Ex. 1.1: Tic-Tac-Toe's shortage of available moves and the existence of a *best policy* would likely lead to policy convergence.

Ex. 1.2: May incorporate less *states* (as symmetrical states are equivalent). This is likely to speed up policy convergence. The opponent may have inherent position biases if they do not take advantage of symmetries. The learning algorithm may therefore take advantage of these biases by treating these *states* as separate (i.e. have different values).

Ex. 1.3: Worse, after a certain period of time. A greedy player may not sufficiently explore (or learn) high-value (initially) non-greedy state-action pairs, and thus may not be aware of more rewarding opportunities.

Ex. 1.4:

- Set 1 (Learn from exploratory moves): Probabilities that are conflated/impaired by inferior action selection (i.e. probability distribution of highest value and exploratory action selection).

- Set 2 (Not learn from exploratory moves): Probabilities that reflect/represent a purely greedy policy (i.e. probability distribution of highest value).

Set 2 is better and would result in more wins.

Ex. 1.5:

1. More informed Q initialisation.
2. Play different sub-games (and amalgamate updates) that emanate from a particular state.
3. Tic-Tac-Toe Problem: Deterministic algorithm based on player moves.