Non-Deterministic Policy Improvement Stabilizes Approximated Reinforcement Learning

Wendelin Böhmer, Rong Guo and Klaus Obermayer
<{wendelin,rong,oby}@ni.tu-berlin.de>
Neural Information Processing Group — Technische Universität Berlin

PPR 1527

Abstract

This paper investigates a type of instability that is linked to the greedy policy improvement in approximated reinforcement learning. We show empirically that *non-deterministic policy improvement* can stabilize methods like *least-squares policy iteration* (LSPI, Lagoudakis and Parr, 2003) by controlling the improvements' stochasticity. Additionally we show that a suitable representation of the value function also stabilizes the solution to some degree. The presented approach is simple and should also be easily transferable to more sophisticated algorithms like *deep reinforcement learning*.

Non-Deterministic Policy Improvement

- replace greedy policy improvement

$$\Gamma_*[f|q](x) \;=\; f\Big(x, \operatorname*{argmax}_{a\in\mathcal{A}} q(x,a)\Big)$$

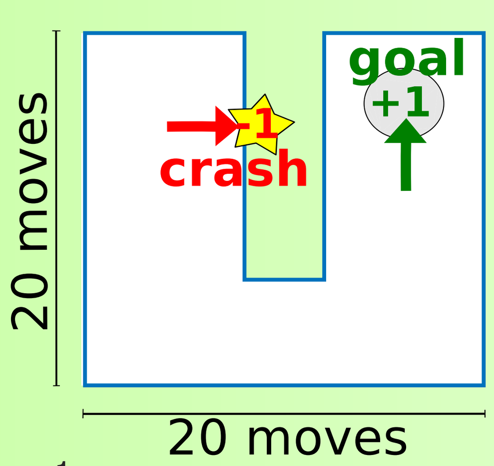
- with an improved non-deterministic policy

$$\Gamma_\beta[f|q](x) \;=\; \sum_{a\in\mathcal{A}} \pi^q_\beta(a|x)\, f(x,a)$$

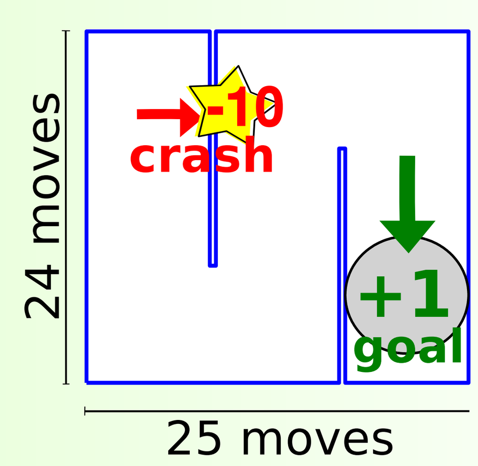
- e.g. the softmax with *inverse tempterature* β

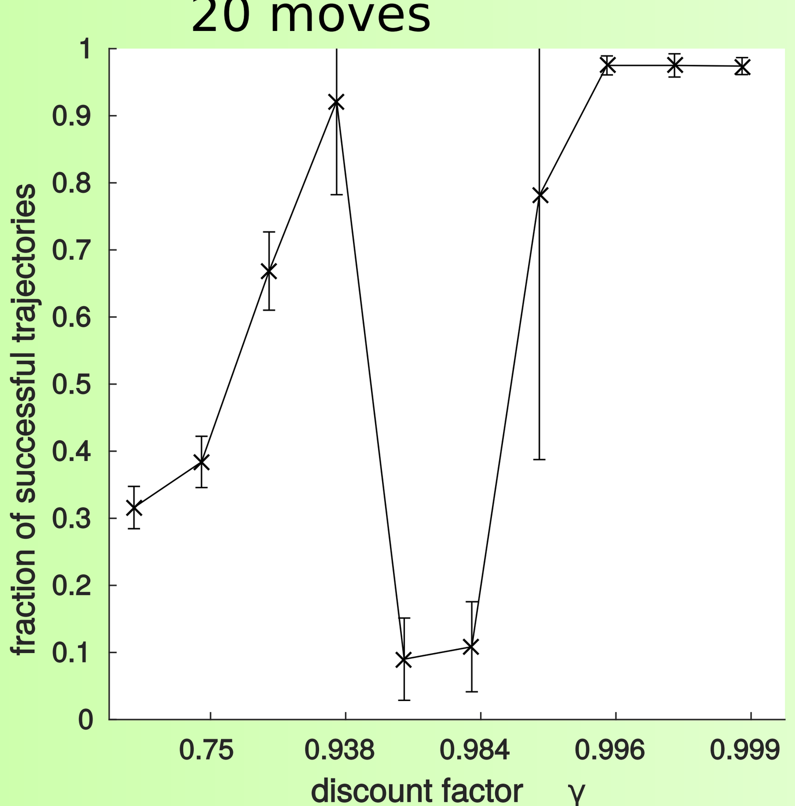
$$\pi^q_\beta(a|x) = \frac{\exp\big(\beta q(x,a)\big)}{\sum\limits_{a'\in\mathcal{A}} \exp\big(\beta q(x,a')\big)}$$

