Large feature dot products arise when out-of-sample actions are used in TD-learning compared to SARSA, despite similar Q-values

Large feature dot products *eventually* correlate with divergent Q-values

Out-of-sample (TD-learning)   In-sample (SARSA)

High dot products

Dot-product similarity

Breakout

Seaquest

Breakout

Gradient Updates (x 62.5k)

Similar Q-values

Average Q-value

Breakout

Seaquest

Breakout

Incorrect Q-values found eventually; dot-products increase throughout training