1. Learning by Example

1.1. $V^{\pi}(s) = 0$ for any state

Episode 1: $V^{\pi}(A) = (1-0.25) * V^{\pi}(A) + 0.25[R(A, right, B) + V^{\pi}(B)]$
$V^{\pi}(A) = (0.75) * 0 + 0.25[16 + 0] = 4$

Episode 2: $V^{\pi}(B) = (1-0.25) * V^{\pi}(B) + 0.25[R(B, right, C) + V^{\pi}(C)]$
$V^{\pi}(B) = (0.75) * 0 + 0.25[4 + 0] = 1$

Episode 3: $V^{\pi}(B) = (1-0.25) * V^{\pi}(B) + 0.25[R(B, down, E) + V^{\pi}(E)]$
$V^{\pi}(B) = (0.75) * 1 + 0.25[-12 + 0] = -2.25$

Episode 4: $V^{\pi}(C) = (1-0.25) * V^{\pi}(C) + 0.25[R(C, down, F) + V^{\pi}(F)]$
$V^{\pi}(C) = (0.75) * 0 + 0.25[-16 + 0] = -4$

Episode 5: $V^{\pi}(F) = (1-0.25) * V^{\pi}(F) + 0.25[R(F, stay, F) + V^{\pi}(F)]$
$V^{\pi}(F) = (0.75) * 0 + 0.25[8 + 0] = 2$

Episode 6: $V^{\pi}(C) = (1-0.25) * V^{\pi}(C) + 0.25[R(C, down, F) + V^{\pi}(F)]$
$V^{\pi}(C) = (0.75) * -4 + 0.25[-9 + 2] = -4.75$

State	$U^{\pi}(state)$
A	4
B	-2.25
C	-4.75
D	0
E	0
F	2

2. Reinforcements

2.1. $Q(A, right) = (1-0.5) * Q(A, right) + 0.5(r_1 + maxQ(s_B, a'))$
$= 0.5 * 0 + 0.5(4 + 0) = 2$

$Q(C, left) = (1-0.5) * Q(C, left) + 0.5(r_1 + maxQ(s_B, a'))$
$= 0.5 * 0 + 0.5(4 + 0) = 2$

$Q(B, right) = (1-0.5) * Q(B, right) + 0.5(r_1 + maxQ(s_C, a'))$
$= 0.5 * 0 + 0.5(-4 + 2) = -1$

$Q(A, right) = (1-0.5) * Q(A, right) + 0.5(r_1 + maxQ(s_B, a'))$
$= 0.5 * 2 + 0.5(8 + 0) = 5$

2.1.1. Q(A right) = 5
Q(B, right) = -1

2.1.2. π_Q(A) = right
π_Q(B) = left

2.2. T(A, right, B) = 1.0 (was at state A two times and went right to state B twice)
R(A, right, B) = (4 + 8) / 2 = 6

T(B, right, A) = 0 (was at state B once and went right and the next state was state C, not state A)
R(B, right, A) = N/A (since there was no observed transition from B, going right, to A, there was no reward)

T(B, left, A) = 0 (was at state B once, but went right to state C)
R(B, left, A) = N/A (since there was no observed transition from B, going left, to A, there was no reward)

3. Reinforcement Learning Background

3.1. True or false

i) False, Temporal difference learning updates it's value estimates incrementally, made from interactions with the environment

ii) True, while optimal, exploration involves uncertainty which, in some cases, can lead to regret due to suboptimal decisions (from exploring new actions).

iii) True, since there is no randomness in a deterministic MDP, with a learning rate of $\alpha = 1$, it will eventually converge to the optimal policy

iv) False, having a large discount (close to 1) means that the agent values the future rewards almost as much as the immediate rewards. Greedy behavior, on the other hand, places an emphasis on immediate rewards and not the future rewards.

v) True, a large, negative living reward incentivizes the agent to avoid prolonged interaction with the environment because it keeps losing rewards over time. Thus, the agent is encouraged prioritize immediate rewards, which is a form of greedy behavior

vi) False, a negative living reward isn't always expressed by using a discount < 1. The discount value impacts how future rewards are valued. It doesn't penalize the agent for staying in the environment, it makes the agent less motivated to

consider future rewards. A negative living reward is conceptually different from the discount value.

vii) True, similar to above, a negative living reward is different from a discount value < 1, thus it cannot always be expressed as the same.

3.2. a)
Direct Evaluation to estimate U(s)
Temporal Difference learning to estimate U(s)

b)
A fixed policy taking actions uniformly at random
An ε-greedy policy