* Optimality Criteria

- 설명
  - policy를 결정하게 되면, 이에 해당되는 reward sequence를 random으로 반환받음.
  - 이 reward sequence가 최대한 커지도록 policy를 선택하는 문제.
  - random 시퀀스들 가운데 최대 reward 선택
- Notations
  - $h_{t+1} = (h_t, a_t, s_{t+1})$ 이며 $h_t = (s_1, a_1, s_2, \dots, a_{t-1}, s_t)$
  - X_t는 t 시점에서 state에 대한 확률변수
  - Y_t는 t 시점에서 action에 대한 확률변수
  - 보상에 대한 확률 변수 $R = \begin{cases} R_t = g_t(X_t, Y_t) & (t<N) \\ R_N = g_N(X_N) & (t=N) \end{cases}$

* π^k에 따라 달라지는 optimal Criteria

- $\pi \in d_t^{HR} \sim R =$
- $\pi \in d_t^{HD} \sim R = (g_1(X_1, d_1(h_1)), \dots, g_{N-1}(X_{N-1}, d_{N-1}(h_{N-1})), g_N(X_N))$
- $\pi \in d_t^{MD} \sim R = (g_1(X_1, d_1(x_1)), \dots, g_{N-1}(X_{N-1}, d_{N-1}(X_{N-1})), g_N(X_N))$

* Total Reward Criterion의 기댓값.

- Optimality Criteria는 모든 v에 대해서 $\varphi(u)$고 $\varphi(v)$한 reward 시퀀스 $u = (u_1, \dots, u_N)$을 찾을 일 → 이는 사람마다 다를 수 있음.
- $E^{\pi}[\varphi(R)] \geq E^{\pi'}[\varphi(R)]$을 만족하면 π'보다 π 더 선호
- $E^{\pi}[\varphi(R)] = \sum_r \varphi(r) \cdot p^{\pi}(R=r)$
- policy의 기대효용을 계산이 어려우므로 φ 함수는 선형이라고 가정.
  - $\varphi(g_1, g_2, \dots, g_N) = g_1 + g_2 + \cdots + g_N$

* π^k에 따라 달라지는 기대효용값

- $\pi \in dt^{HR} \sim V_N^{\pi}(s) = E_s^{\pi}\left[\sum_{t=1}^{N-1} g_t(X_t, Y_t) + g_N(X_N)\right]$
- $\pi \in dt^{HD} \sim V_N^{\pi}(s) = E_s^{\pi}\left[\sum_{t=1}^{N-1} g_t(X_t, d_t(h_t)) + g_N(X_N)\right]$

* Discount factor

- 시간적 요소를 고려하기 위해 discount factor를 도입.
- t+1 시기에 받은 보상를 t시기의 개념으로 측정.
- $\pi \in dt^{HR} \sim V_{N,\lambda}^{\pi}(s) = E_s^{\pi}\left[\sum_{t=1}^{N-1} \lambda^{t-1} g_t(X_t, Y_t) + \lambda^{N-1} g_N(X_N)\right]$

* Tolerance

- $s \in S$, $\pi \in \Pi^k$ 일 때 $V_N^{\pi^*}(s) \geq V_N^{\pi}(s)$, $k = \{MD, MR, HD, HR\}$ 하면 π^*를 optimal policy 라고 함.
- $s \in S$, $\pi \in \Pi^k$ 이고 tolerance $\varepsilon > 0$ 에 대해 $V_N^{\pi^*}(s) \geq V_N^{\pi}(s) - \varepsilon$ 하면 π_ε^*를 ε-optimal policy 라고 함.

* Value of a MDP under the expected total reward criterion.

- $V_N^*(s) = \sup_{\pi \in \Pi^k} V_N^{\pi}(s)$ 이며, $V_\pi^* : S \to (-\infty, \infty)$ 임.
- 만약 $V_N^*(s) = V_N^{\pi^*}(s)$ 이면 π^*는 optimal policy임.
- $V_N^{\pi_\varepsilon^*}(s) + \varepsilon \geq V_N^*(s)$ 일 때 π_ε^*는 ε-optimal policy 임.