Slide 1

Dobrý den, jmenuji se Jakub Rada a téma mojí bakalářské práce je porovnávání metod explorace v částečně pozorovatelných stochastických hrách.

Slide 2

Stochastické hry a částečně pozorovatelné stochastické hry jsou formální modely dvouhráčových her s nulovým součtem se simultánními tahy a různou mírou informovanosti hráčů. Algoritmy pro přibližné řešení těchto typů her existují, ale většinou využívají lineární programování, které způsobuje špatnou škálovatelnost pro velmi velké instance. Cílem mojí práce bylo nahradit lineární programování banditími algoritmy a porovnat je mezi sebou. Nejprve jsem se zaměřil na stochastické hry s plnou informací, které jsou podmnožinou částečně pozorovatelných her, a následně jsem bandity porovnal i na obecnějším modelu ze zadání.

Slide 3

Nejprve krátký úvod do mnohorukých banditů. Narozdíl od lineárního programování, které v případě her pokaždé hledá optimální strategii v daný moment, mnohorucí bandité v každém kole vybírají jednu akci, kterou zahrají.

Stochastičtí bandité používají Q funkci, která reprezentuje pozorovanou kvalitu jednotlivých akcí a ovlivňuje následující výběry.

Navíc, pro doménu stochastických her, kde hráč má informaci o akcích hraných oponentem, jsem přidal ještě takzvané pozorovatelné bandity, kteří si udržují průměrnou strategii oponenta a to jim umožňuje lepší výběr dobrých akcí.

Naopak, adversariální banditi neudržují kvalitu jednotlivých akcí, ale upravují si rozdělení pravděpodobnosti přes tyto akce, pomocí které pak akce náhodně vybírají.

Slide 4

Stochastické hry jsou jednodušší z dvou testovaných modelů. Hráči jsou umístěni do stochastického prostředí, přechází mezi stavy a získávají odměnu. Hráč 1 se snaží maximalizovat celkovou získanou odměnu, jeho protivník ji minimalizuje. Hráči pozorují svůj i protivníkův stav a všechny již zahrané akce.

Konkrétně ve hře Tag se hráči pohybují bludištěm a hráč 1 se snaží co nejdříve trefit laserem hráče 2, který se snaží uniknout.

Slide 5

Stochastické hry se typicky řeší algoritmem value iteration, který hledá pro každý stav očekávanou odměnu hráče 1, pokud by hra začínala v tomto stavu. Iterativní aplikací Bellmanova operátoru se zlepšuje aproximace neznámé očekávané hodnoty.

Ve stochastických hrách odpovídá aplikaci Bellmanova operátoru vyřešení maticové hry u, která se řeší lineárním programováním. Tuto maticovou hru lze nahradit banditím algoritmem, který vybere akci, zkusí jí zahrát a výsledná hodnota je použita místo řešení z lineárního programu. Jelikož banditím algoritmus vybere vždy jen jednu akci, ale optimální strategie maticové hry nemusí být jen jedna akce, tak jako novou hodnotu value funkce použijeme průměr těchto výsledků.

Slide 6

Ve výsledcích se objevují dva typy situací. V první je optimální hrát pouze jednu stejnou akci pokaždé. Tato situace není tak zajímavá, protože většina banditů rychle dokonverguje ke správné hodnotě.

Druhý typ situace je zajímavější pro porovnání a jedna z takových situací je vidět v grafu. Osa x určuje počet iterací algoritmu v logaritmickém měřítku a osa y znamená odchylku aproximované hodnoty od pravé hodnoty, kterou nalezla originální verze value iteration s lineárním programováním. Jde o situaci, kdy není žádná jedna akce nejlepší a hráč musí vždy vybírat alespoň ze dvou různých a v tomto typu stavu mají někteří bandité problémy.

Nejlepší je konzistentně Exp3 společně s pozorovatelnou variantou UCB bandity, normálnímu UCB se nepodaří od určité hodnoty konvergovat dál. Nejhůře na tom pak jsou Best of N a Successive elimination, kteří málokdy dokonvergují blízko k správné hodnotě.

Slide 7

Jako další jsem testoval bandity na částečně pozorovatelných stochastických hrách a speciálně na jejich podtřídě jednostranných her.

Narozdíl od předchozího modelu hráč 1 nemá plnou informaci. Nemůže pozorovat stav sebe ani protihráče a ani akce zahrané protihráčem. Místo toho dostává pozorování, která mu alespoň nějakou informaci dávají. Aby mohl odhadovat, ve kterém stavu se hra nachází, udržuje si pravděpodobnostní rozdělení, kterému se říká belief.

Konkrétně na příkladu hry Pursuit-evasion, hráče 1 reprezentují černé tečky a hráče 2, který má plnou informaci jako ve stochastických hrách, reprezentuje distribuce ve zbývajících stavech. V této hře se pursuer snaží dostat alespoň jednu ze svých jednotek na stejné políčko jako je evader.

Slide 8

HSVI pro tento typ her je založený na podobné myšlence jako stejně pojmenovaný algoritmus pro částečně pozorovatelné markovské procesy.

Místo toho se metoda používá dvě funkce, jedna omezující opravdovou value funkce zespoda, druhá shora. Opět se iterativně aplikuje Bellmanův operátor, tentokrát na obě meze, a to je přibližuje k sobě. V momentě, kdy jsou meze v inicálním beliefu blíže, než předem daný práh ϵ, algoritmus končí a získali jsme přibližnou hodnotu hry.

Bandité opět nahrazují lineární program, který aplikuje Bellmanův operátor.

Slide 9

Tento graf ukazuje běh algoritmu na instanci hry Pursuit-evasion. Osa x reprezentuje počet updatů mezních funkcí, osa y ukazuje hodnotu mezí pro různé bandity. Obě meze se přibližují k pravé hodnotě vyznačené černou konstantní funkcí.

V horní mezi nedochází k velkému zlepšení, ale to je způsobeno i tím, že iniciální hodnota byla zvolena mnohem blíže optimální hodnotě než pro dolní mez. Z experimentů vychází, že nejlépe se chovají ϵ-greedy a Exp3 bandité, ostatní zaostávají. Celkově by se dalo řict, že bandité se silnou explorací přináší lepší výsledky.

Otázka