Instead of predicting expected returns, we could train the method to directly predict the policy

$$\pi(a|s; \theta).$$

Obtaining the full distribution over all actions would also allow us to sample the actions according to the distribution $\pi$ instead of just $\varepsilon$-greedy sampling.

However, to train the network, we maximize the expected return $v_\pi(s)$ and to that account we need to compute its gradient $\nabla_\theta v_\pi(s)$. 
In addition to discarding $\epsilon$-greedy action selection, policy gradient methods allow producing policies which are by nature stochastic, as in card games with imperfect information, while the action-value methods have no natural way of finding stochastic policies (distributional RL might be of some use though).

$$J(\theta) = v_{\pi_{\theta}}(S)$$

Policy Gradient Theorem

Let $\pi(a|s; \theta)$ be a parametrized policy. We denote the initial state distribution as $h(s)$ and the on-policy distribution under $\pi$ as $\mu(s)$. Let also $J(\theta) \overset{\text{def}}{=} \mathbb{E}_{h,\pi} v_\pi(s)$.

Then

$$
\nabla_\theta v_\pi(s) \propto \sum_{s' \in S} P(s \rightarrow \ldots \rightarrow s'|\pi) \sum_{a \in A} q_\pi(s', a) \nabla_\theta \pi(a|s'; \theta)
$$

and

$$
\nabla_\theta J(\theta) \propto \sum_{s \in S} \mu(s) \sum_{a \in A} q_\pi(s, a) \nabla_\theta \pi(a|s; \theta),
$$

where $P(s \rightarrow \ldots \rightarrow s'|\pi)$ is probability of transitioning from state $s$ to $s'$ using 0, 1, ... steps.
Proof of Policy Gradient Theorem

\[ \nabla v_\pi(s) = \nabla \left[ \sum_a \pi(a|s; \theta) q_\pi(s, a) \right] \]

\[ = \sum_a \left[ \nabla \pi(a|s; \theta) q_\pi(s, a) + \pi(a|s; \theta) \nabla \left( \sum_{s'} p(s'|s, a)(r + v_\pi(s')) \right) \right] \]

\[ = \sum_a \left[ \nabla \pi(a|s; \theta) q_\pi(s, a) + \pi(a|s; \theta) \left( \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right) \right] \]

We now expand \( v_\pi(s') \).

\[ = \sum_a \left[ \nabla \pi(a|s; \theta) q_\pi(s, a) + \pi(a|s; \theta) \left( \sum_{s'} p(s'|s, a) \nabla v_\pi(s') \right) \right] \]

Continuing to expand all \( v_\pi(s'') \), we obtain the following:

\[ \nabla v_\pi(s) = \sum_{s' \in S} \sum_{a \in A} \nabla \theta \pi(a|s'; \theta). \]
Proof of Policy Gradient Theorem

Recall that the initial state distribution is \( h(s) \) and the on-policy distribution under \( \pi \) is \( \mu(s) \).

If we let \( \eta(s) \) denote the number of time steps spent, on average, in state \( s \) in a single episode, we have

\[
\eta(s) = h(s) + \sum_{s'} \eta(s') \sum_a \pi(a|s') p(s'|s,a).
\]

The on-policy distribution is then the normalization of \( \eta(s) \):

\[
\mu(s) \overset{\text{def}}{=} \frac{\eta(s)}{\sum_{s'} \eta(s')}.
\]

The last part of the policy gradient theorem follows from the fact that \( \mu(s) \) is

\[
\mu(s) = \mathbb{E}_{s_0 \sim h(s)} P(s_0 \rightarrow \ldots \rightarrow s|\pi).
\]
The REINFORCE algorithm (Williams, 1992) uses directly the policy gradient theorem, minimizing \(-J(\theta) \overset{\text{def}}{=} -\mathbb{E}_{h,\pi} v_\pi(s)\). The loss gradient is then

\[
\nabla_\theta - J(\theta) \propto - \sum_{s \in S} \mu(s) \sum_{a \in A} q_\pi(s, a) \nabla_\theta \pi(a|s; \theta)
\]

\[
= -\mathbb{E}_{s \sim \mu} \sum_{a \in A} q_\pi(s, a) \nabla_\theta \pi(a|s; \theta).
\]

However, the sum over all actions is problematic. Instead, we rewrite it to an expectation which we can estimate by sampling:

\[
\nabla_\theta - J(\theta) \propto \mathbb{E}_{s \sim \mu} \mathbb{E}_{a \sim \pi} q_\pi(s, a) \nabla_\theta \pi(a|s; \theta) - \ln \pi(a|s; \theta),
\]

where we used the fact that

\[
\nabla_\theta \ln \pi(a|s; \theta) = \frac{1}{\pi(a|s; \theta)} \nabla_\theta \pi(a|s; \theta).
\]
REINFORCE therefore maximizes the expected return by ascending the gradient

\[ \mathbb{E}_{s \sim \mu} \mathbb{E}_{a \sim \pi} q_\pi(s, a) \nabla_\theta - \ln \pi(a|s; \theta), \]

estimating the \( q_\pi(s, a) \) by a single sample.

