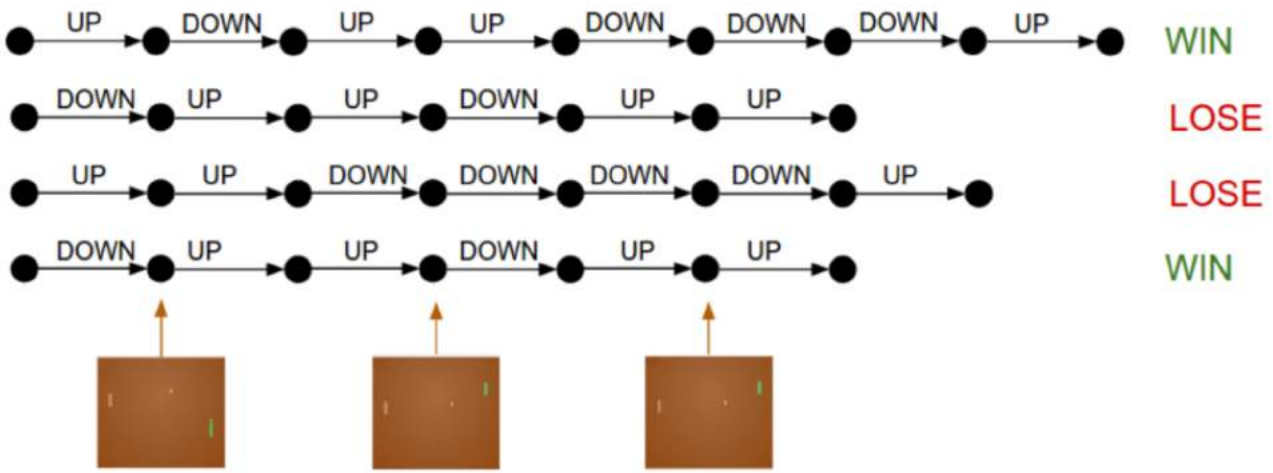
Policy Gradient (PG): Training



1. Run a policy for a while
2. Increase probability of actions that lead to high rewards
3. Decrease probability of actions that lead to low/no rewards

```
function REINFORCE
 Initialize θ
 for episode ~ π_θ
  {s_i, a_i, r_i}_{i=1}^{T-1} ← episode
  for t = 1 to T-1
     ∇ ← ∇_θ log π_θ(a_t|s_t) R_t
     θ ← θ + α∇
 return θ
```

log-likelihood of action

$$\nabla_\theta \log \pi_\theta(a_t|s_t)\, R_t$$

reward