Masterprüfung

Increasing Horizon Policy Neural Networks for Finite-Horizon MDP’s

David Sutter

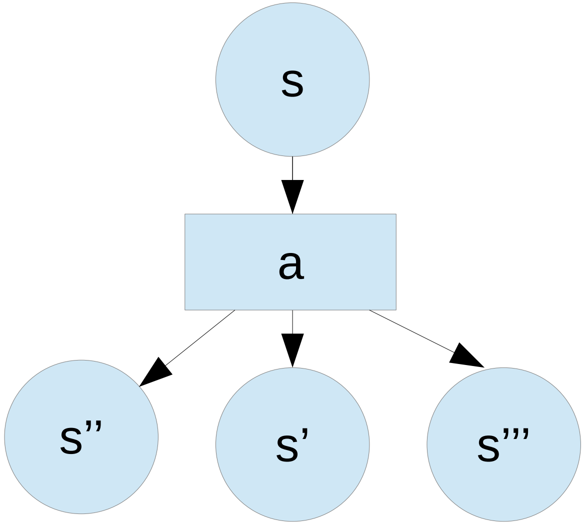
Outline

1. **Fundamentals**
   - Markov Decision Process
   - Neural Networks
2. **Intention**
   - General approach
   - Shortcommings
   - Policy NN
3. **Evaluation**
   - Results
   - Adjustments
   - Future work

Part 1 Markov Decision Process (MDP)

$$\mathcal{S} = \langle S, A, cost, T, s_0, S_* \rangle$$

$$\mathcal{T} = \langle S, A, P, R, s_0, H \rangle$$



Bellman Equation

- Bellman Equation:

$$V_\pi(s,d) = \begin{cases} R(s,\pi(s)) + \sum_{s' \in S} P(s'|s,\pi(s)) \cdot V_\pi(s',d-1)), & \text{if } d > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$Q_\pi(s,d,a) = R(s,a) + \sum_{s' \in S} P(s'|s,a) \cdot Q_\pi(s',d-1,\pi(s'))$$

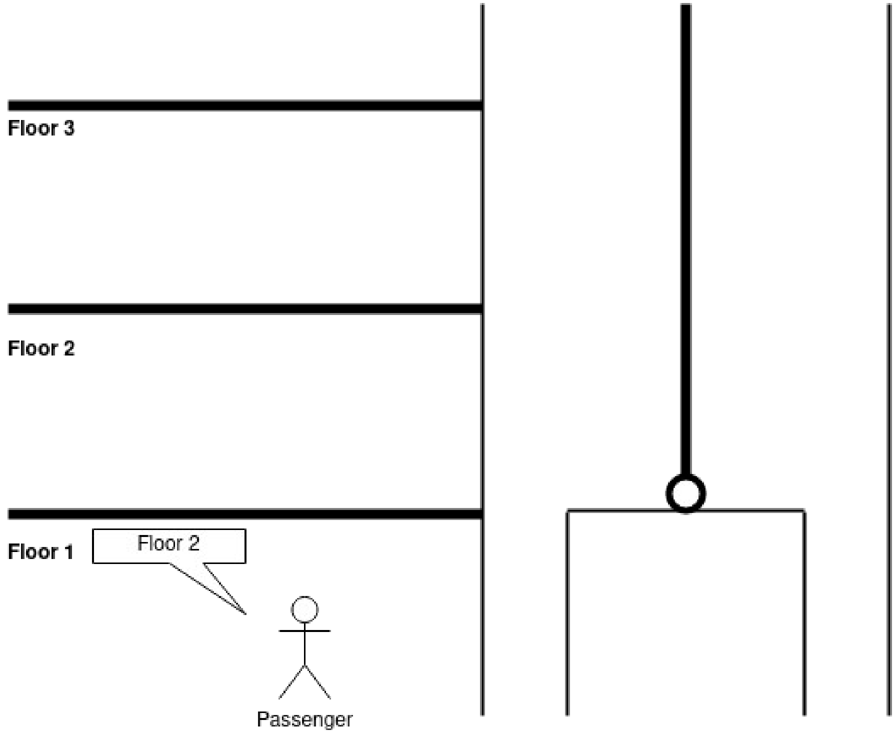
Bellman Equation

- Bellman Equation:

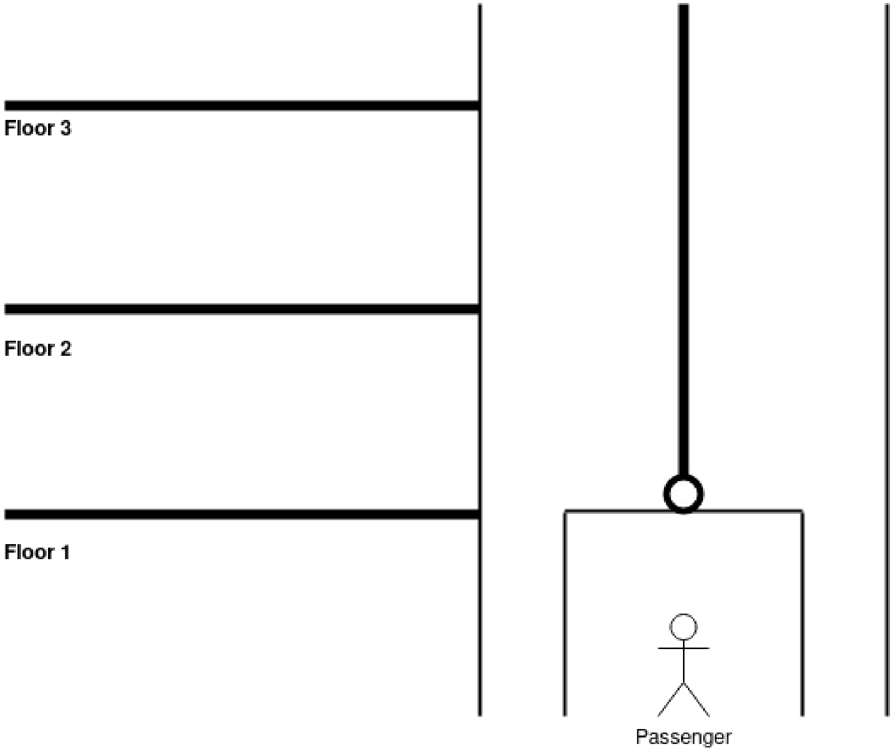
$$V_*(s,d) = \begin{cases} \max\limits_{a \in A} Q_*(s,d,a), & \text{if } d > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$Q_*(s,d,a) = R(s,a) + \sum_{s' \in S} P(s'|s,a) \cdot V_*(s',d-1)$$

