Abstract

In reinforcement learning, an agent interacts repeatedly with its environment by selecting an action and receiving a reward while the environment transits from the current state to the next one. The reward helps the agent to decide whether its action is profitable or not. During this interaction, the agent faces some interesting challenges. This dissertation investigates two challenges for the reinforcement learning agent acting in Markov decision processes (MDPs).

The first challenge is the trade-off between exploration and exploitation, where the agent has to decide between new action or current best action. We examine the first challenge by examining empirically well-known policies like ϵ-greedy and an online version of the knowledge gradient (KG) policy on a number of multi-armed bandit problems and then for a test bed of infinite horizon MDPs where now the action selection policy is incorporated into online-$LSPI$. Although the KG does not have parameters to be tuned, it performs as well as or even better than the other well-tuned action selection policies.

The second challenge is representation discovery in MDPs. We propose the shortest path Gaussian kernels basis functions defined on either the state or state-action graph. Using offline-$LSPI$, we empirically demonstrate that these shortest path basis functions outperform other basis functions, for instance the hand engineered basis functions. For the online-$LSPI$, we either experimented with basis functions derived for offline-$LSPI$ or a relevant subset. To avoid selecting a subset, we propose online kernel-based $LSPI$ with KG as action-selection policy. Here, the basis functions are kernels and new ones are added online by using the approximate linear dependency method.

In reinforcement leren interageert een agent voortdurend met zijn omgeving door middel van het kiezen van een actie, het verkrijgen van een beloning of straf terwijl de omgeving evolueert van toestand naar toestand. De beloning of straf helpt de agent te beslissen of een actie al dan niet nuttig is. De interactie met de omgeving levert een aantal interessante uitdagingen op. In deze dissertatie onderzoeken we twee uitdagingen voor een reinforcement lerende agent die ageert in een Markov beslissingsproces (of MDP).

De eerste uitdaging heeft te maken met de balans tussen exploratie en exploitatie: de agent moet kiezen tussen een nieuwe of de huidige beste actie. We onderzoeken deze uitdaging door welbekende keuzebeslissingsstrategieën zoals ϵ-'greedy' en een online versie van de kennisgradient empirisch te vergelijken op eerst een aantal veelarmige bandietproblemen en daarna op een testbatterij van MDPs met oneindige horizon waar nu de keuzebesliss-

ingsstrategie geïncorporeerd is in online LSPI . Alhoewel de kennisgradient geen parameters heeft die eerst afgesteld moeten worden is zijn performantie even goed zo niet better dan de andere goed afgestelde keuzebeslissingsstrategieën.

De tweede uitdaging is het ontwikkelen van representaties, ook basisfuncties genoemd, voor MDPs. We stellen hier kortste pad gaussiaanse basisfuncties voor die gedefinieerd zijn op ofwel de toestandsgraf of the toestand-actiegraf. We tonen empirisch aan dat deze basisfuncties anderen overtreffen die worden gecombineerd met offline LSPI en we hebben het gebruik van korste pad gaussiaanse basisfuncties ook mogelijk gemaakt voor online LSPI. In het geval van online LSPI, hebben we experimenten uitgevoerd met de basisfuncties die we hebben voorgesteld voor offline LSPI en met een relevante deelverzameling hiervan. Om het bepalen van deze deelverzameling te vermijden stellen we online LSPI met kernfuncties voor waar de kennisgradient wordt gebruikt als keuzebeslissingsstrategie. In dit geval zijn de basisfuncties kernfuncties en nieuwe kernfuncties worden online toegevoegd als basisfunctie door middel van de benaderende lineaire afhankelijkheidsmethode.