I-280 N & 19th Ave

Reroute

No thanks

How can we provide such en route navigation?

Stochastic Shortest Paths

Introduction

When travel times on links in a network are not deterministic, we can fit probability distributions from past data.

These distributions can be time-dependent, i.e., they can be different depending on when we arrive at each intermediate node.

For simplicity, assume that they are not. Assume that when we reach a node in the network, the travel time on the downstream arcs is not known until we traverse it. What is the optimal policy to go from point A to B?

Just replace travel times with expected travel times and use any shortest path algorithm.

Stochastic Shortest Paths

Introduction

However, when we arrive at a node, the levels on congestion on immediate downstream links is fairly predictable.

We can also update our distributions of what happens two or more hops later. But for the purpose of illustration, suppose that when we reach a node, we know the travel times on downstream nodes with certainty.

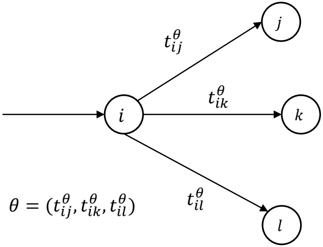
The travel times on all the other links are still random with known probability distributions. How would you model this as a total cost MDP?

Stochastic Shortest Paths

State and Action Space

Suppose the graph is represented by $G = (N, A)$, where N is the set of nodes and A is the set of links/arcs.

Upon arriving at a node i, a traveler observes a information vector $\theta \in \Theta_i$ drawn with probability q^θ informing him or her of the travel time of each link leaving node i.



Thus, the states are tuples (i, θ). Policies are functions $\mu(i, \theta)$ which tell us which node to go to next. Note that this is a problem with uncontrollable state components like Tetris.

Stochastic Shortest Paths

Value Functions and Policies

Suppose t represents the destination node and let $\Gamma(i)$ represent the nodes adjacent to node i. Thus, one can expect the value iteration algorithm to proceed as

$$J_{k+1}(i,\theta) = \min_{j \in \Gamma(i)} \left\{ t_{ij}^{\theta} + \sum_{\theta' \in \Theta_j} q^{\theta'} J_k(j,\theta') \right\}$$

and $J_k(t,\theta) = 0$ for all $\theta \in \Theta_t$.

Likewise, given a policy μ, we expect the cost of the policy μ to be a solution to

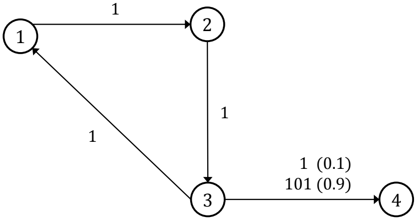
$$J_\mu(i,\theta) = t_{i,\mu(i,\theta)}^{\theta} + \sum_{\theta' \in \Theta_j} q^{\theta'} J_\mu(\mu(i,\theta),\theta')$$

and $J_\mu(t,\theta) = 0$ for all $\theta \in \Theta_t$. Since we have uncontrollable state components, can you write analogues of these equations using the ex ante value functions? Do we need any assumptions on μ?

Stochastic Shortest Paths

Example

What is the optimal policy in the network shown below. The origin and destination are 1 and 4 respectively.



Consider the policy: Take arc (3,4) only if its cost is 1, else return to node 3 via nodes 1 and 2. The expected cost of the policy can be written as a arithmetico geometric sequence.

$$3(0.1) + 6(0.9)(0.1) + 9(0.9)^2(0.1) + \ldots = 30$$

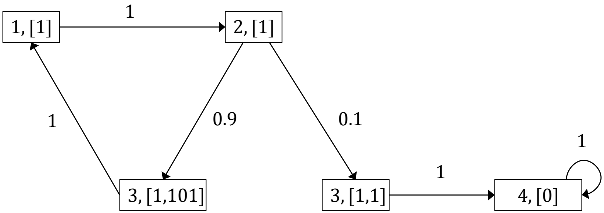
Cycling occurs here but because of the Markovian assumption (not to be confused with the cycling example due to discounting). This is also referred to as the full-reset case.

Stochastic Shortest Paths

Markov Chain Associated With a Policy

For any given policy, we can construct a Markov chain using the states (i, θ) where $i \in N$, $\theta \in \Theta_i$.

The transition diagram for the previous example looks as follows. Once, we reach the destination, we remain there forever. Is this Markov Chain reducible/irreducible?



We can thus partition the state space into a closed communicating class containing the state $(4, [0])$ and a set of transient states (that contains all the remaining states).