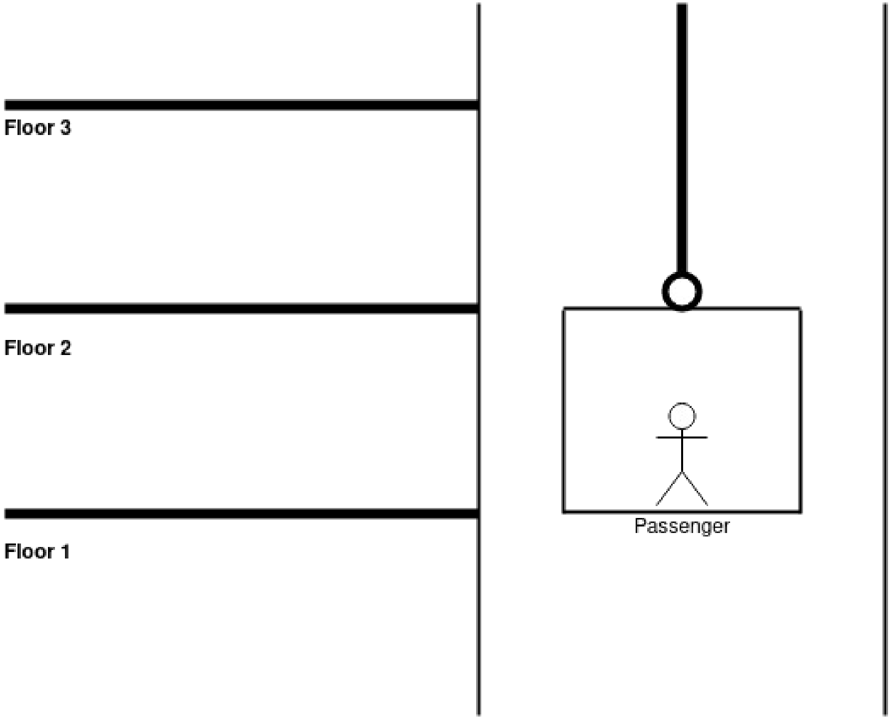
Elevators Domain



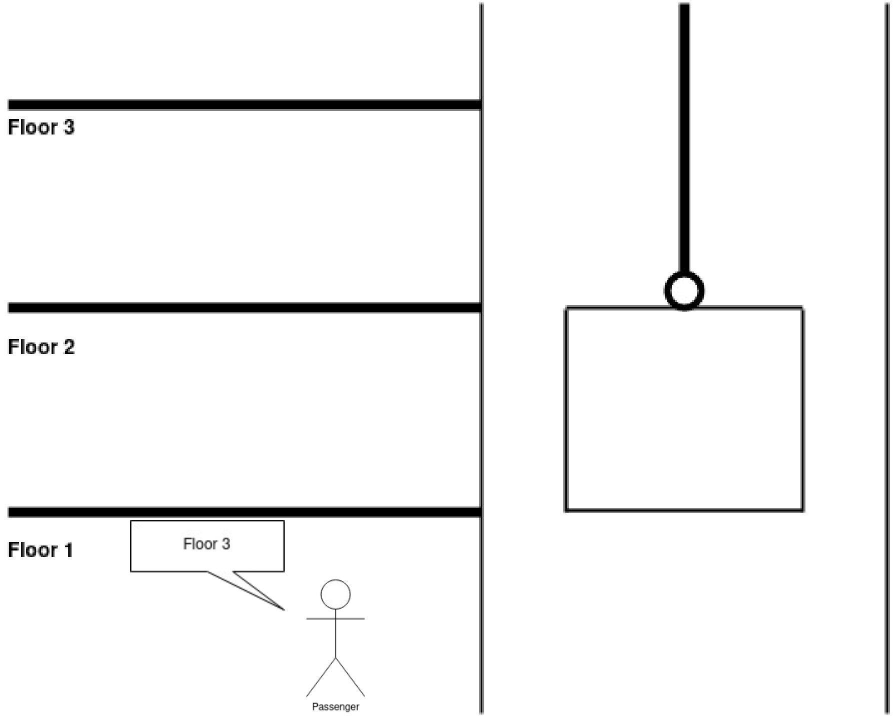
Elevators Domain



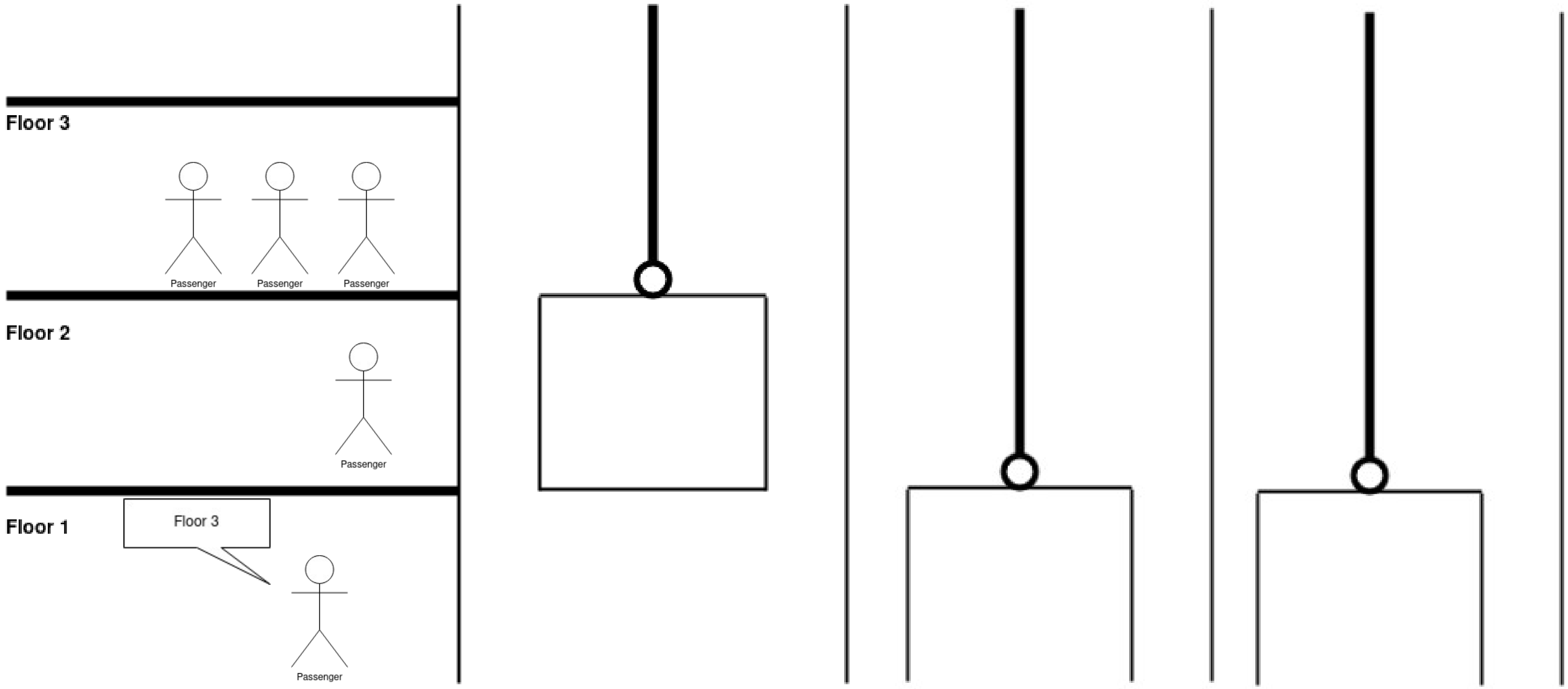
Elevators Domain



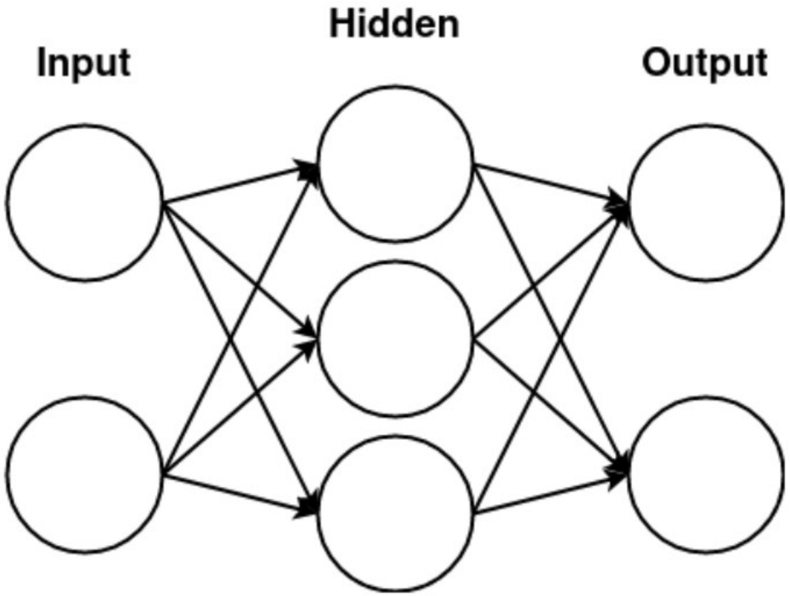
Elevators Domain



Elevators Domain



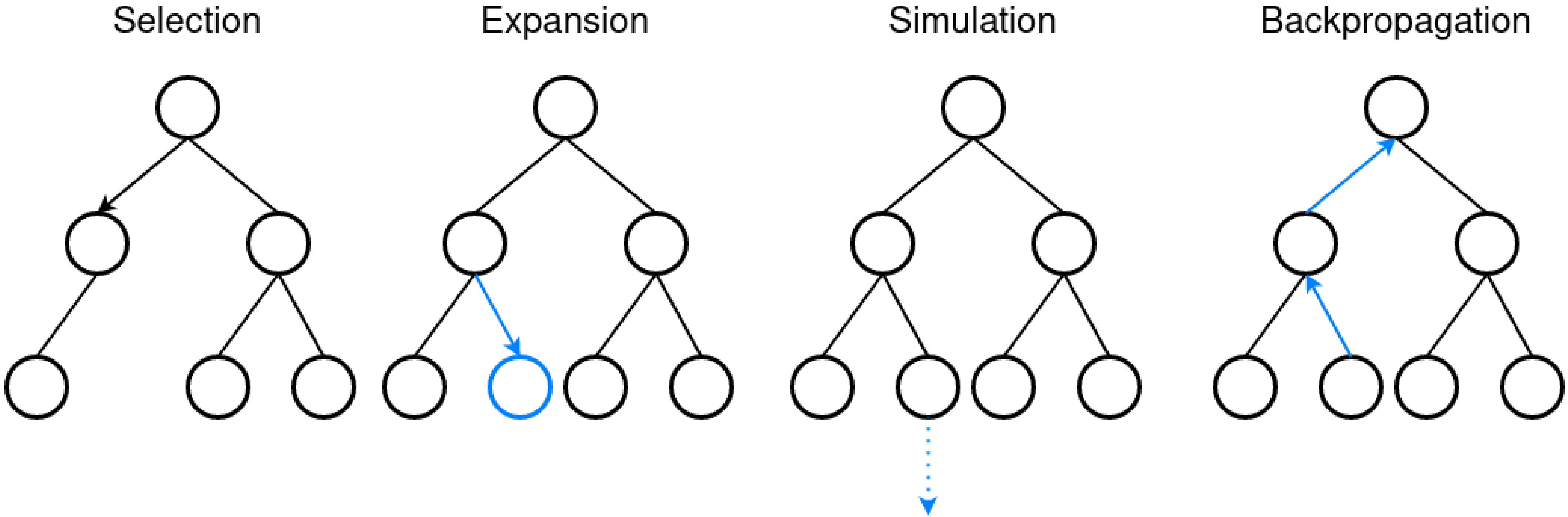
Neural Networks (NN)



