پیش‌نوشت : لطفا توجه فرمایید که علاوه بر حل مسئله ، روش استدلال و فکر کردن شما بر روی مسئله و نزدیک شدن به راه حل نیز اهمیت زیادی برای ما دارد. در صورت نیاز به فرض اضافه برای حل مسئله به صورت مستدل آن فرضیات را مشخص و استفاده نمایید.

سوال‌ها

۱. یک عامل باید از طریق یادگیری ،تقویتی کنترل سه آسانسور موازی را در یک ساختمان اداری پرتردد بهینه کند. منظور از کنترل آسانسور تعیین طبقات توقف آسانسورها قبل از رسیدن به طبقات تعیین شده توسط مسافران داخل آسانسور و همچنین تعیین طبقاتی است که آسانسور خالی در آنها منتظر مسافر بماند. جزای عامل میزان معطل شدن هر مسافر و پاداش آن تعداد مسافر حمل شده در هر ساعت است. رستوران اداره در طبق همکف است و پارکینگها در 1- و 2- و مدیریت و اتاقهای جلسات در طبقه آخر قرار دارد. حالت و عمل را برای عامل تعریف کنید و روش کدینگ مناسب را برای ابعاد پیوسته حالت (در صورت وجود) مشخص کنید.

۲. عاملی با یک مسئله یادگیری تک حالته و تک قدمه با عمل پیوسته $a \in [a_1, a_2]$ مواجه است. عامل میداند که دو عملِ هم ارزش و بهینه برای این مسئله وجود دارد. یک روش یادگیری مناسب برای این عامل توسعه دهید. شبه کد روش را ارایه و پشتوانه ریاضی آنرا بیان نمایید.

۳. رباتی در هر حالت یک MDP گسسته باید به n>1 ماهیچه خود فرمان دهد $(f_1, f_2, \dots, f_n)$ اما آنچه باعث حرکت و تغییر حالت ربات میشود تابع حاصل جمع این فرمانها است

($a = G(f_1, f_2, \dots, f_n)$) و پاداش بیرونی صرفا به ازای تغییر حالت تحت اثر a به عامل داده میشود. به این سیستم Redundant in actuation گفته میشود. عامل یک تابع پاداش درونی نیز دارد که تابع حالت فعلی و مقادیر $f_1, f_2, \dots, f_n$ است. یک روشی یادگیری برای حداکثر کردن حاصل جمع این دو پاداش ارایه دهید.

۴. میخواهیم ترکیب ایده Thompson Sampling و SARSA را در یک MDP با حالت پیوسته و اعمال گسسته استفاده کنیم. ریاضیات و شبه کد مربوطه را توسعه دهید.

۵. در مدل ما از یک MDP گسسته در $P^a_{s,\acute{s}}$ خطا وجود دارد و خطا در هر $P^a_{s,\acute{s}}$ بین ۰ و ۱ است. راه حل مناسب برای یافتن بهترین سیاست در این MDP چیست؟ لازم به ذکر است که ما به مسئله واقعی دسترسی نداریم و صرفا باید از این مدل استفاده کنیم.