Stochastic Shortest Paths

Markov Chain Associated With a Policy

The transition matrix can be written as

$$P_\mu = \begin{array}{c|ccccc} & 4,[0] & 1,[1] & 2,[1] & 3,[1,1] & 3,[1,101] \\ \hline 4,[0] & 1 & 0 & 0 & 0 & 0 \\ 1,[1] & 0 & 0 & 1 & 0 & 0 \\ 2,[1] & 0 & 0 & 0 & 0.1 & 0.9 \\ 3,[1,1] & 1 & 0 & 0 & 0 & 0 \\ 3,[1,101] & 0 & 1 & 0 & 0 & 0 \end{array}$$

We will soon redefine the transition matrices of total cost MDPs to include only the green sub-matrix and evaluate the cost of the policy using $(I - P_\mu)^{-1} g_\mu$.

$$\left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 30 \\ 29 \\ 1 \\ 31 \end{bmatrix}$$

Note that the row sums of these matrices need not sum to 1!

Stochastic Shortest Paths

Ex ante Value Functions and Policies

We can simplify the problem by defining ex ante value functions (this the value at a node before we observe the information vector) using

$$\hat{J}(i) = \sum_{\theta \in \Theta_i} q^\theta J(i, \theta)$$

Thus, one can hypothesize that the value iteration algorithm looks like

$$\hat{J}_{k+1}(i) = \sum_{\theta \in \Theta_i} q^\theta \min_{j \in \Gamma(i)} \left\{ t_{ij}^\theta + \hat{J}_k(j) \right\}$$

with $\hat{J}_k(t) = 0$ for all k. The problem is easier to solve because we have a smaller number of states.

Let's say this algorithm converges. How do we find the optimal policies from $\hat{J}^*$?

$$\mu^*(i, \theta) \in \arg \min_{j \in \Gamma(i)} \left\{ t_{ij}^\theta + \hat{J}^*(j) \right\}$$

Stochastic Shortest Paths

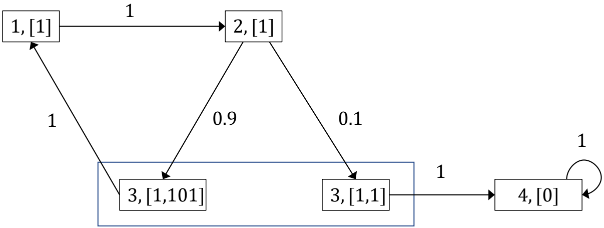
Ex ante Value Functions

Likewise, given a policy μ, we expect the ex ante cost of the policy μ to be a solution to

$$\hat{J}_\mu(i) = \sum_{\theta \in \Theta_i} q^\theta \left\{ t^\theta_{i,\mu(i,\theta)} + \hat{J}_\mu \left(\mu(i,\theta)\right) \right\}$$

and $\hat{J}_\mu(t) = 0$. How do the Markov chains look like when we deal with the ex ante value functions?

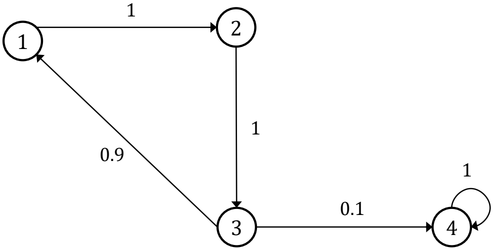
The transition diagram will have nodes as states. This is equivalent to aggregating states in the previous Markov chain



Stochastic Shortest Paths

Aggregated Markov Chains

The aggregated transition diagram (with states as nodes) looks as follows:



The policies we've seen so far are deterministic. Another way to look at policies for the reduced state space is to think of them as the probability with which we pick different downstream links at each node.

Stochastic Shortest Paths

Aggregated Markov Chains

The transition matrix can be written as

$$P_\mu = \begin{array}{c} \\ 4 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 4 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.1 & 0.9 & 0 & 0 \end{bmatrix} \end{array}$$

Again, the transition matrices of total cost MDP will be assumed to include only the green sub-matrix and we evaluate the cost of the policy using $(I - P_\mu)^{-1} g_\mu$.

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.9 & 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 0.9(1) + 0.1(1) \end{bmatrix} = \begin{bmatrix} 30 \\ 29 \\ 28 \end{bmatrix}$$

Lecture Outline

Proper Policies

Proper Policies

Introduction

We can formally prove the VI and PI ideas discussed so far. But we do need some restrictions on the policies and a new definition of the transition matrix.

Recall that $\alpha = 1$ hence we cannot use Banach fixed point theorem. In fact, one can define a new norm and use some contraction mapping results. But we will take a simpler route.

Let the state space be $X = \{1, 2, \ldots, n, t\}$ where t represents a termination state. Let as before, $p_{ij}(u)$ represent the probability of reaching state j when u is chosen in state i. We further assume that

- The terminal state is absorbing, i.e., $p_{tt}(u) = 1, \forall\, u \in U(t)$.
- The terminal state is cost-free, i.e., $g(t, u) = 0 \,\forall\, u \in U(t)$.