Part 2 – Intention & General Approach

Intention

- Produce heuristic for MCTS based algorithm

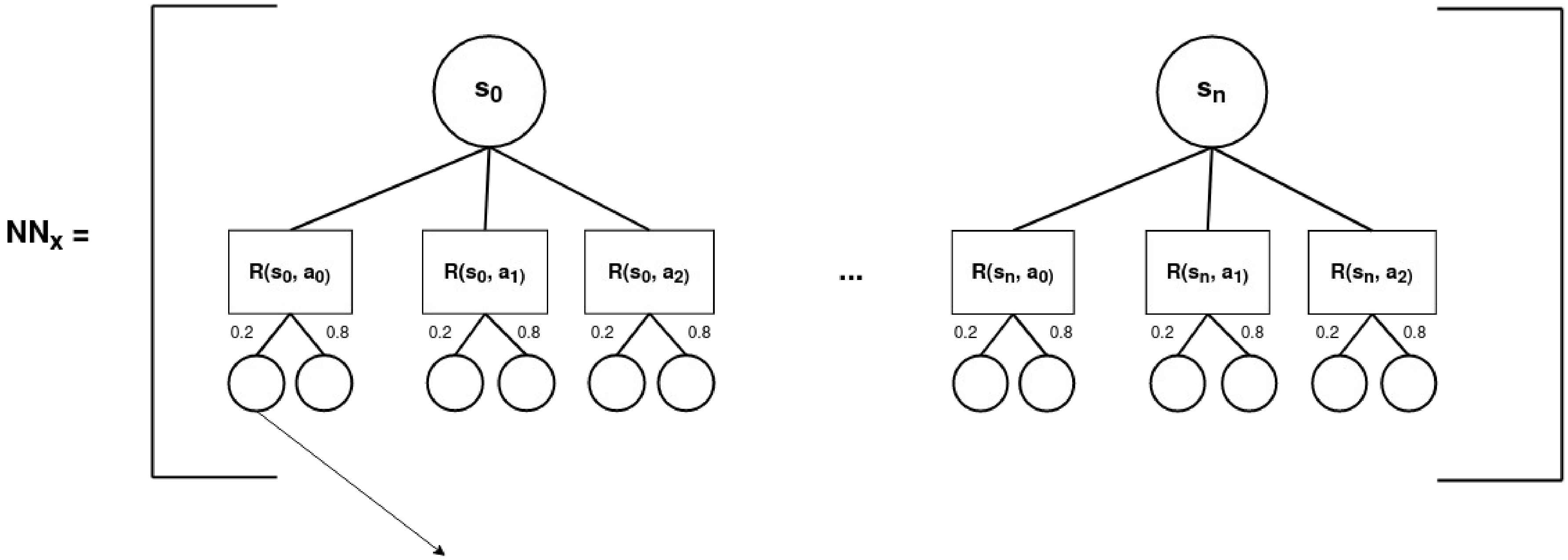


Increasing Horizon NN

- Provide Q-value NN for each depth
- Depth of 1 is equal to reward function
- Depth of 2:

$$Q_*(s, d, a) = R(s, a) + \sum_{s' \in S} P(s'|s, a) \cdot \max_{a \in A}(R(s', a))$$

