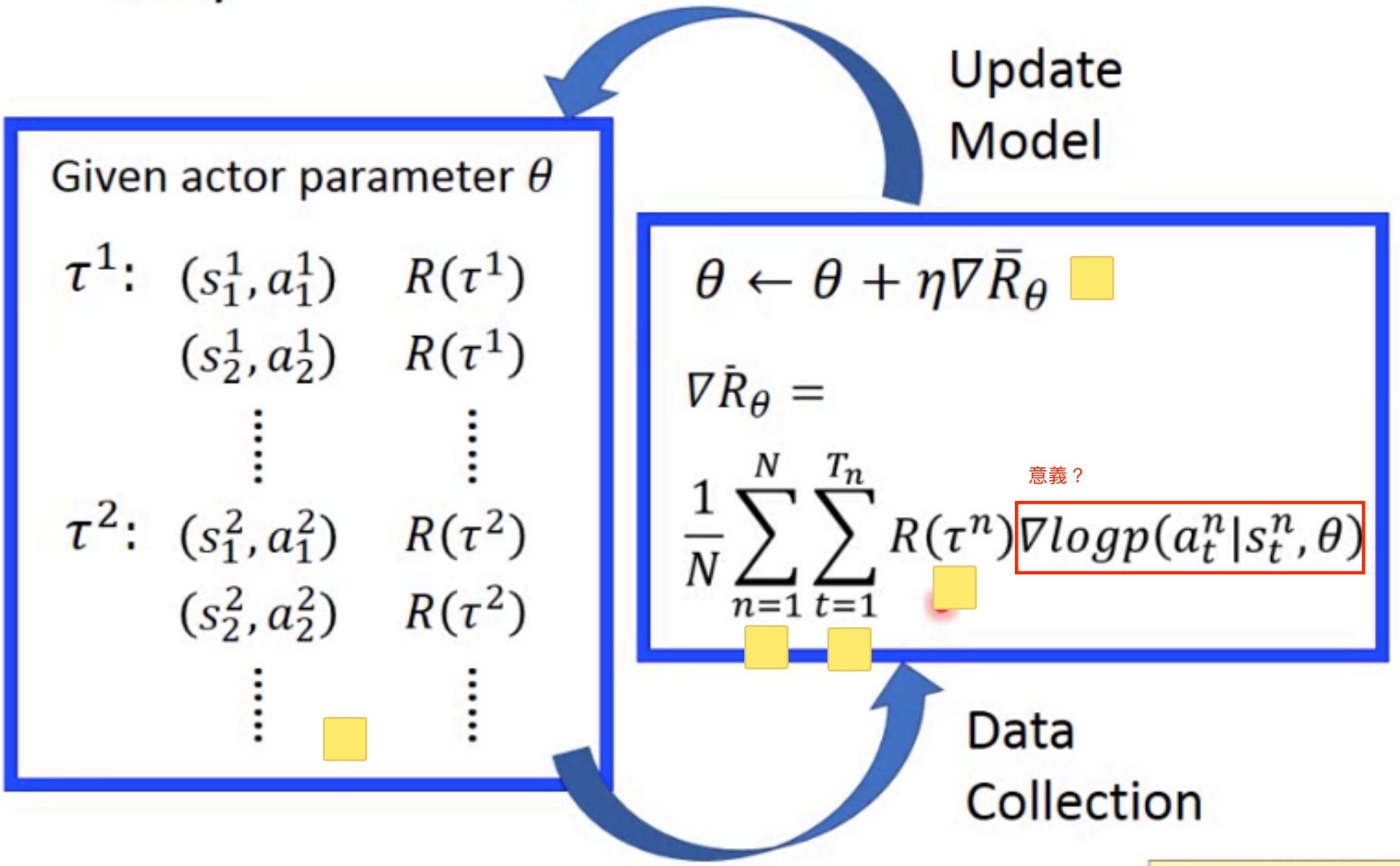
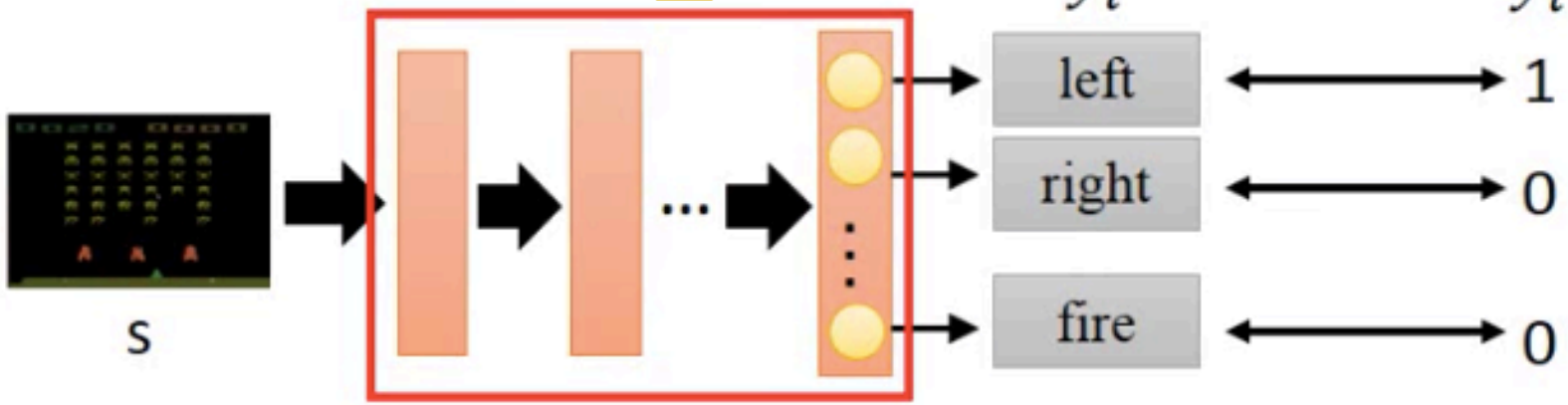
Policy Gradient



