Follow your Nose: Using General Value Functions for Directed Exploration in Reinforcement Learning

Somjit Nath [1] Omkar Shelke [1] Durgesh Kalwar [1] Hardik Meisheri [1] Harshad Khadilkar [1, 2]

[1]TCS Research, Mumbai, India [2]IIT Bombay, India

Contribution

- We extend upon temporally extended version of the ϵ-greedy exploration strategy by using auxiliary task learning with the help of General Value Functions (GVF) to perform directed exploration thereby further improving state space coverage during exploration.
- This is generalized formulation to include domain knowledge about the environment by providing GVF cumulant which also improves latent representation.

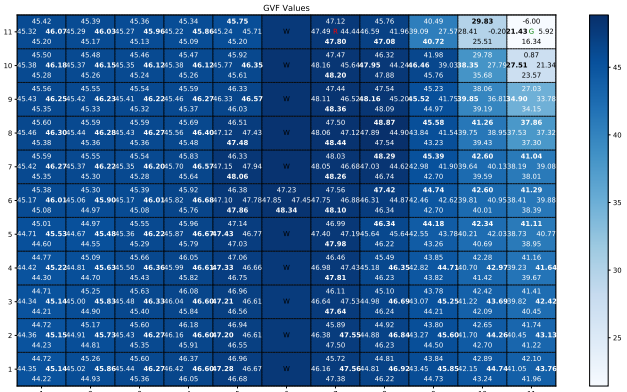
Pseudocode

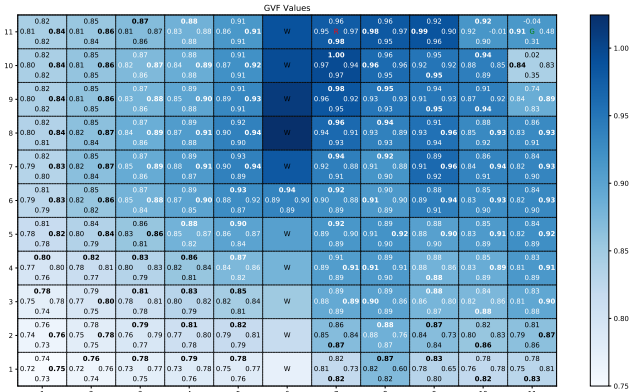
Function *DEZ-greedy*(ϵ, Z_{max}):
- $z \leftarrow 0$
- $w \leftarrow -1$
- $g \leftarrow 0$
- while *True* do
  - Observe state s
  - if $z == 0$ then
    - if $random() < \epsilon$ then
      - Sample duration: $z \sim [1, Z_{max}]$
      - Sample GVF: $g \sim [0, M]$
      - if $g == 0$ then
        - Sample action: $w \leftarrow U(A)$
        - $a \leftarrow w$
      - else
        - $a \leftarrow argmax(Q_g^{GVF})$
    - else
      - $a \leftarrow argmax(Q^{Main})$
  - else
    - if $g == 0$ then
      - $a \leftarrow w$
    - else
      - $a \leftarrow argmax(Q_g^{GVF})$
      - $z \leftarrow z - 1$
  - Take Action a

Z_{max} is maximum persistence value

Off-policy Divergence

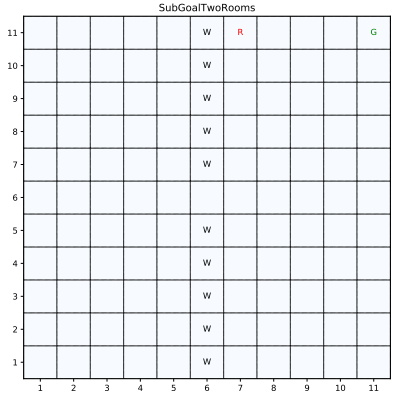
Off-policy divergence with generic GVF algorithms for 11x11 SubGoal Two Rooms. The heatmap shows Q values of the GVF that gets a reward of +1 on collecting the red dot. The values are much more bounded for DEZ-greedy exploration.



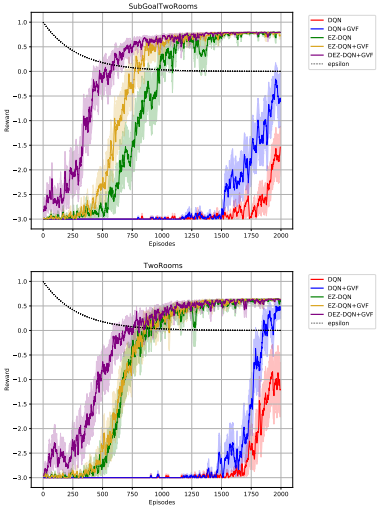


Environment

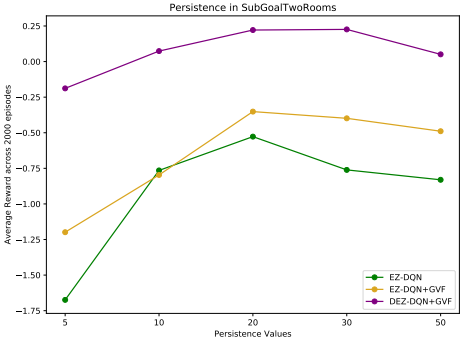
SubGoal Two Rooms Environment



Results



Senstivity to Z_{max}



tcs research