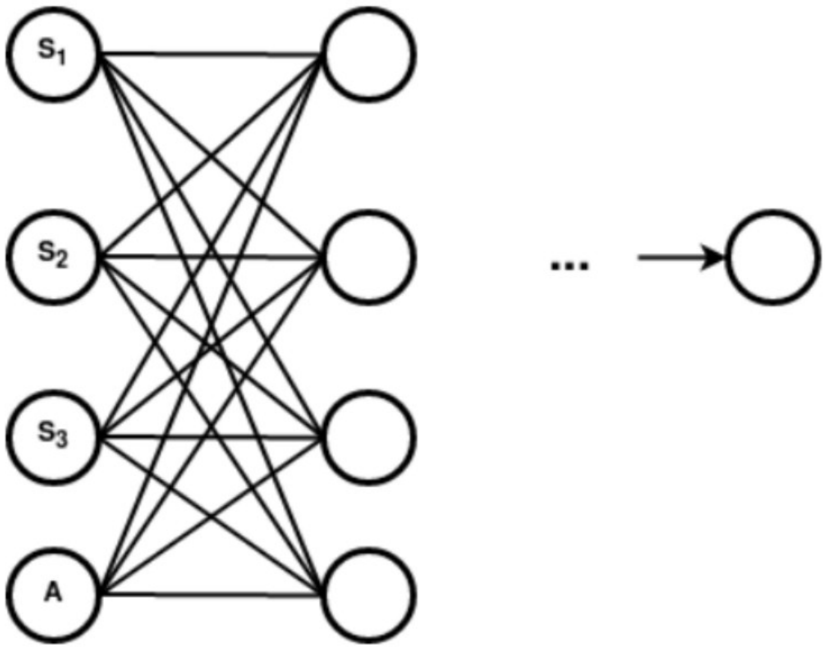
Increasing Horizon NN



$$\max_{a \in A}(NN_{x-1}(s_0', a))$$

Q-value NN

- Input state variables and applied action
- Output Q-value

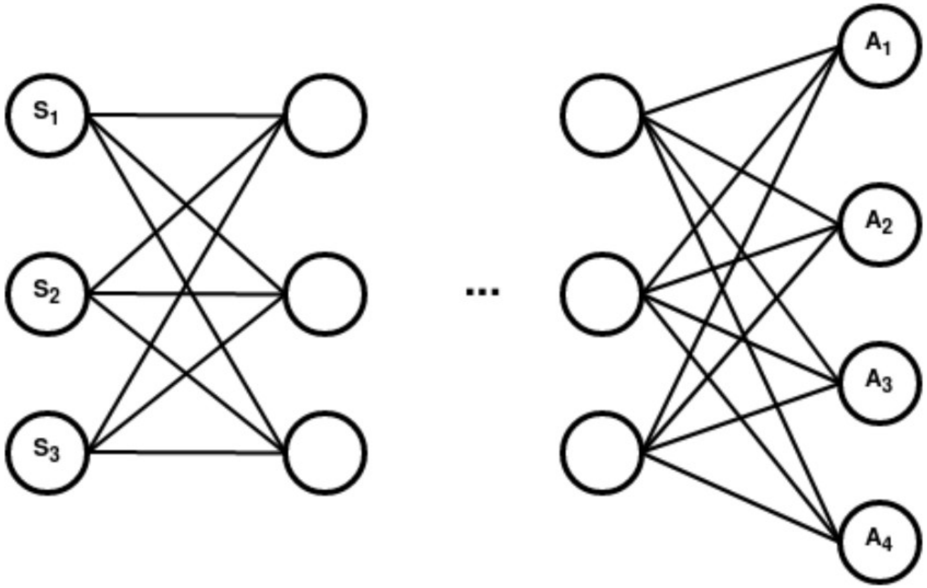


Shortcomings

- Training requires much time
- Split NN generation and search process
- Use last NN if not all NN’s prepared
- Have to use Q-value NN for each applicable action to get respective Q-value

Policy NN

- Provides Distribution over Q-values
- Less NN computations than Q-value



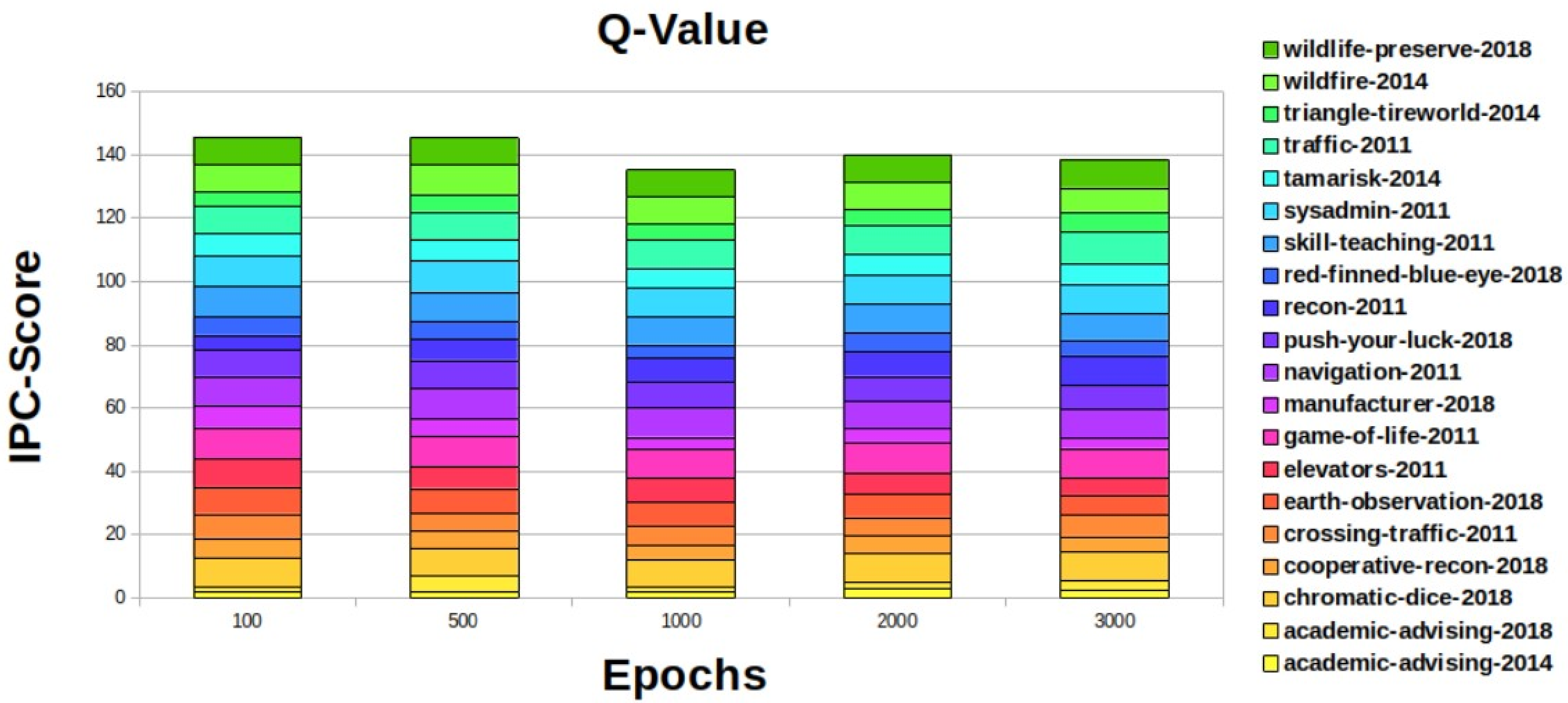
Part 3 Evaluation

Improving Parameters

- Need to find suitable hyper parameters:

  batch size; number of hidden layers; learning rate; epochs
- Hyper parameters are interdependent
- Perform local parameter search

