Lecture-10

MBRL

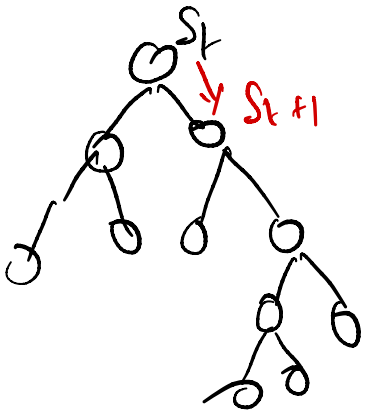
$P(s', r \mid s, a)$

use the model to learn the value fn / Policy

- Dyna-Q.

online Planning to improve the policy

MCTS



Stochastic optimization:

Model-Predictive Control (MPC):

$s_t \rightarrow a_t$

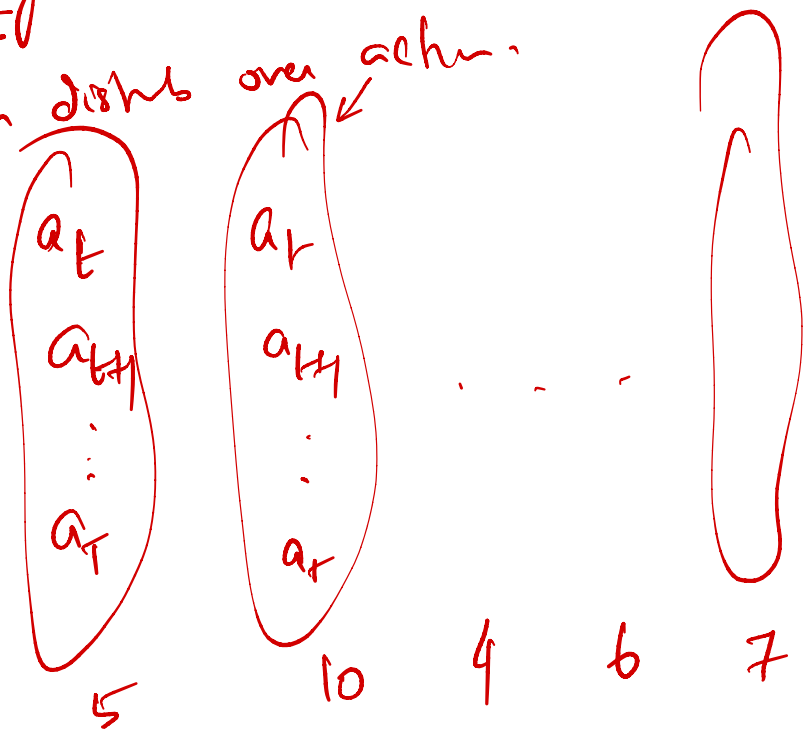
$s_t \quad a_t, a_{t+1} \ldots a_T$

$$\underbrace{a_t, a_{t+1} \ldots a_T}_{A} = \underset{a_t \ldots a_T}{\operatorname{argmax}} \; J(a_t \ldots a_T)$$

Return.

Random Shooting:

① Uniform distrb over actn.



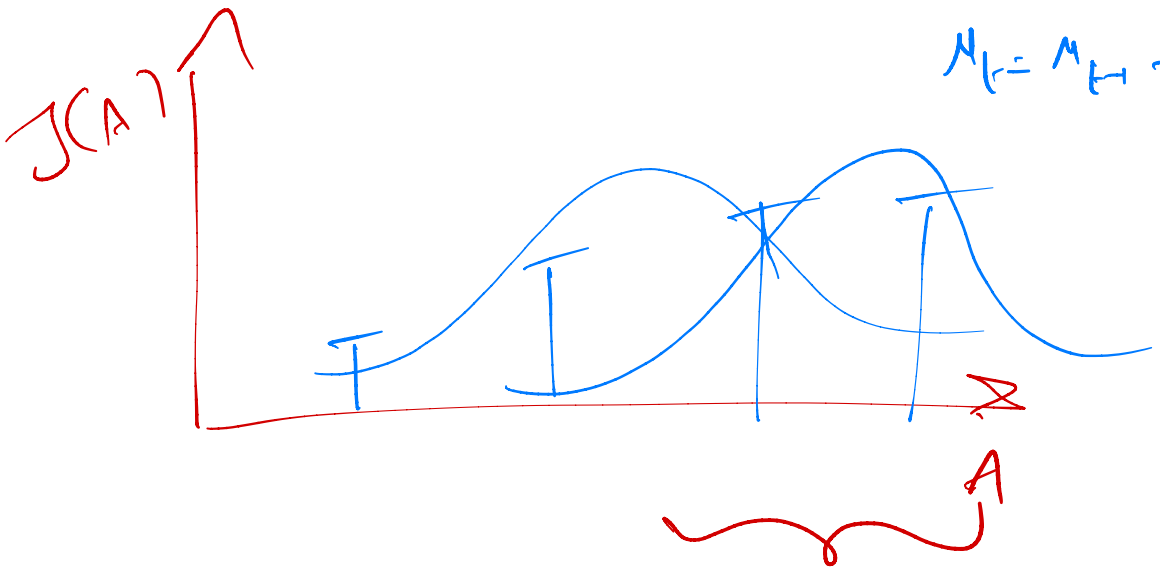
Cross-entropy method:- (CEM)

