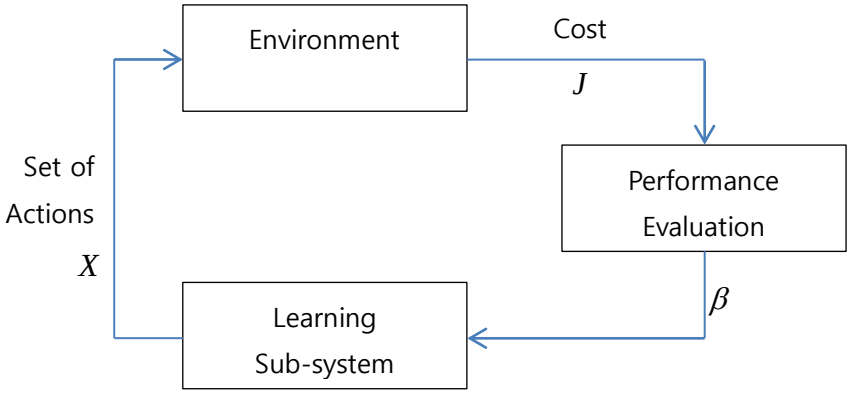
CARLA(continuous action reinforcement learning automata)

강화학습으로 최적의 파라미터 탐색

참조: M. N. Howell의 "'Continuous Action Reinforcement Learning Applied to vehicle suspension Control"와 남동균의 "저가형 자이로 센서의 드리프트 감소 방안에 대한 고찰"

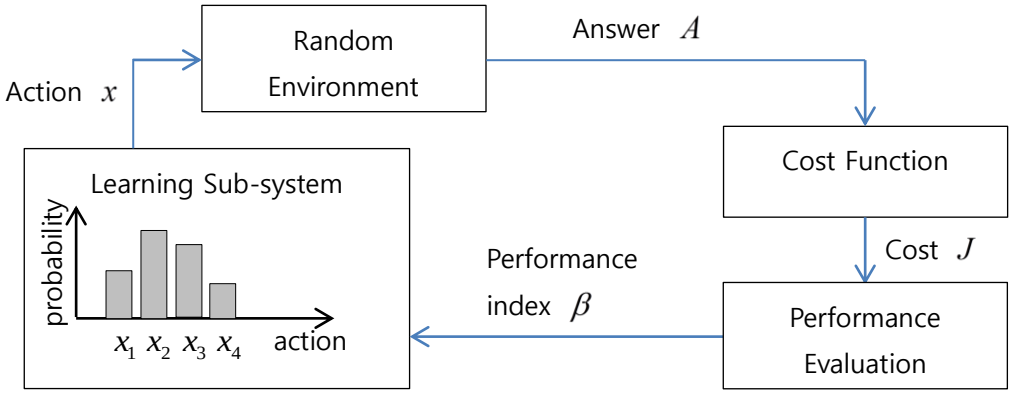
이전 문서에서 자이로 센서의 드리프트를 최소화 하기위한 HDR 알고리즘을 소개하였는데, 여기서는 자이로 센서의 최적의 파라미터를 찾는 방법에 대해 소개한다.

Typical reinforcement learning system:



Learning sub-system은 environment로 액션 X 를 보내고 비용 J 을 측정한다. 그리고 성능평가 함수로부터 성능지수 β 를 계산하여 learning sub-system을 학습한다.

Discrete Action Learning Automata



Stochastic learning automata는 유한한 크기의 이산 액션집합 $\{x_1, x_2, \cdots, x_n\} \in X$ 에서 임의로 선

택된 액션에 의해 동작하는 비연관 강화학습법(non-associative reinforcement learning) 형식이다. 각각의 액션 x_i는 선택될 확률 p_i를 가진다. 최초 학습시에는 모든 확률은 동일하다.