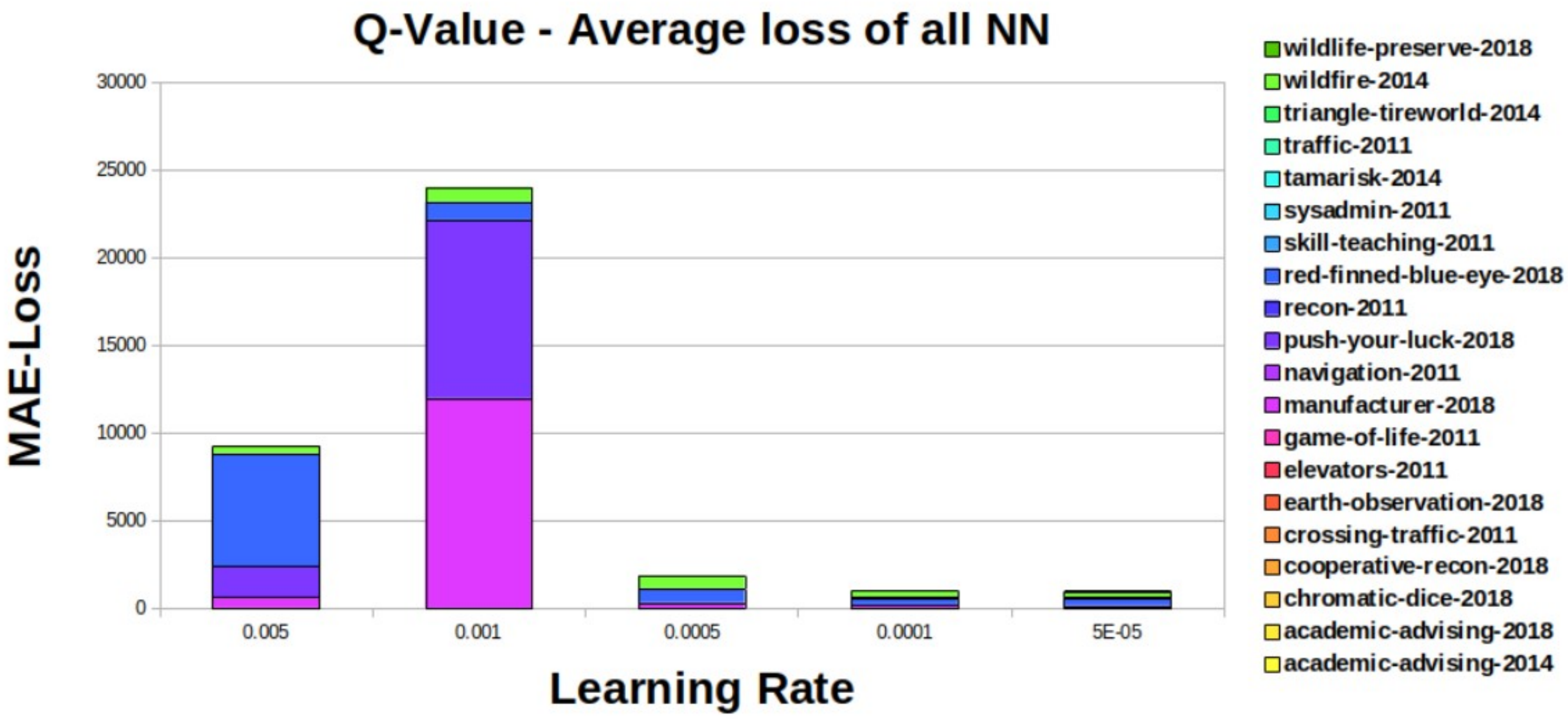
Epochs



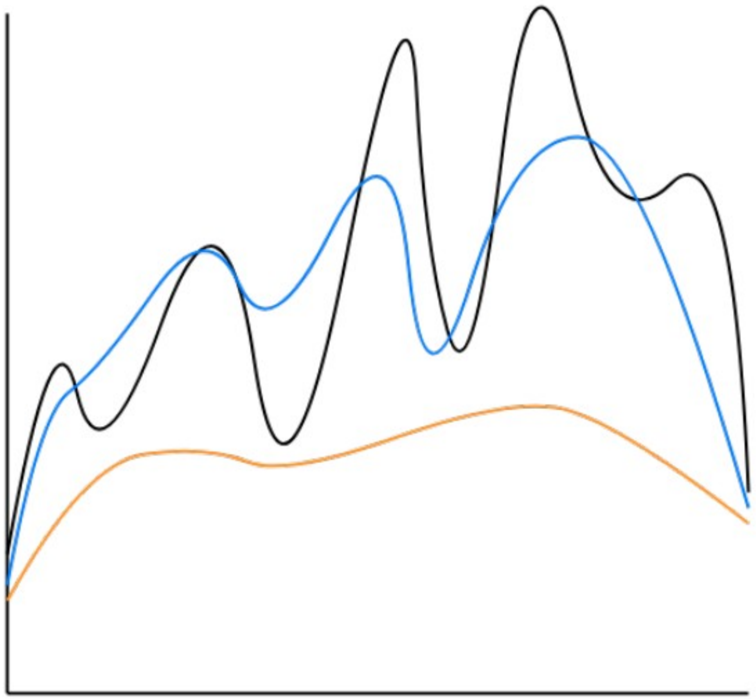
Q-Value
IPC-Score
160
140
120
100
80
60
40
20
0
100
500
1000
2000
3000
Epochs
wildlife-preserve-2018
wildfire-2014
triangle-tireworld-2014
traffic-2011
tamarisk-2014
sysadmin-2011
skill-teaching-2011
red-finned-blue-eye-2018
recon-2011
push-your-luck-2018
navigation-2011
manufacturer-2018
game-of-life-2011
elevators-2011
earth-observation-2018
crossing-traffic-2011
cooperative-recon-2018
chromatic-dice-2018
academic-advising-2018
academic-advising-2014

