بسمه تعالی

دانشکده مهندسی

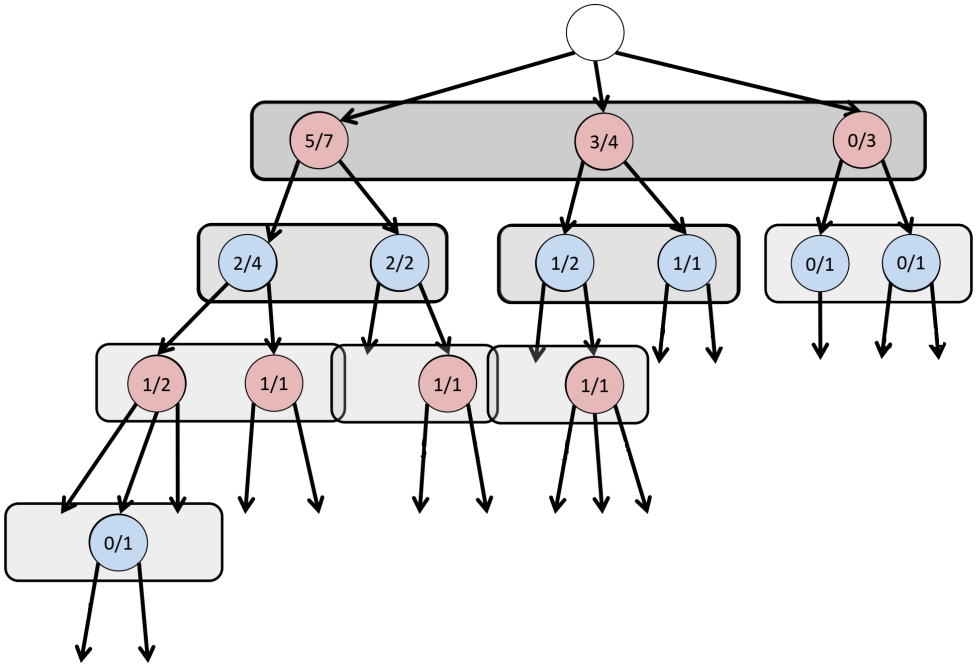
یادگیری تقویتی

نیم‌سال دوم ۱۴۰۱-۱۴۰۲

دکتر احد هراتی

تمرین دوم نظری | زمان تحویل: امتحان پایان‌ترم

۱ جست‌وجوی درختی مونت کارلو



شکل ۱: نمایی از یک وضعیت اجرای الگوریتم درخت جست‌وجوی مونت کارلو

در شکل ۱ درخت یک بازی دو نفره را مشاهده می‌کنید که هر لایه نشانگر نوبت یکی از بازیکنان است. در هر گره نسبت $w_{k,1}$ به n_k ($w_{k,1}/n_k$) نوشته شده است که در آن n_k بیانگر تعداد بازی‌هایی است که شامل حالت k بوده‌اند و $w_{k,p}$ تعداد بازی‌هایی است که توسط بازیکن p برده شده و شامل حالت k بوده است. روش جست‌وجوی درختی مونت کارلو از ۴ مرحله تشکیل یافته است:

۱. درون گره‌هایی که تا به حال ذخیره کرده است پایین می‌رود و بر اساس یک معیار مانند UCB[1] یکی را انتخاب می‌کند.

۲. گره انتخاب شده را گسترش داده و فرزند ساخته شده را به گره‌های ذخیره شده اضافه می‌کند.

[1] upper confidence bound

۳. در گره تازه ساخته شده، ادامه‌ی بازی را تا زمان رسیدن به نتیجه‌ی Δ شبیه‌سازی می‌کند. در مثال ما، اگر ما برديم $\Delta = +۱$ و اگر باختیم یا مساوی کردیم $\Delta = ۰$.

۴. با نتیجه‌ی شبیه‌سازی داده شده‌ی Δ بروزرسانی درخت انجام می‌شود. یعنی برای تمام گره‌های k که جزو اجداد گرهی هستند که برای آن شبیه‌سازی را انجام داده‌ایم $w_{k,۱}^{new} = w_{k,۱}^{old} + \Delta$ و $n_k^{new} = n_k^{old} + ۱$.