Policy Gradient

Considered as Classification Problem

Minimize: $-\sum_{i=1}^{3} \hat{y}_i log y_i$



Maximize: $logy_i = logP("left"|s)$

$$\theta \leftarrow \theta + \eta \nabla logP("left"|s)$$

Policy Gradient

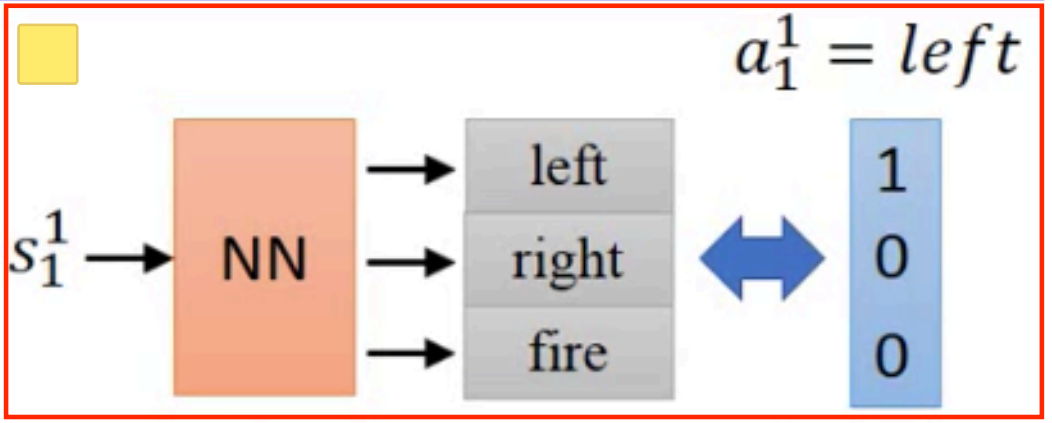
$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

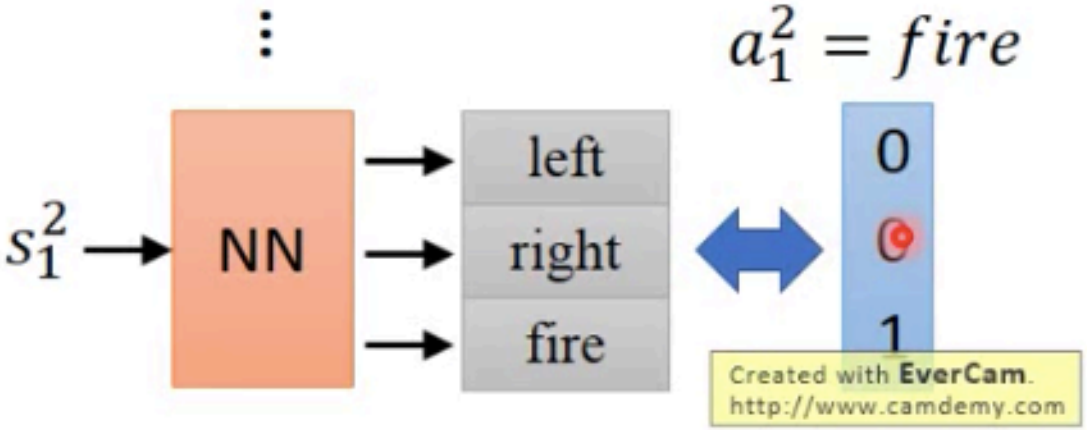
$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \blacksquare \nabla log p(a_t^n | s_t^n, \theta)$$

來自一個分類問題

Given actor parameter θ

$$\begin{array}{lll} \tau^1: & (s_1^1, a_1^1) & R(\tau^1) \\ & (s_2^1, a_2^1) & R(\tau^1) \\ & \vdots & \vdots \\ \tau^2: & (s_1^2, a_1^2) & R(\tau^2) \\ & (s_2^2, a_2^2) & R(\tau^2) \\ & \vdots & \vdots \end{array}$$





Policy Gradient

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau^n) \nabla log p(a_t^n | s_t^n, \theta)$$

Given actor parameter θ

τ^1: (s_1^1, a_1^1) $R(\tau^1)$ **2**

(s_2^1, a_2^1) $R(\tau^1)$ **2**

⋮ ⋮

τ^2: (s_1^2, a_1^2) $R(\tau^2)$ **1**

(s_2^2, a_2^2) $R(\tau^2)$ **1**

⋮ ⋮

Each training data is weighted by $R(\tau^n)$