Gaussian action. μ, σ

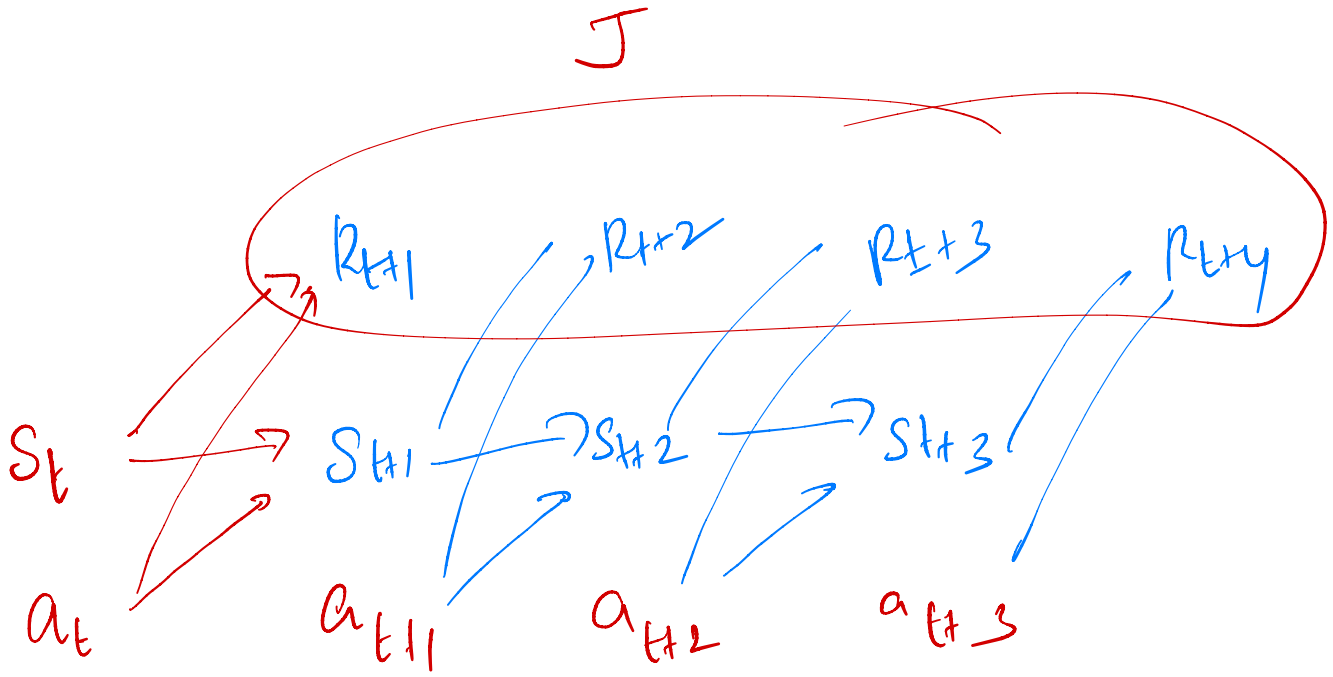
$S_t \rightarrow a_t \; a_{t+1} \; a_{t+2} \; a_{t+3}$

$\mu_1 \sigma_1 \quad \mu_2 \sigma_2 \quad \mu_3 \sigma_3 \quad \mu_4 \sigma_4$

1. Sample $A_1 \ldots A_N$ from $p(A)$
2. evaluate $J(A_1) \ldots J(A_N)$
3. pick M elite candidate $A_{i_1} \ldots A_{i_M}$ with highest value.
4. refit $p(A)$ to Refine $A_1 \ldots A_N$



$\mu_t = \mu_{t-1} + \alpha(\;)$


J
R_{t+1}
R_{t+2}
R_{t+3}
R_{t+4}
S_t
S_{t+1}
S_{t+2}
S_{t+3}
a_t
a_{t+1}
a_{t+2}
a_{t+3}