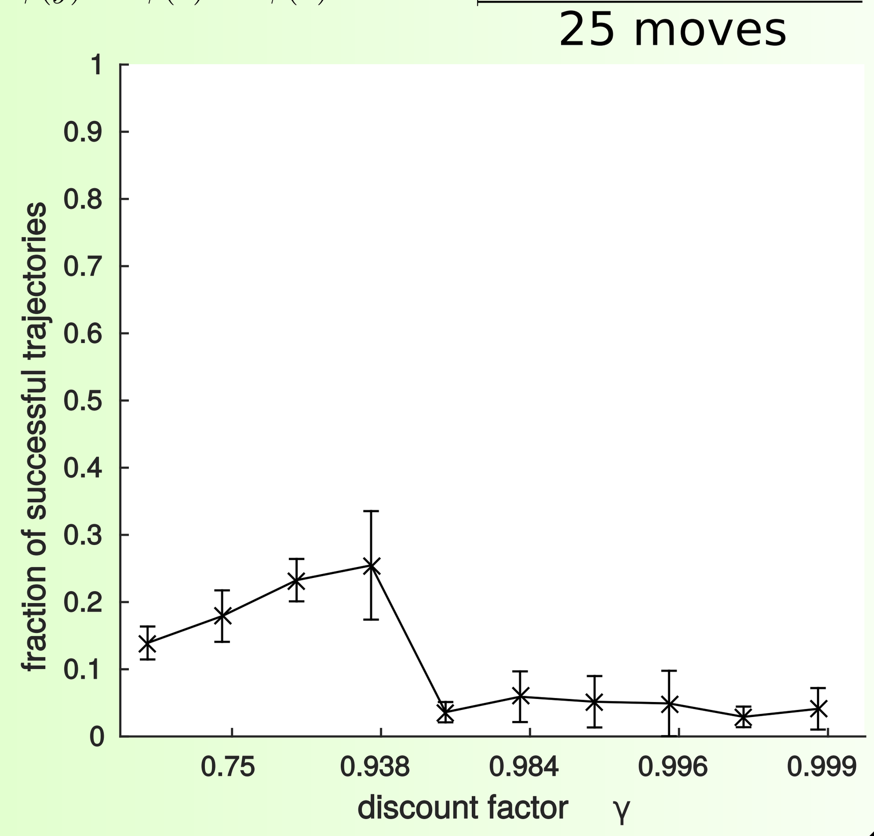
Approximated Batch PI is not Stable



- navigation experiment with LSPI
- randomwalk with 50,000 samples
- continuous state: $(x,y,\theta)\in\mathbb{R}^3$
- 3 discrete actions a:
  - move forward
  - turn left/right 45°
- 1500 Fourier bases:
  - $10\phi(x)\times 10\phi(y)\times 5\phi(\theta)\times 3\phi(a)$





