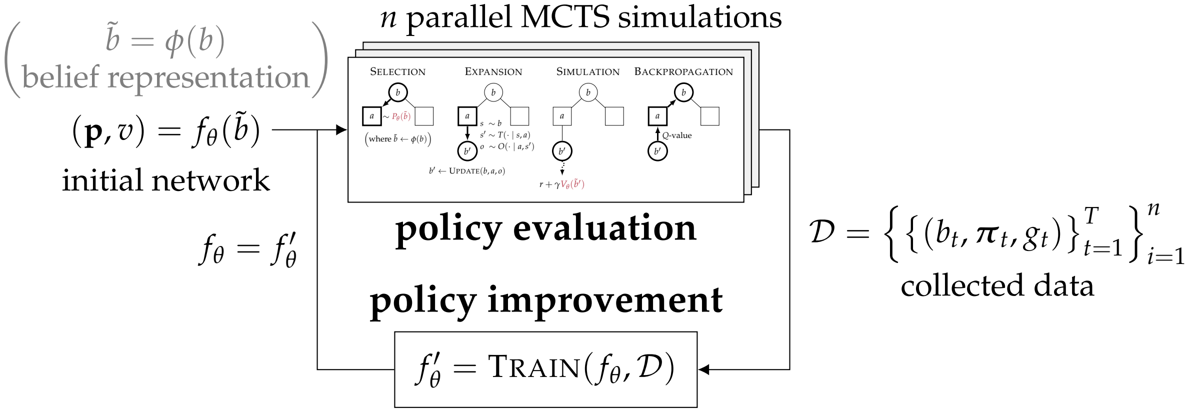n parallel MCTS simulations
$\begin{pmatrix}\tilde{b} = \phi(b) \\ \text{belief representation}\end{pmatrix}$
$(\mathbf{p}, v) = f_\theta(\tilde{b})$
initial network
SELECTION
EXPANSION
SIMULATION
BACKPROPAGATION
b
a
$\sim P_\theta(\tilde{b})$
$\left(\text{where } \tilde{b} \leftarrow \phi(b)\right)$
b
a
$s \sim b$
$s' \sim T(\cdot \mid s, a)$
$o \sim O(\cdot \mid a, s')$
b'
$b' \leftarrow \textsc{Update}(b, a, o)$
b
a
b'
$r + \gamma V_\theta(\tilde{b}')$
b
a
Q-value
b'
policy evaluation
$f_\theta = f'_\theta$
$\mathcal{D} = \left\{ \left\{ (b_t, \boldsymbol{\pi}_t, g_t) \right\}_{t=1}^{T} \right\}_{i=1}^{n}$
collected data
policy improvement
$f'_\theta = \textsc{Train}(f_\theta, \mathcal{D})$