Proper Policies

Introduction

Define the T and T_μ mappings similarly but only for the set of states $\{1, 2, \ldots, n\}$. The cost of starting from the terminal state is zero.

$$(TJ)(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j) \right\} \forall\, i = 1, 2, \ldots, n$$

$$(T_\mu J)(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j) \,\forall\, i = 1, 2, \ldots, n$$

Does the monotonicity lemma hold for T and T_μ? Yes.

Proper Policies

Introduction

As before, for a given policy μ, we can write the one-step transition probability matrix as

$$P_\mu = \begin{pmatrix} p_{11}(\mu(1)) & \dots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \dots & p_{nn}(\mu(n)) \end{pmatrix}$$

and the cost vector for a fixed policy μ as

$$g_\mu = \begin{pmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{pmatrix}$$

Thus, the T-mu operator in matrix form can be written as

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

Notice that there is no α in the above expression.

Proper Policies

Introduction

Suppose, J_0 represents a vector of zeros. What happens when we apply the T_μ repeatedly? We would accumulate the one-step costs and hence get the total cost of associated Markov chain.

$$\begin{aligned} T_\mu J_0 &= g_\mu \\ T_\mu^2 J_0 &= g_\mu + P_\mu g_\mu \\ T_\mu^3 J_0 &= g_\mu + P_\mu g_\mu + P_\mu^2 g_\mu \\ &\vdots \end{aligned}$$

$$J_\mu = \lim_{N\to\infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu$$

We will soon extend this by proving that we get J_μ by applying T_μ repeatedly on **any** initial guess J_0.

Proper Policies

Definitions

Definition

A policy μ is proper if there is a positive probability that the terminal state will be reached after at most k stages, starting from any initial state, i.e.,

$$\max_{i=1,\ldots,n} \mathbb{P}\left[x_k \neq t \middle| x_0 = i, \mu\right] < 1$$

A simpler way to think of proper policies is that $i \to t$ for all $i = 1, \ldots, n$ in the Markov chain associated with μ.

Thus, the terminal state will be reached w.p. 1 under a proper policy. Any policy that is not proper is said to be an improper policy.

Proper Policies

Assumptions

We make two main assumptions for the analysis of total cost MDPs:

Assumption 1: There exists at least one proper policy

Assumption 2: For all improper policies μ, $J_\mu(i)$ is ∞ for at least one i

Since, $J_\mu = \lim_{N\to\infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu$, the second assumption implies that some component of $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ diverges to ∞ as $N \to \infty$.

For stochastic shortest paths, the above conditions are met if the destination is reachable from all nodes and the link travel times are positive.

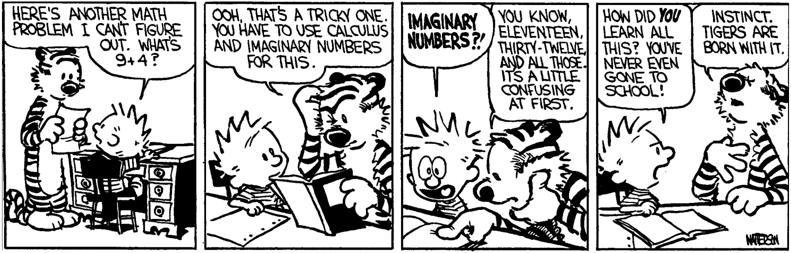
Additional Reading

Waller, S. T., & Ziliaskopoulos, A. K. (2002). On the online shortest path problem with limited arc cost dependencies. Networks, 40(4), 216-227.

Provan, J. S. (2003). A polynomial-time algorithm to find shortest paths with recourse. Networks, 41(2), 115-125.

Boyles, S. D., & Rambha, T. (2016). A note on detecting unbounded instances of the online shortest path problem. Networks, 67(4), 270-276.

Your Moment of Zen


HERE'S ANOTHER MATH PROBLEM I CAN'T FIGURE OUT. WHAT'S 9+4?
OOH, THAT'S A TRICKY ONE. YOU HAVE TO USE CALCULUS AND IMAGINARY NUMBERS FOR THIS.
IMAGINARY NUMBERS?!
YOU KNOW, ELEVENTEEN, THIRTY-TWELVE, AND ALL THOSE. IT'S A LITTLE CONFUSING AT FIRST.
HOW DID YOU LEARN ALL THIS? YOU'VE NEVER EVEN GONE TO SCHOOL!
INSTINCT. TIGERS ARE BORN WITH IT.
WATERSON