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. در هنگام راه رفتن سریع در هر زمان یک پا بروی زمین (در فاز support) و یک پا در حال حرکت در هوا (در فاز swing) است. به پایی که برروی زمین است Support leg و به پای دیگر swing leg میگوییم. با یک ساده سازی نسبتا خوب میتوان فرض کرد که انرژی لازم برای راه رفتن در هر قدم را در لحظه بلند شدن support leg از روی زمین به صورت یک فشار موج مربعی با عرض کم از طریق قسمت جلوی کف پا (پنجه پا) تامین میکنیم و پس از آن support leg وارد فاز swing میشود و در این فاز پا نیاز به دریافت انرژی جدید ندارد. برای یک ربات دو پا، تعیین عرض، زمان شروع و ارتفاع این موج مربعی اصلی ترین مسئله کنترلی است. این مسئله را به صورت یک مسئله یادگیری تقویتی صورت بندی کنید (حالت اعمال و کدینگ آن، تابع پاداش را تعیین کنید) و یک الگوریتم مناسب برای یادگیری آن پیشنهاد دهید.

۲. در یک n-armed-bandit تعداد اعمال زیاد است اما عامل میداند که توزیع پاداش برخی از اعمال به یکدیگر بسیار نزدیک است اما این اعمال را از قبل نمی شناسد روشی برای استفاده از این اطلاعات برای کاهش پشیمانی (regret) ارایه دهید. توجه کنید که حافظه عامل محدود است.

۳. دیده شده است که در برخی مسایل میزان پشیمانی روش Sarsa از روش Q-learning پایینتر است. دلیل این امر چه مسایلی میتواند باشد؟ آیا شرایطی وجود دارد که on-policy MC از Sarsa بهتر عمل کند؟

۴. میخواهیم ترکیب ایده UCB و on-policy MC را در یک MDP با حالت و اعمال گسسته استفاده کنیم. ریاضیات و شبه کد مربوطه را توسعه دهید.

۵. در یک MDP با حالت پیوسته و عمل گسسته، عامل از یک کدینگ RBF برای یادگیری Sarsa متکی بر Discounted return استفاده میکند. پس از پایان یادگیری عامل میخواهد با یک کدینگ

Fourier یادگیری را متکی بر Average reward تکرار کند. روشی برای استفاده از نتیجه یادگیری اول در یادگیری دوم ارایه دهید.

۶. در یک MDP عامل $P^{a}_{s,ś}$ را میداند و علاوه بر پاداشی که از نقاد (critic) بیرونی میگیرد یک نقاد دورنی نیز دارد. عامل از $R^{a}_{s,ś}$ نقاد درونی آگاه است اما $R^{a}_{s,ś}$ نقاد بیرونی را نمیداند. روشی برای یافتن سیاست بهینه

a. متکی بر Discounted Return

b. متکی بر Average Return

ارائه دهید. پشتوانه ریاضی روش خود را در قالب توسعه معادله بلمن بیان کنید.