الف) با فرض

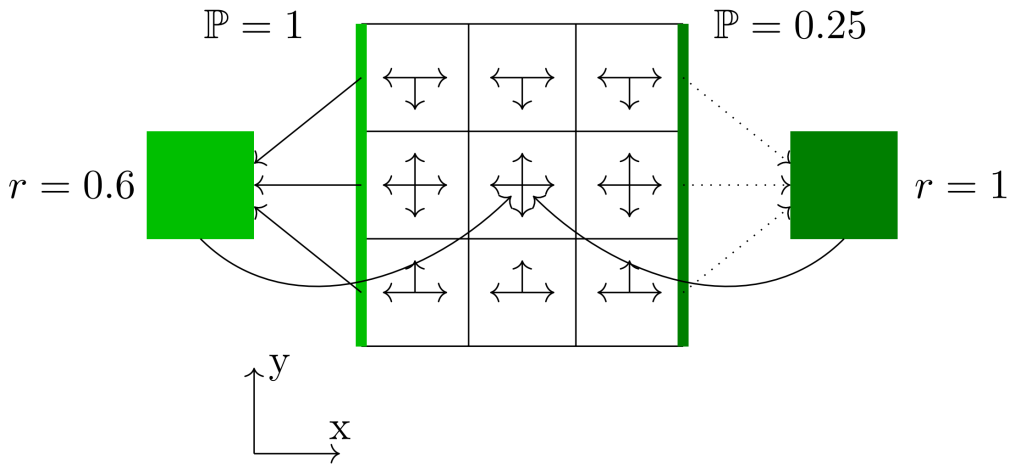
$$UCB۱(k,p) = E[win|k,p] + c\sqrt{\frac{۲\ln(n_{parent(k)})}{n_k}} \approx \frac{w_{k,p}}{n_k} + c\sqrt{\frac{۲\ln(n_{parent(k)})}{n_k}}$$

مشخص کنید در درخت شکل ۱ کدام گره برای گسترش انتخاب می‌شود. (c را برابر ۲ در نظر بگیرید)

ب) اگر نتیجه‌ی شبیه‌سازی برای گره انتخاب شده‌ی بخش قبل $\Delta = +۱$ شود، پس از بروزرسانی، درخت جدید را رسم کنید. کدام گره برای تکرار بعدی الگوریتم جست‌وجوی درختی مونت کارلو انتخاب خواهد شد؟

۲ انتخاب ویژگی

MDP شکل ۲ با ۱۱ حالت و ۴ عمل را در نظر بگیرید. هر یک از حالات مربع ۳×۳ وسط بوسیله‌ی یک دوتایی (x,y) که $x \in \{۰,۱,۲\}$ و $y \in \{۰,۱,۲\}$ می‌باشند بازنمایی می‌شود. کارگزار از حالت مرکزی $(۱,۱)$ شروع کرده و در هر گام یکی از ۴ عمل مربوط به رفتن در جهت بالا، پایین، چپ، و راست را انتخاب می‌کند. هر عمل منجر به انتقال قطعی به خانه‌ی مجاور می‌شود به استثناء مواردی که تلاش در جهت خارج شدن از نقشه است که در این صورت کارگزار سر جای خود باقی می‌ماند. در سمت چپ، کارگزار با انتخاب عمل چپ رفتن به طور قطعی وارد حالتی می‌شود که در آن انتخاب هرگونه عملی متناظر با دریافت پاداش ۰/۶ و انتقال به خانه‌ی مرکزی است. در سمت راست مربع، کارگزار با انتخاب عمل راست رفتن به احتمال ۲۵٪ وارد حالتی می‌شود که در آن انتخاب هرگونه عملی متناظر با دریافت پاداش ۱ و انتقال به خانه‌ی مرکزی است. پاداش در مابقی حالت‌ها برابر صفر است.