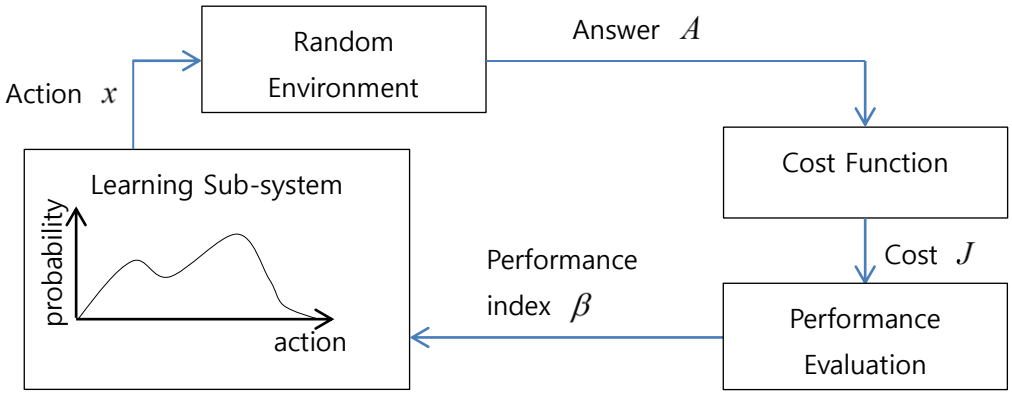
학습 과정에서 얻어지는 성능평가 신호 β에 대하여 Linear reward inaction 알고리즘은 다음과 같이 확률값들을 업데이트 한다.

$$p_i(n+1) = p_i(n) + \theta\beta(n)(1 - p_i(n))$$
$$p_j(n+1) = p_j(n) - \theta\beta(n)p_j(n) \quad \text{if } i \neq j$$

여기서 θ는 학습 비율 파라미터로 $0 < \theta < 1$이고, $\beta(n) \in [0,1]$는 보상값이다.

CARLA

CARLA는 이산형 액션 공간을 연속형으로 대체한 것이다.



적절한 액션은 Gaussian distribution function을 통해 선택 확률을 높여간다.

액션 변수 x는 구간 $X = [x_{min}, x_{max}] \subset \mathbb{R}$에서 정의되고, 확률밀도함수(*pdf*)는 다음 조건을 만족한다.

$$\int_X f(x,n)dx = 1,$$

$$f(x,n) \geq 0 \ \forall n, \forall x$$

Initial distribution

초기 확률밀도함수는 전 구간에 동일하게 설정한다.

$$f(x,1) = \begin{cases} \dfrac{1}{x_{min} - x_{max}} & \text{for } x \in X \\ 0 & \text{otherwise} \end{cases}$$

Action selection

$x(n)$ 은 non-uniform distribution function $f(x,n)$ 에 기반하여 $U[x_{min}, x_{max}]$ 범위 내의 랜덤 값으로 선택한다. 이 선택방법은 uniform variable $z(n)$ 을 통해 계산할 수 있다. 즉, n번째 반복횟수에서 균등 변수 $z(n) \sim U[0,1]$ 이 주어지면 다음 식을 만족하기 위한 액션 값 $x(n)$ 을 찾는다.

$$\int_{x_{min}}^{x(n)} f(x,n)dx = z(n)$$

Cost Function

선택된 액션이 environment에 적용된 후 사용자가 정의한 비용함수 $J(n)$ 에 의해 액션의 효과를 측정한다. 그리고 과거부터 현재까지의 비용은 집합 R 에 저장한다.

$$R = \{J_{min}, \cdots, J_{med}, \cdots, J_{max}\}$$

여기서 J_{med} 와 J_{min} 은 집합에 저장되어 있는 비용들의 중간값과 최소값이다. 특히, 집합 R 이 무한정 커지는 문제를 피하기 위해 계산에 사용되는 비용들은 최근의 m 개의 크기로 제한하여야 한다.

Performance index

상기 비용은 선택된 액션의 우수성을 판별하는 지표가 되며, 비용 집합과 현재 상태에서 계산된 비용으로부터 보상값(β)은 다음과 같이 계산된다.

$$\beta(n) = \max\left(0, \frac{J_{med} - J(n)}{J_{med} - J_{min}}\right)$$

Update probability density function

보상값이 결정되면 이를 기반으로 *pdf* 함수 $f(x,n)$을 다음과 같은 규칙으로 업데이트 한다.

$$f(x,n+1)=\alpha\left(f(x,n)+\beta(n)H(x,r)\right)$$

여기서 α는 정규화 조건을 만족시키는 파라미터이고, $H(x,r)$은 중심이 $r=x(n)$인 대칭형 가우시안 함수로 다음과 같이 표현할 수 있다.

$$H(x,r)=\lambda\exp\left(-\frac{(x-r)^2}{2\sigma^2}\right)$$

$$\int_{x_{min}}^{x_{max}} f(x,n+1)dx=1$$

λ와 σ는 다음과 같이 정의한다.

$$\lambda=\frac{g_h}{x_{max}-x_{min}}$$

$$\sigma=g_w(x_{max}-x_{min})$$

여기서 두 파라미터 g_h와 g_w는 학습속도와 분해능을 조절하는 값이다.