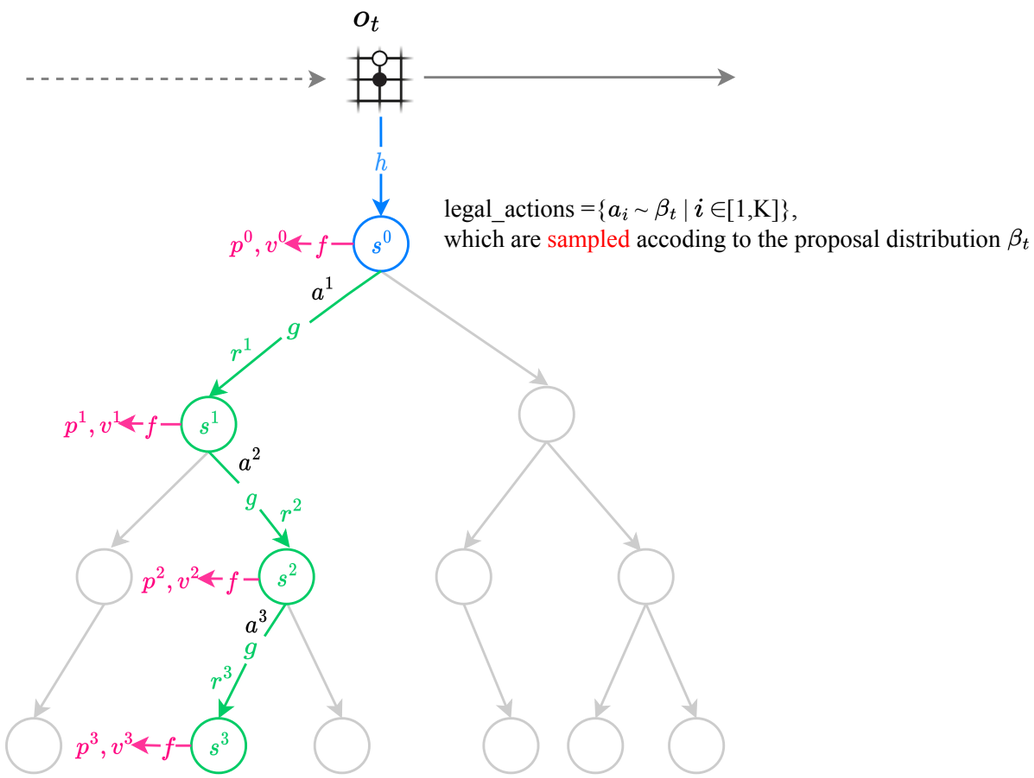
Sampled MuZero: Learning and Planning in Complex Action Spaces (High Dimensional Discrete or Continuous)

A. Planing



legal_actions = $\{a_i \sim \beta_t \mid i \in [1,K]\}$,
which are sampled accoding to the proposal distribution β_t

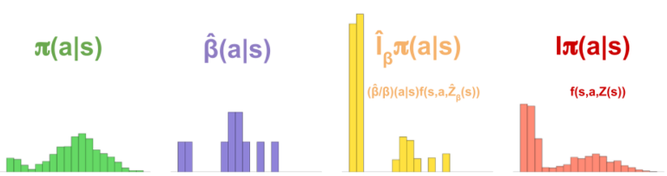
Selection:

In each node, agent select action a^k accordingto the UCB score:

$$a^k = \arg\max_a \left[Q(s,a) + \hat{\beta}/\beta p(s,a) \cdot \frac{\sqrt{\sum_b N(s,b)}}{1+N(s,a)} \left(c_1 + \log \left(\frac{\sum_b N(s,b) + c_2 + 1}{c_2} \right) \right) \right]$$

