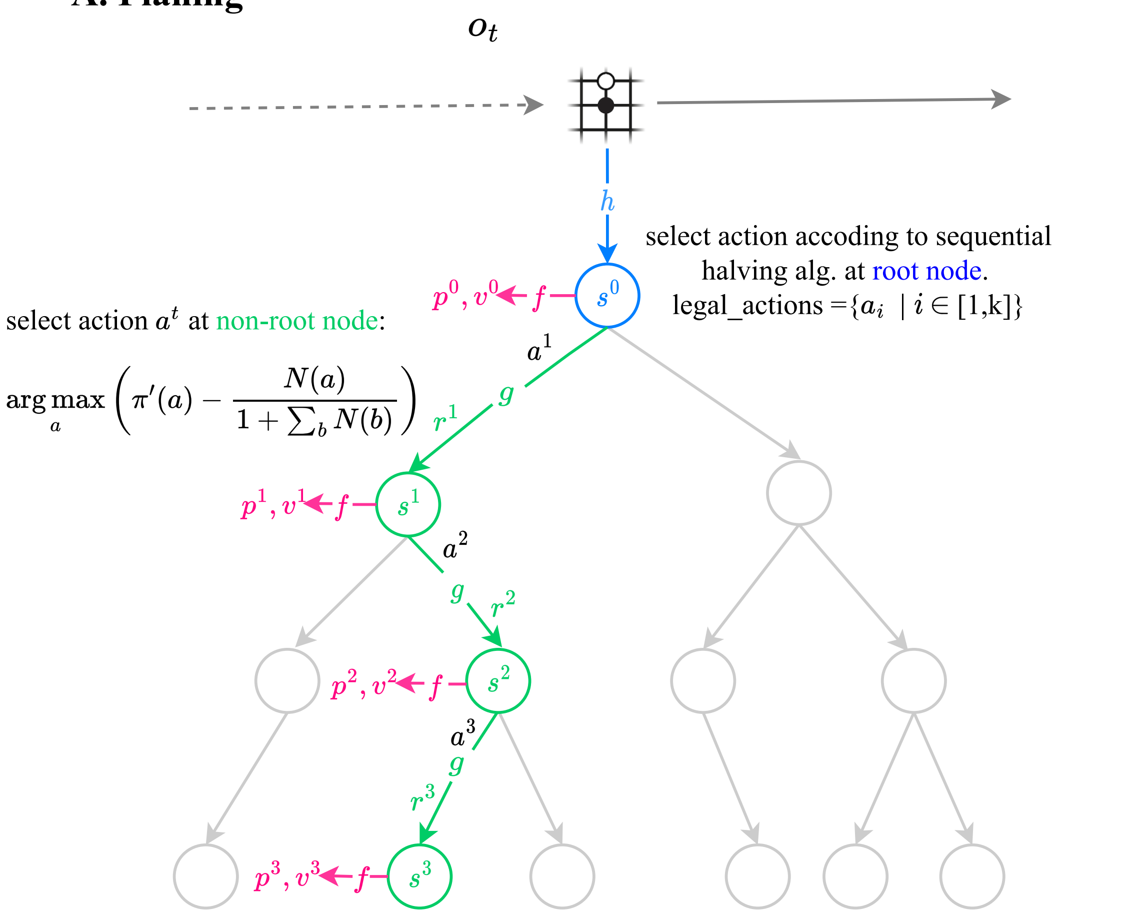
Gumbel MuZero: Planning with few simulations in high dimensional discrete action space

