Value Iteration Optimisation

Es gibt eine Variation des Policy Iteration Algorithmus, welche die Policy für mehrere Schritte konstant hält und nur die Utility werte verändert.[1] (Zitiert nach[2]) Durch diese verbesserung wird die Policy gleichmäßiger der optimalen angenähert, jedoch wird das Optimum erst nach mehr Iterationen erreicht.

Weiterhin kann bei der Policy iteration die Berechnung der Utility Werte als Gleichungssystem gesehen werden. Es gibt eine Gleichung für jeden State, da für alle States ein Utility-Wert berechnet werden muss und in jeder Gleichung eventuell die möglichkeit in jeden anderen State zu kommen, weshalb in jeder Zeile all die anderen States als Variable vorkommen können.
Dieses Gleichungssystem kann mit dem Gauß-Jordan-Verfahren optimal gelößt werden, allerdings fürt das zu einer Komplexität von $O(n^3)$ (bei n Zuständen) was jedoch für große n schnell zu lange dauert. Deshalb kann man numerische Verfahren verwenden um die Lösung des Gleichungssystems zu approximieren. Dies Reduziert die Komplexität des Schrittes auf $O(n^2)$. Die reduzierte genauigkeit der Utility-Werte könnte jedoch dazu führen, dass etwas mehr Iterationen benötigt werden oder im Extremfall die optimale Policy nicht gefunden wird.[3]

[1] Martin L. Puterman and Moon Chirl Shin. Modified policy iteration algorithms for discounted Markov decision processes. Management Science, 24:1127-1137, 1978.

[2] https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/node21.html

[3] http://www.mit.edu/~dimitrib/Adaptive_aggr.pdf