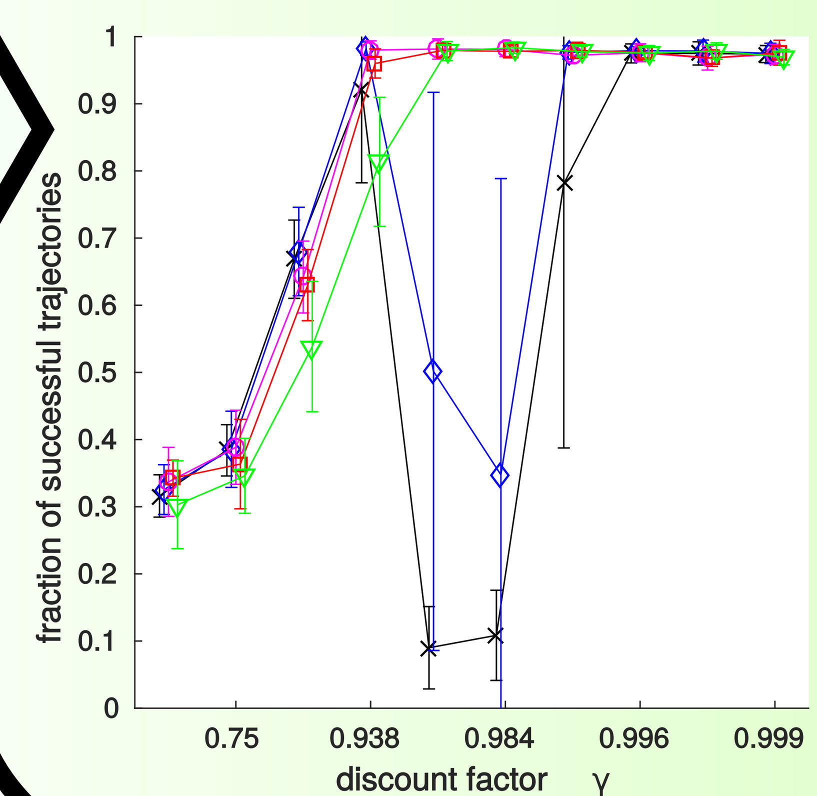


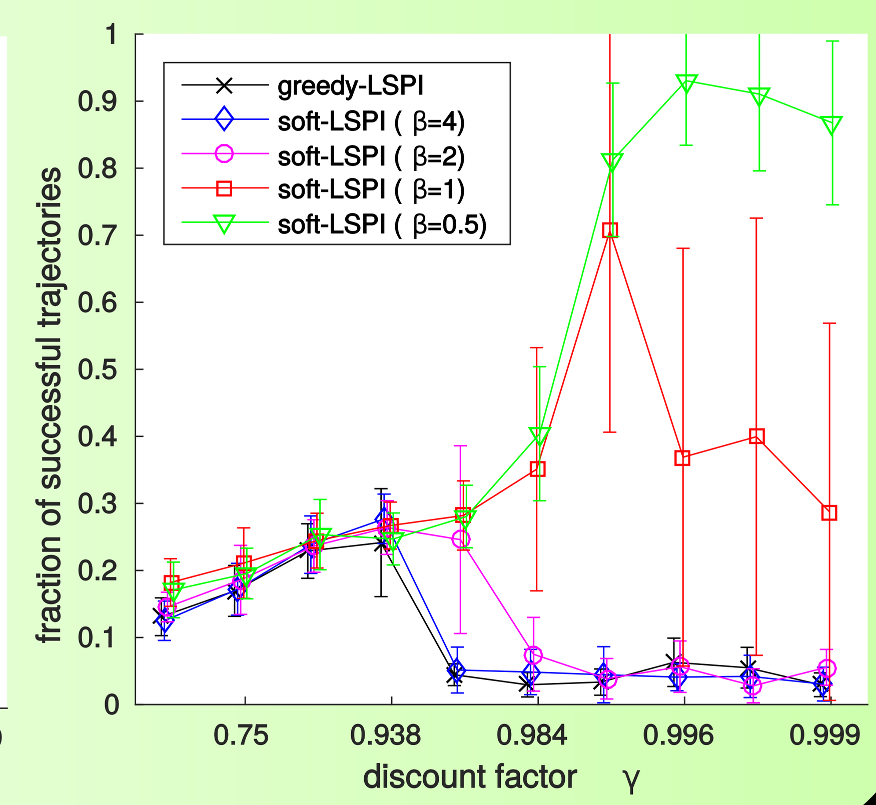
Stochasticity Stabilizes LSPI

- normalize Q-values for equal stoachsticity

$$\bar{q}(x,a) = \frac{q(x,a) - \langle q(x,\cdot)\rangle}{\sqrt{\langle q^2(x,\cdot)\rangle - \langle q(x,\cdot)\rangle^2}}$$

