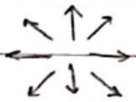
سوال ۱:

الف) ص) هر MDP را می‌توان بصورت یک expectimax دید که هر استیت ها یک نود max است و نود chance ها را هم یک نود حس می‌کنیم که در آنجا با توجه به احتمالات هر استیت به یک استیت جدید می‌رویم ولی یک تفاوت آن با expectimax این است که utility ها هستند که در درخت expectimax روی برگ ها بود ولی در اینجا رفتن از هر جفت به یک استیت یک utility دارد که باید در نظر گرفته شود و همچنین چون مانند MDP به استیت های تکراری می‌رویم ممکن است هیچ وقت به برگ نرسیم ولی در کل برای محاسبه کردن optimal policy می‌توان یک دور درخت expectimax را طی کرد

ب) ص) با توجه به مدل های MDP می‌دانیم دلیل هر استیت فقط وابسته به احتمال رفتن به آن استیت از استیت قبلی و ریوارد رفتن به آن از استیت قبلی است در نتیجه فقط استیت های قبلی هستند که در حل مسئله MDP مهم می‌شوند

ج) ص) regret مجموع منهای discounted reward ها از زمان شروع یادگیری تا آخر کار است به علاوه‌ی حالتی که از اول optimal policy را داشتیم چون ما در اول کار مشغول یادگیری هستیم و استیت های بدی را امتحان می‌کنیم تا policy بهینه را پیدا کنیم reward های منفی هم می‌گیریم تا زمانی که کم کم پالیسی خوب را پیدا کنیم در نتیجه این regret بزرگتر باشد یعنی دیرتر به پالیسی بهینه رسیده‌ایم چون در طول یادگیری reward ها کم هستند.

سوال ۲:

Action : چون زاویه‌های ماز ۴۵° هستند. action ها را این می‌گیریم که ایجنت ما در یکی از جهت های  یک قدم به سمت جلو حرکت کند.

State : ایجنت ما یک سنسور دارد که وضعیت محیط را به ما می‌دهد. هر استیت را وضعیت داده شده توسط ایجنت با استفاده از آن سنسورها در نظر می‌گیریم که بسته به اینکه کجای ماز باشد و محیط اطراف آن چه شکلی باشد متفاوت است. ترمینال استیت های ما دیوار ها و خروجی ماز است

Reward : چون می‌خواهیم با کمترین میزان برخورد به دیوارها بیرون رود برای دیوار ها reward منفی می‌گذاریم و خروجی ماز هم یک reward مثبت

s.a.m

Scanned with CamScanner