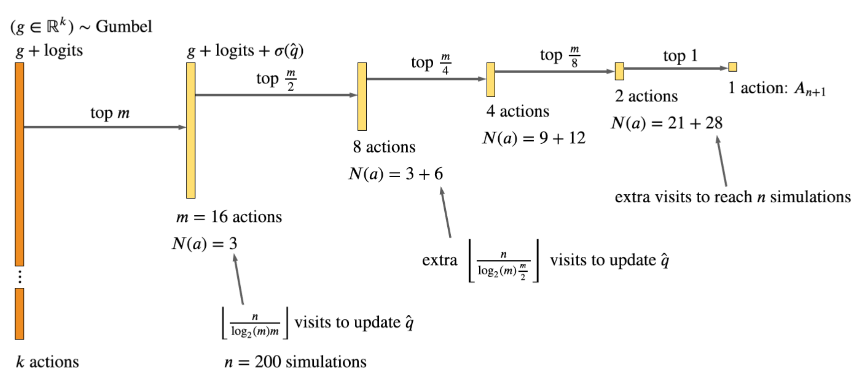
A. Planing



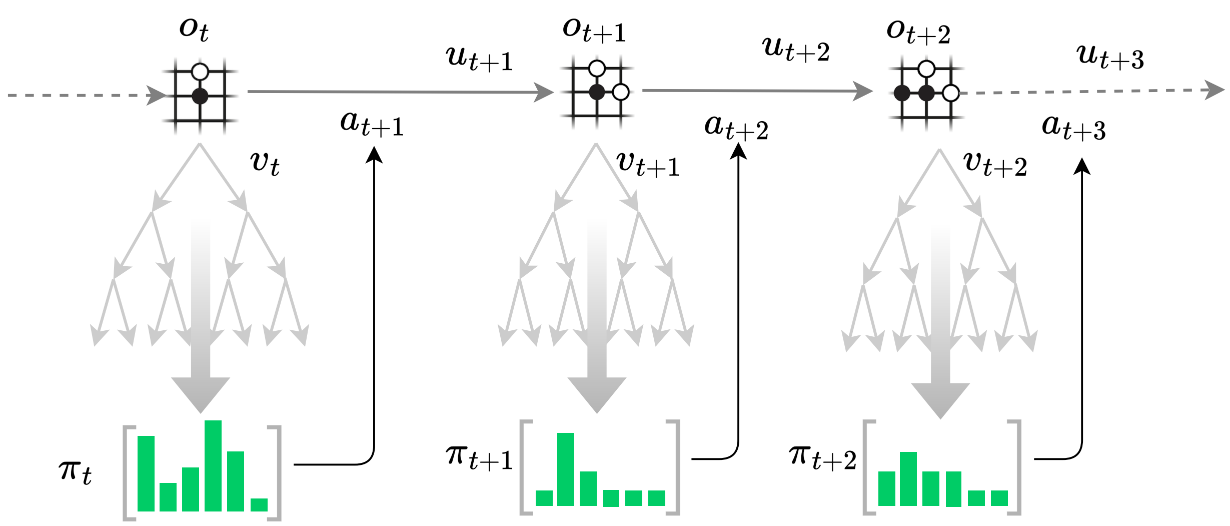
The key issue when the num of simulations n < num of total actions k in MCTS search is **how to choose which actions to vistis and how many times**.

1. we can control the number of actions sampled without replacement.
2. we can use a bandit algorithm to efficiently explore the set of sampled actions.

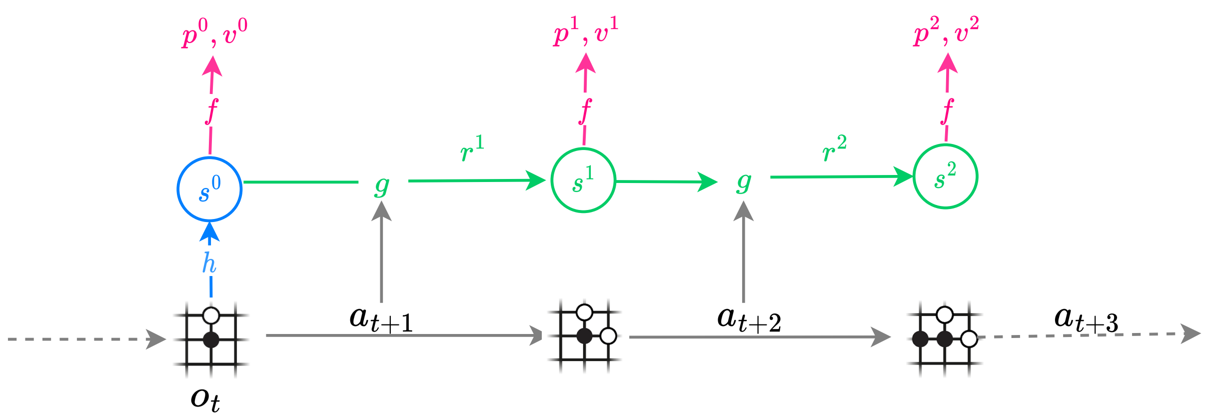
Sequential Halving is used to identify the action with the highest $g(a) + logits(a) + \sigma(\hat{q}(a))$.



B. Acting



C. Training



D. Loss

$$l_t(\theta) = \sum_{k=0}^{K} l^r\left(u_{t+k}, r_t^k\right) + l^v\left(z_{t+k}, v_{t,mix}^k\right) + KL\left(\pi'_{t+k}, \mathbf{p}_t^k\right) + c\|\theta\|^2$$

mixed value approximate:

$$v_{\text{mix}} = \frac{1}{1+\sum_b N(b)}\left(\hat{v}_\pi + \frac{\sum_b N(b)}{\sum_{b\in\{b:N(b)>0\}}\pi(b)}\sum_{a\in\{a:N(a)>0\}}\pi(a)q(a)\right)$$

improved policy distribution: $\pi' = \text{softmax}(\text{logits} + \sigma(\text{ completedQ }))$

$$\text{completedQ}(a) = \begin{cases} q(a) & \text{if } N(a) > 0 \\ v_\pi, & \text{otherwise} \end{cases}$$

NOTE: p is policy distribution , π' is the improved policy distribution ,
completed$Q(a)$ is the completed Q values, $q(a)$ is the MCTS estimated value (for visited actions).