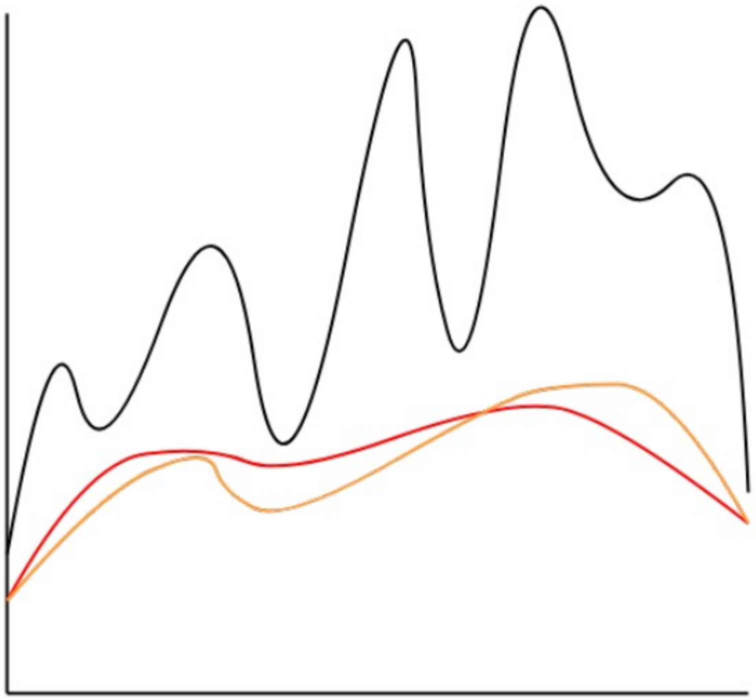
Learning Rate

Learning Rate

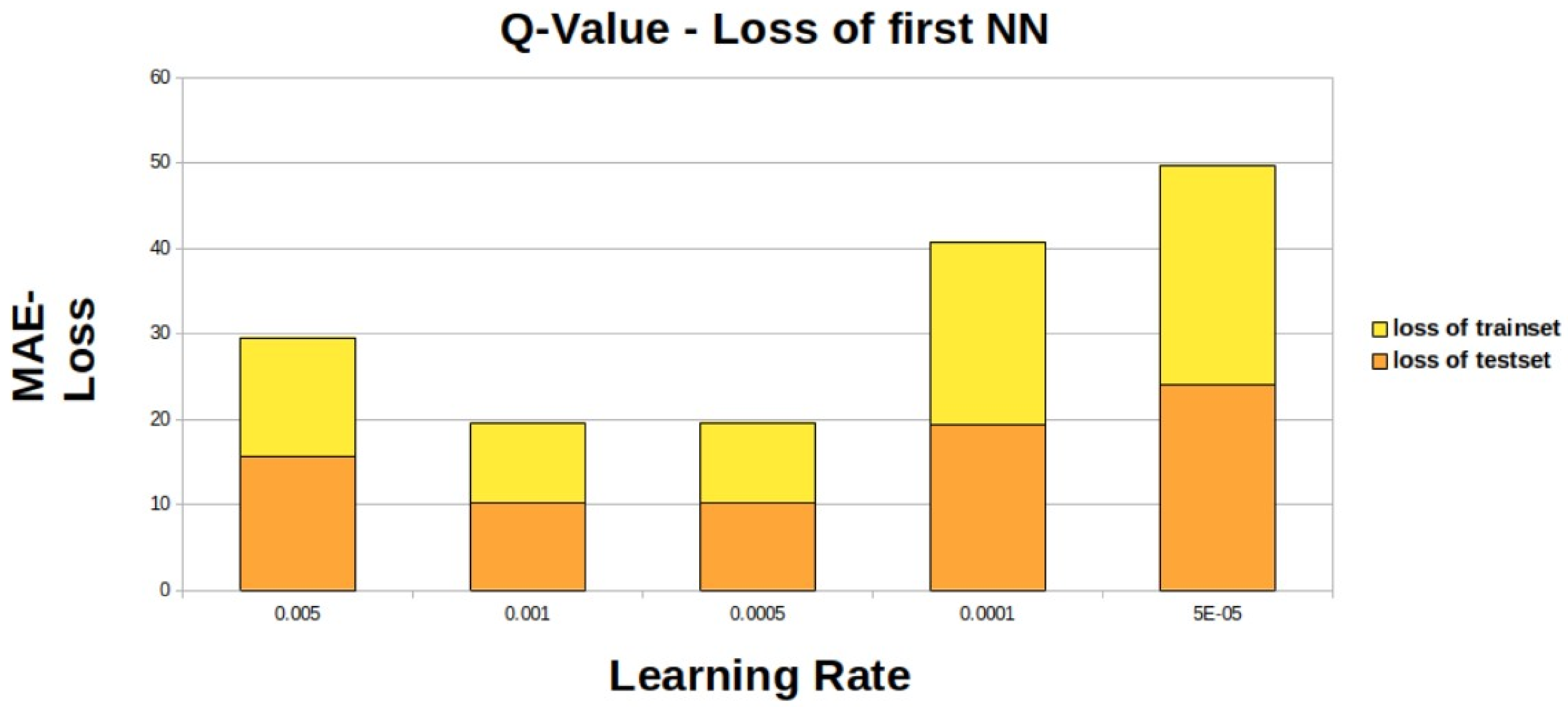


Learning Rate

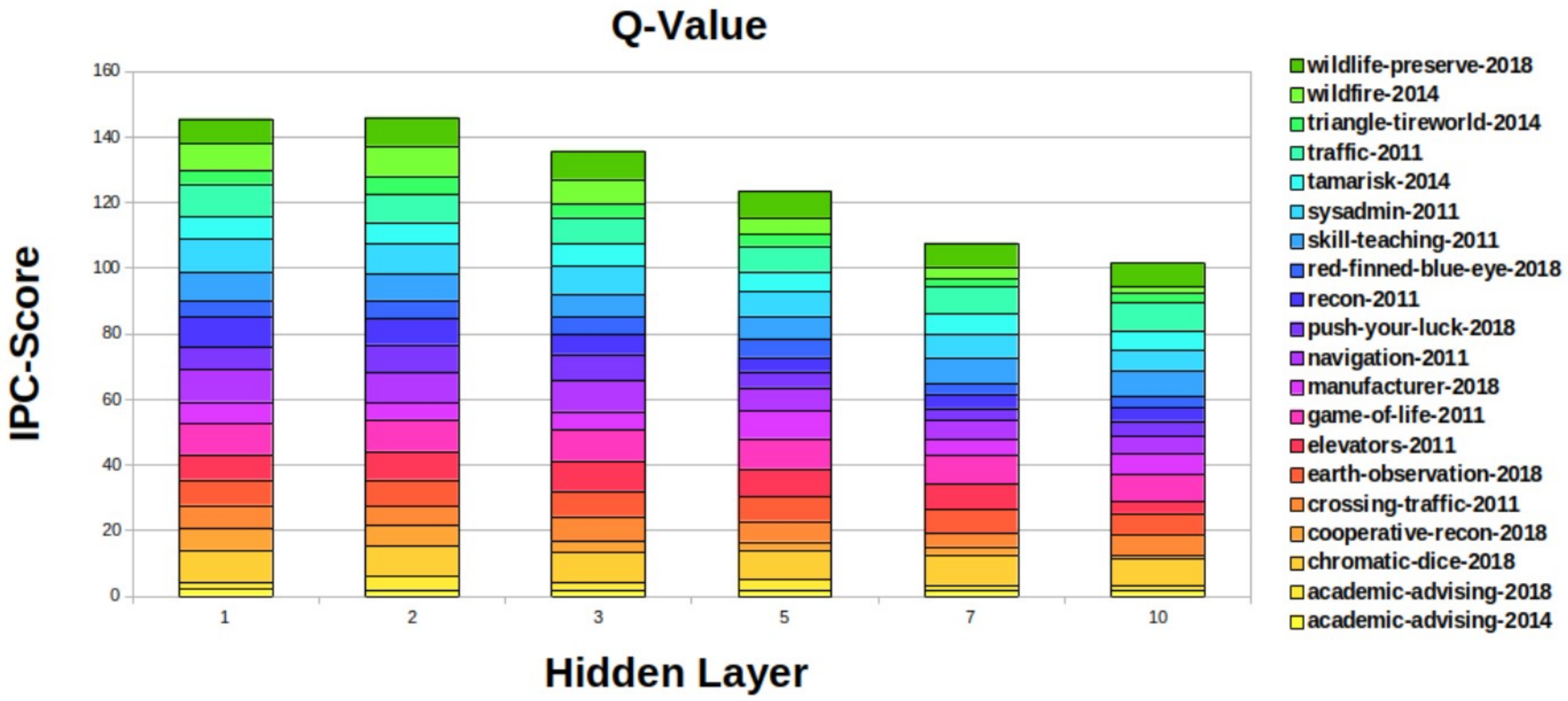




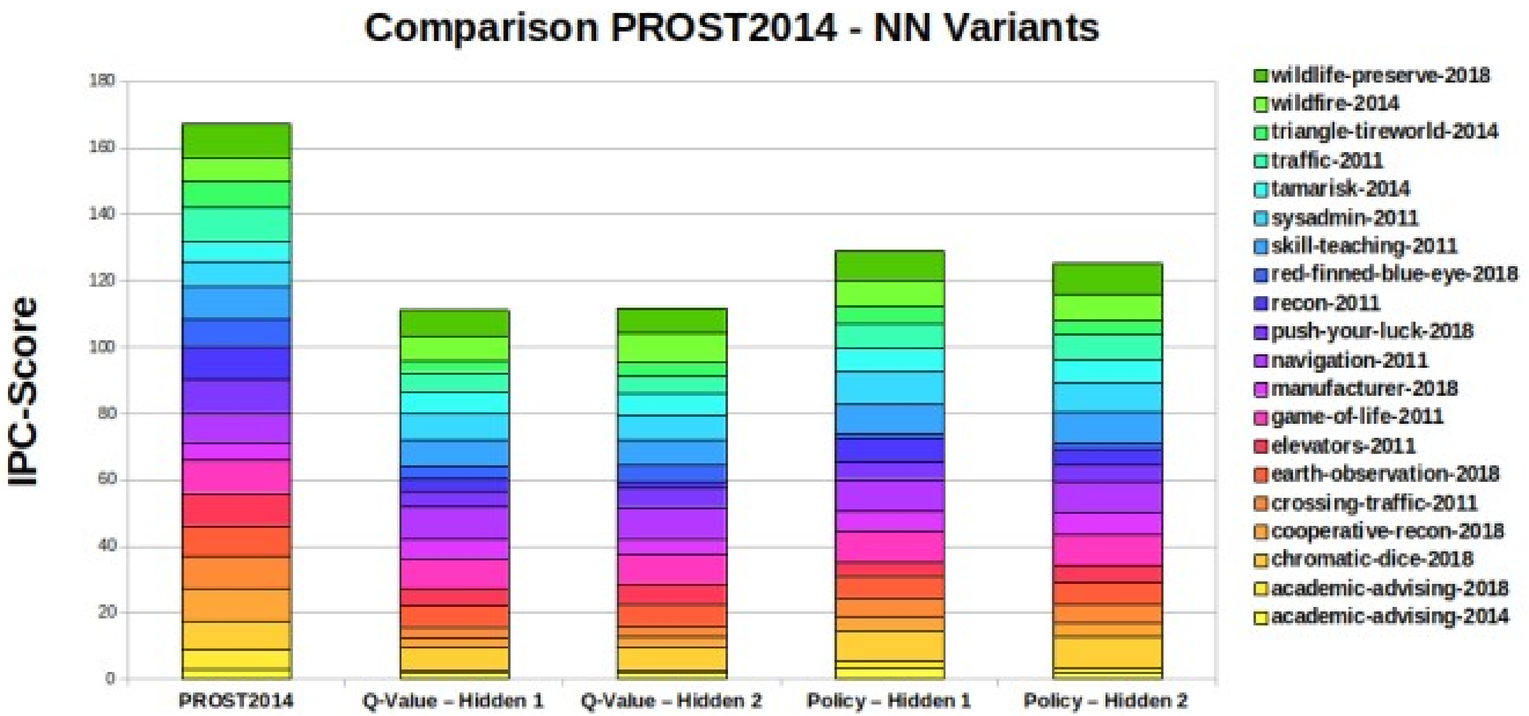
Learning Rate



Hiddenlayers and Batchsize



Comparison


Comparison PROST2014 - NN Variants
IPC-Score
180
160
140
120
100
80
60
40
20
0
PROST2014
Q-Value – Hidden 1
Q-Value – Hidden 2
Policy – Hidden 1
Policy – Hidden 2
wildlife-preserve-2018
wildfire-2014
triangle-tireworld-2014
traffic-2011
tamarisk-2014
sysadmin-2011