- policy improvement with constant stochasticity β



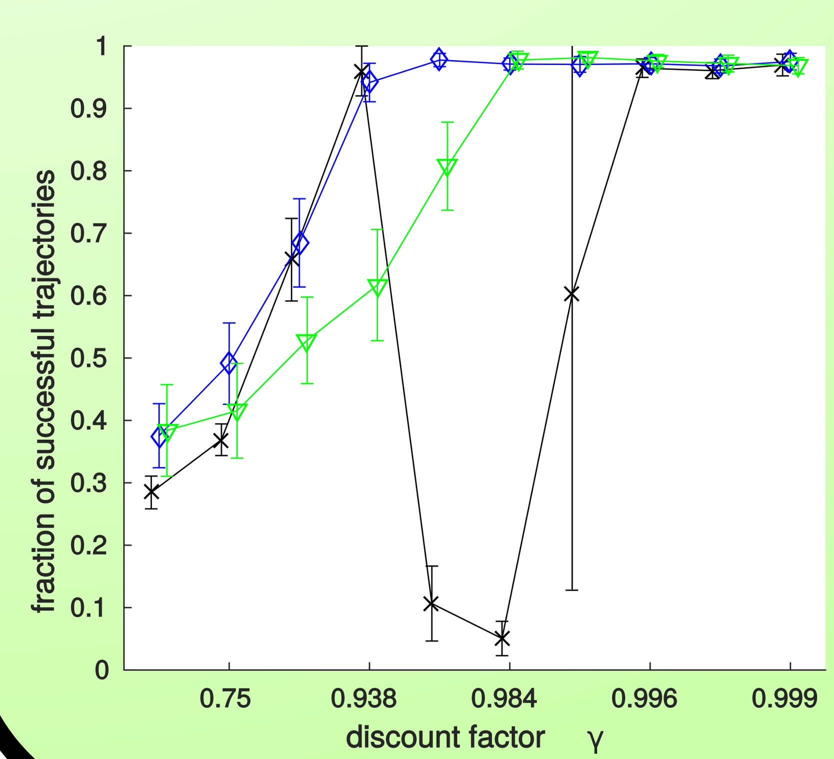


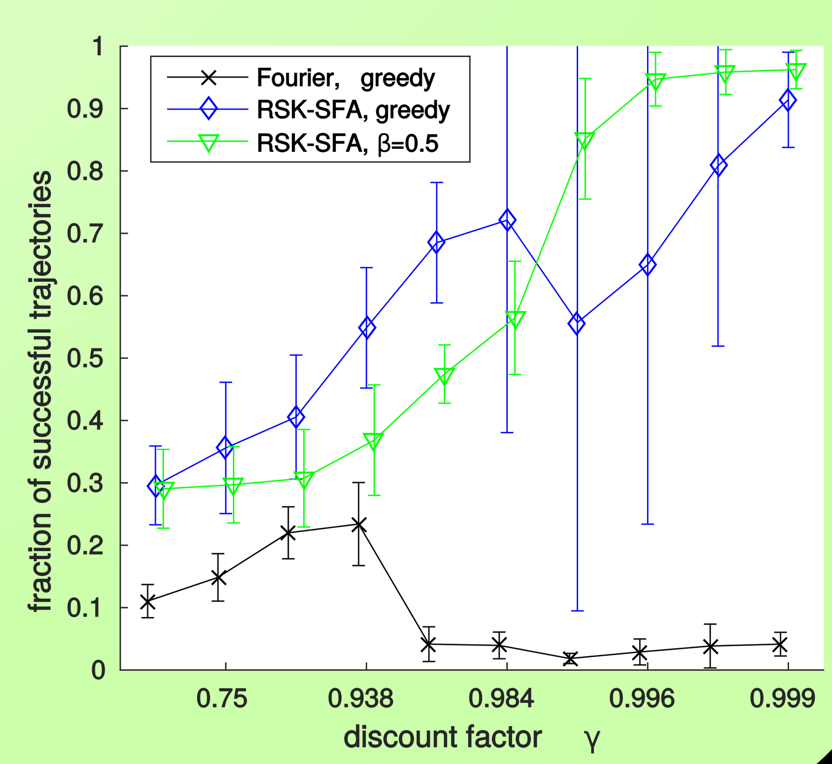
Conclusion

- approximated RL can become instable
- non-deterministic PI can stabilize solution
- even good representations must be stabilized
- proposed heuristic is easy to implement

Stabilization and Representation

- deep RL learns representations in lower layers
- good representations may influence stabilization
- learn rep. with non-linear SFA (Böhmer et al., 2012)
- SFA close to optimal for LSPI (Böhmer et al., 2013)





Acknowledgements

This work was funded by the *German science foundation* (DFG) within SPP 1527 "autonomous learning".