شکل ۲: بازنمایی MDP نیازمند انتخاب ویژگی

الف) مقدار پاداش میانگین[۲] را به ازای دو سیاست فقط راست رفتن و فقط چپ رفتن بدست آورید. سیاست بهینه در این MDP چیست؟(ضریب تنزیل $\gamma = ۱$)

[۲] average reward

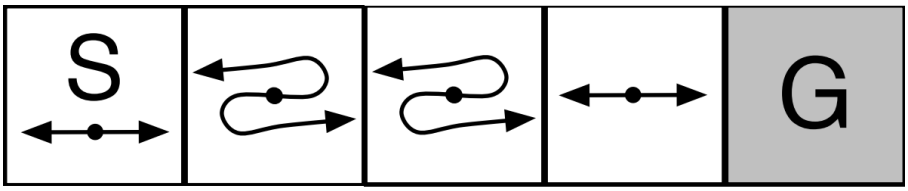
حال تصور کنید که اطلاعات محدودی برای این MDP جمع‌آوری گشته است و فقط یک نمونه تجربه‌ی $< s, a, r, s' >$ برای هر زوج $< s, a >$ داریم.

ب) احتمال اینکه حداقل یکی از گذارهای سمت راست، دسترسی قطعی به پاداش $r = ۱$ را وانمود کند چقدر است؟ در چنین موردی آیا رویکردهای مبتنی بر مدل یا بدون مدل می‌توانند بدون اجازه‌ی تعامل بیشتر با محیط سیاست بهینه را بیابند؟

ج) تابع تقریب خطی‌ای را پیشنهاد دهید که الگوریتم‌های یادگیری تقویتی بتوانند با استفاده از آن و فقط با داشتن اطلاعات محدود مذکور سیاست بهینه را محاسبه کنند.

۳ گرادیان سیاست

دنیای جدولی راهروی کوچک شکل ۳ را در نظر بگیرید. پاداش در هر گام برابر -۱ هست به غیر از خانه‌ی پایانی G که پاداش در آن صفر است. در هرکدام از چهار حالت غیرپایانی فقط دو عمل وجود دارد، راست و چپ. این اعمال در حالت‌های اول و چهارم پیامدهای معمول خود را دارند(در حالت اول چپ رفتن منجر به هیچ حرکتی نمی‌شود)، اما در حالت‌های دوم و سوم برعکس می‌شوند یعنی عمل راست به چپ می‌رود و عمل چپ به راست. بردار ویژگی $\mathbf{x}$ را برای تمام حالت‌های s تعریف می‌کنیم: $\mathbf{x}(s, \text{راست}) = [۱, ۰]^\intercal$ و $\mathbf{x}(s, \text{چپ}) = [۰, ۱]^\intercal$. برای پاسخ‌دهی به سوالات ضریب تنزیل را برابر ۱ بگیرید.



شکل ۳: دنیای جدولی راهروی کوچک

الف) از دانش خود درباره‌ی دنیای جدولی و پویایی آن استفاده کنید تا به عبارت پارامتری دقیقی برای احتمال بهینه‌ی انتخاب عمل راست برسید.

ب) سیاست بهینه‌ی بدست آمده را با سیاست εـحریصانه‌ای که با احتمال $۱ - \varepsilon/۲$ در تمام گام‌ها عمل راست را انتخاب می‌کند و همچنین سیاستی که با احتمال مشابه در تمام گام‌ها عمل چپ را انتخاب می‌کند مقایسه کنید.

ج) اکنون فرض کنید که در حالت G انتخاب هرگونه عملی منجر به انتقال به حالت S شده و پاداش $+۲۰$ دریافت می‌کند. در نتیجه MDP پیش‌رو تبدیل به یک وظیفه‌ی ادامه‌دار می‌شود. حال، تابع ارزش تفاضلی[۳] را برای حالت S و به ازای پیروی از سیاست بهینه حساب کنید.

د) سه مورد از مزیت‌های پارامتری‌سازی سیاست بر پارامتری‌سازی ارزش‌ـعمل را بیان کنید.

[۳] differential value