skill-teaching-2011
red-finned-blue-eye-2018
recon-2011
push-your-luck-2018
navigation-2011
manufacturer-2018
game-of-life-2011
elevators-2011
earth-observation-2018
crossing-traffic-2011
cooperative-recon-2018
chromatic-dice-2018
academic-advising-2018
academic-advising-2014

Adjustments

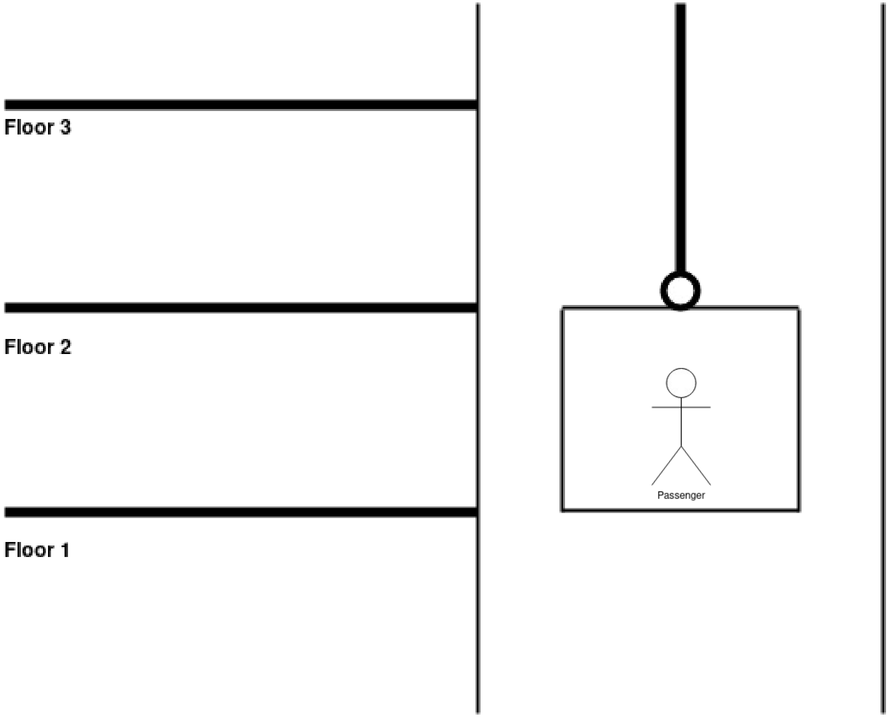
- Bounded NN
- Using the entire state-space as trainingset

Bounded Neural Network

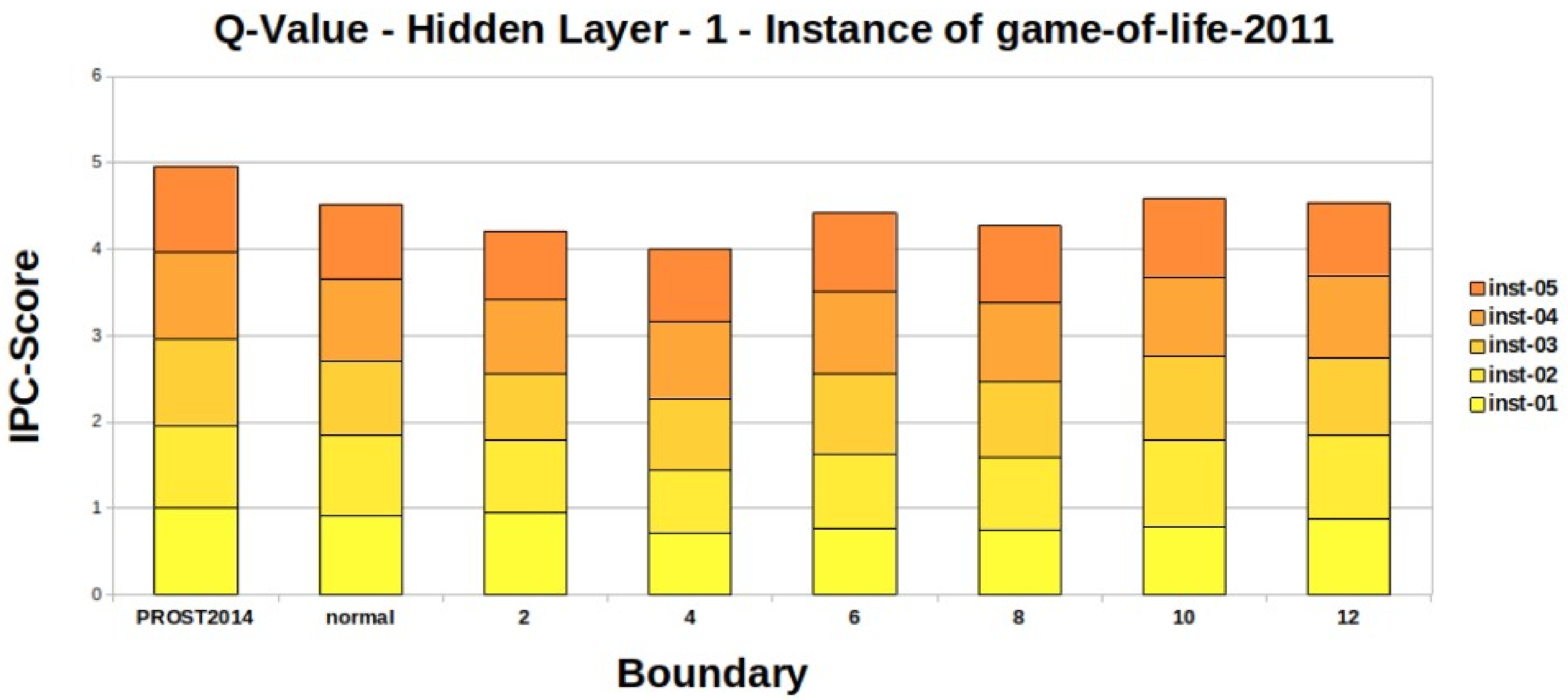
- Limit number of used NN depth d
- Last NN is used for all depths > d
- Assume optimum depth