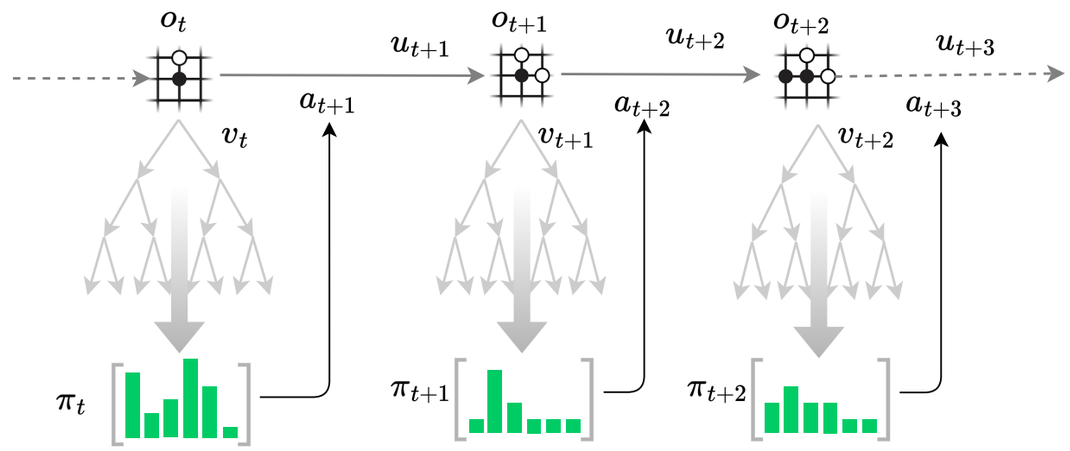
We usually use $\beta = p$, thus $\hat{\beta}/\beta p(s,a) = \hat{\beta}(s,a)$.

Sampled-based Policy Improvement.

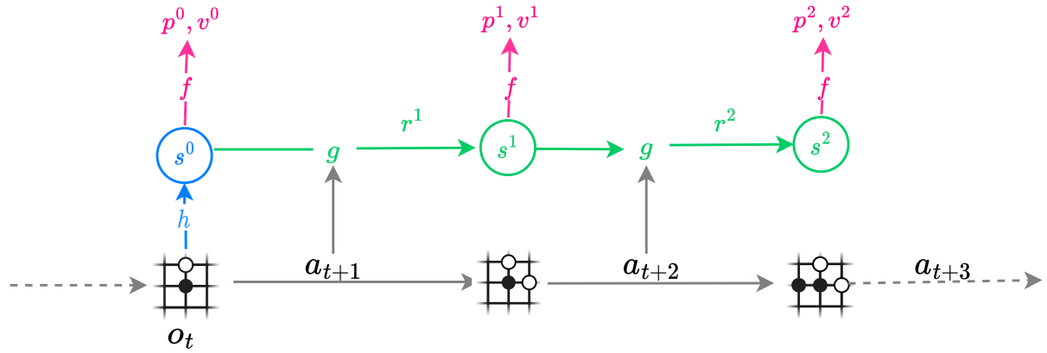


From left to right, current policy $\pi(a|s)$, the empirical distribution $\hat{\beta}(a|s)$, the sample-based improved policy $\hat{I}_\beta \pi(a|s)$, the improved policy $I\pi(a|s)$, respectively.As the number of samples K increases, $\hat{I}_\beta \pi(a|s)$ converges to $I\pi(a|s)$.

B. Acting



C. Training



D. Loss

$$l_t(\theta) = \sum_{k=0}^{K} l^r \left(u_{t+k}, r_t^k \right) + l^v \left(z_{t+k}, v_t^k \right) + KL \left(\pi_{t+k}, \mathbf{p}_t^k \right) + c\|\theta\|$$

where, π_{t+k} is the MCTS search policy (normalized visit counts), also called the sampled-based improved policy $\hat{I}_\beta \pi(a|s)$, which is a discrete categorical distribution.

$\mathbf{p}_t^k$ is the predicted (potentially continuous) policy distribution. The policy loss (KL divergence) is calculated in the **sampled** actions.

NOTE： $\mathbf{p}$ (aka. π)is predicted policy (potentially continuous) distribution , β is the proposal policy distribution (typically equals π), $\hat{\beta}$ is the empirical policy distribution.