1. (4.3) vπ(s) = 𝔼π [Rt+1 + γvπ(St+1) | St = s]

qπ(s, a) = 𝔼π [Rt+1 + γGt+1 | St = s, At = a]

(4.4) vπ(s) = ∑a π(a|s) ∑s’, r p(s’, r|s,a) [r + γvπ(s’)]

qπ(s, a) = ∑s’, r p(s’, r|s, a) ∑a’ π(a’|s’) [r + γqπ(s’, a’)]

(4.5) vk+1(s) = ∑a π(a|s) ∑s’, r p(s’, r|s,a) [r + γvk(s’)]

qk+1(s, a) = ∑s’, r p(s’, r|s, a) ∑a’ π(a’|s’) [r + γqk(s’, a’)]

**Initialization:**

Q(s, a)∈ℝ and π(s)∈A(s) arbitrarily for all s∈S

**Policy Evaluation:**

Repeat

### [Δ](https://en.wiktionary.org/wiki/%CE%94) ← 0

For each s∈S:

q ← Q(s, π(s))

Q(s, π(s)) ← ∑s’, r p(s’, r|s, π(s)) ∑π(s)’ π(π(s)’|s’) [r + γqπ(s’, π(s)’)]

[Δ](https://en.wiktionary.org/wiki/%CE%94) ← max([Δ](https://en.wiktionary.org/wiki/%CE%94),|q – Q(s, π(s)))

### until [Δ](https://en.wiktionary.org/wiki/%CE%94) < [θ](http://wordinfo.info/unit/2138/s:alphabet) (a small positive number)

**Policy Improvement:**

policy-stable ← true

For each s∈S:

old-action ← π(s)

π(s) ← argmaxa’ ∑s’, r p(s’, r|s, a) ∑a’ π(a’|s’) [r + γqπ(s’, a’)]

If old-action ≠ π(s), then policy-stable ← false

If policy-stable, then stop and return Q ≈ q\* and π ≈ π\*; else go to **Policy Evaluation**