Markov Decision Process (MDP)

elevators domain



Bounded Neural Network



Entire State-Space

- Use entire state-space
- Only feasible for smaller state-space instances
- Small state-space is already well explored by MCTS
- Contradiction to planning intention

Future work

Training Data

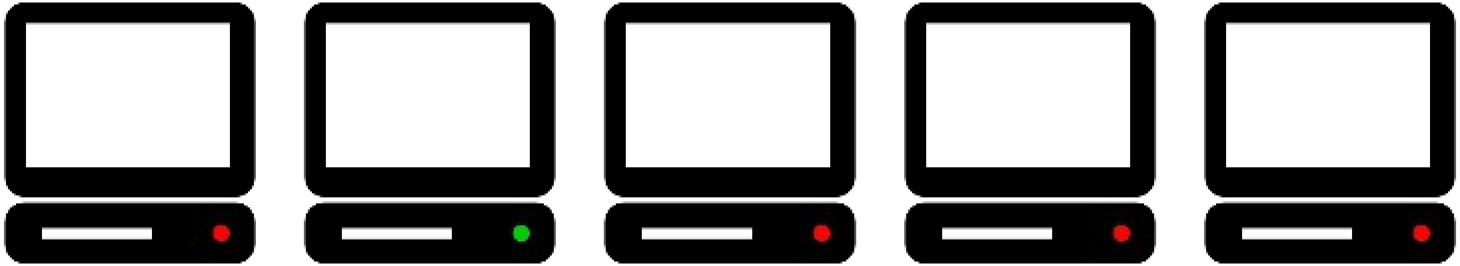
- Bad quality of data
- Randomly generated

Training Data

- Sysadmin domain

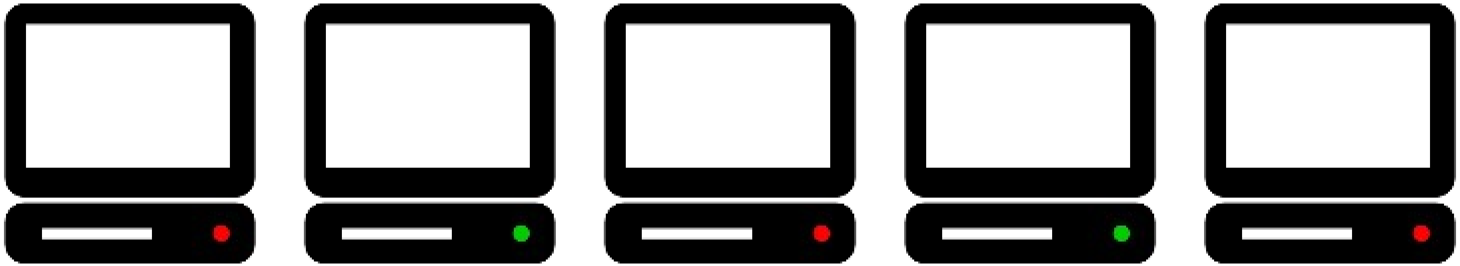
Training Data

- Sysadmin domain



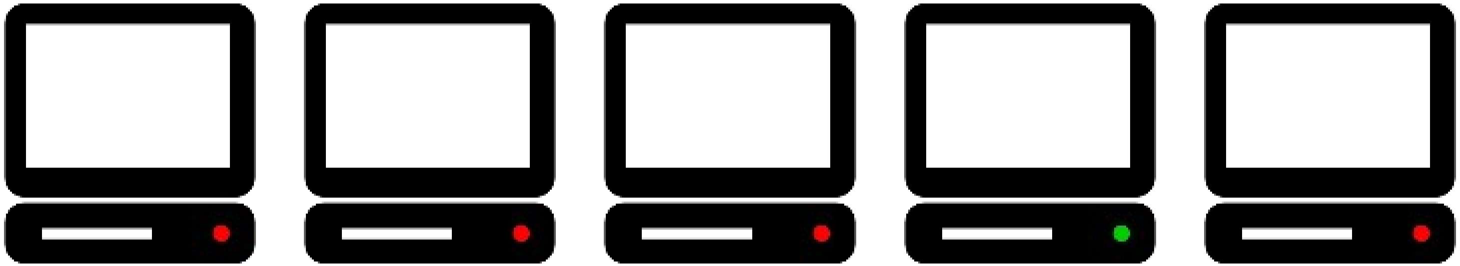
Training Data

- Sysadmin domain



Training Data

- Sysadmin domain

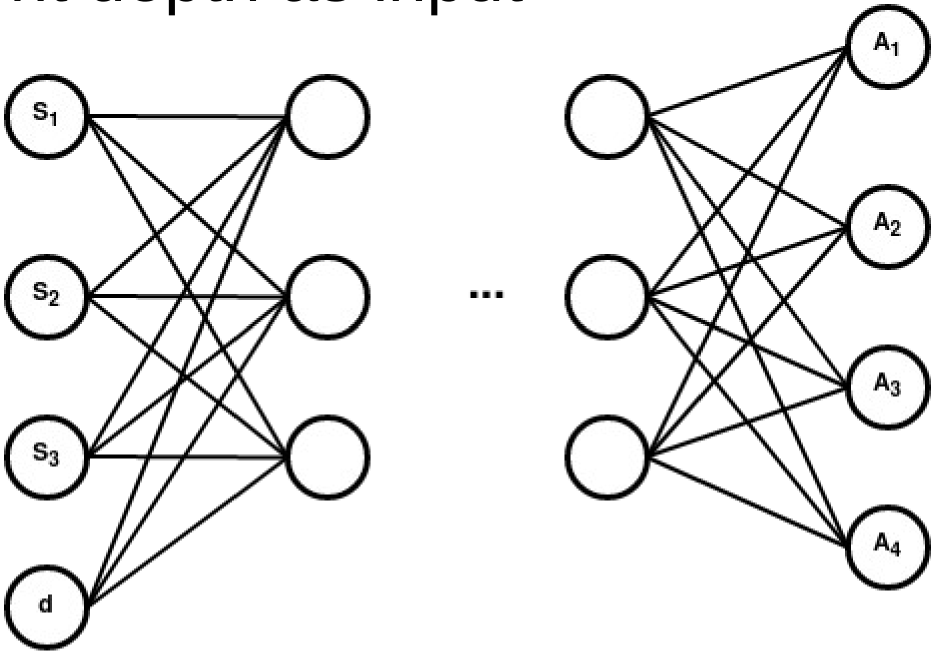


Extended hyper parameter search

- Only local parameter setting
- Extend search for approximating global optimum
- Investigate domain specific setting

Horizon Neural Network

- Use current depth as input



Mitigating accumulating errors

- Use MCTS for each sample point
- Adjust NN results
- Time intense but promising to decrease inaccuracy

Conclusion

- Results nevertheless satisfactory
- Improve data quality
- Extend hyper parameter search