Note that the loss is just a weighted variant of negative log likelihood (NLL), where the sampled actions play a role of gold labels and are weighted according to their return.

**REINFORCE Algorithm**

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for \( \pi_* \)

| Input: a differentiable policy parameterization \( \pi(a|s, \theta) \) |
| Algorithm parameter: step size \( \alpha > 0 \) |
| Initialize policy parameter \( \theta \in \mathbb{R}^{d'} \) (e.g., to 0) |

Loop forever (for each episode):

- Generate an episode \( S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T \), following \( \pi(\cdot|\cdot; \theta) \) |
- Loop for each step of the episode \( t = 0, 1, \ldots, T - 1 \):
  \[
  G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \\
  \theta \leftarrow \theta + \alpha G \nabla \ln \pi(A_t|S_t; \theta)
  \]

Modified from Algorithm 13.3 of "Reinforcement Learning: An Introduction, Second Edition" by removing \( \gamma^t \) from the update of \( \theta \).
REINFORCE with Baseline

The returns can be arbitrary – better-than-average and worse-than-average returns cannot be recognized from the absolute value of the return.

Hopefully, we can generalize the policy gradient theorem using a baseline $b(s)$ to

$$\nabla_{\theta} J(\theta) \propto \sum_{s \in S} \mu(s) \sum_{a \in A} \left( q_{\pi}(s, a) - b(s) \right) \nabla_{\theta} \pi(a \mid s; \theta).$$

The baseline $b(s)$ can be a function or even a random variable, as long as it does not depend on $a$, because

$$\sum_{a} b(s) \nabla_{\theta} \pi(a \mid s; \theta) = b(s) \sum_{a} \nabla_{\theta} \pi(a \mid s; \theta) = b(s) \nabla 1 = 0.$$
A good choice for $b(s)$ is $v_\pi(s)$, which can be shown to minimize variance of the estimator. Such baseline reminds centering of returns, given that

$$v_\pi(s) = \mathbb{E}_{a \sim \pi} q_\pi(s, a).$$

Then, better-than-average returns are positive and worse-than-average returns are negative. The resulting $q_\pi(s, a) - v_\pi(s)$ function is also called an *advantage function*

$$a_\pi(s, a) \overset{\text{def}}{=} q_\pi(s, a) - v_\pi(s).$$

Of course, the $v_\pi(s)$ baseline can be only approximated. If neural networks are used to estimate $\pi(a|s; \theta)$, then some part of the network is usually shared between the policy and value function estimation, which is trained using mean square error of the predicted and observed return.
REINFORCE with Baseline (episodic), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \theta)$
Input: a differentiable state-value function parameterization $\hat{v}(s,w)$
Algorithm parameters: step sizes $\alpha^\theta > 0$, $\alpha^w > 0$
Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state-value weights $w \in \mathbb{R}^d$ (e.g., to 0)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \theta)$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
  
  \[ G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \]
  
  \[ \delta \leftarrow G - \hat{v}(S_t, w) \]
  
  \[ w \leftarrow w + \alpha^w \delta \nabla \hat{v}(S_t, w) \]
  
  \[ \theta \leftarrow \theta + \alpha^\theta \delta \nabla \ln \pi(A_t|S_t, \theta) \]

Modified from Algorithm 13.4 of "Reinforcement Learning: An Introduction, Second Edition" by removing $\gamma^t$ from the update of $\theta$. 
REINFORCE with Baseline

$G_0$
Total reward on episode averaged over 100 runs

Figure 13.2 of "Reinforcement Learning: An Introduction, Second Edition".
Actor-Critic

It is possible to combine the policy gradient methods and temporal difference methods, creating a family of algorithms usually called actor-critic methods.

The idea is straightforward – instead of estimating the episode return using the whole episode rewards, we can use \( n \)-step temporal difference estimation.
One-step Actor–Critic (episodic), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \theta)$
Input: a differentiable state-value function parameterization $\hat{v}(s,w)$
Parameters: step sizes $\alpha^\theta > 0$, $\alpha^w > 0$
Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state-value weights $w \in \mathbb{R}^d$ (e.g., to $0$)
Loop forever (for each episode):
  Initialize $S$ (first state of episode)
  
  Loop while $S$ is not terminal (for each time step):
  
  $A \sim \pi(\cdot|S, \theta)$
  Take action $A$, observe $S', R$
  $\delta \leftarrow R + \gamma \hat{v}(S',w) - \hat{v}(S,w)$  \quad \text{(if $S'$ is terminal, then $\hat{v}(S',w) = 0$)}
  $w \leftarrow w + \alpha^w \delta \nabla \hat{v}(S,w)$
  $\theta \leftarrow \theta + \alpha^\theta \delta \nabla \ln \pi(A|S, \theta)$
  $S \leftarrow S'$

Modified from Algorithm 13.5 of "Reinforcement Learning: An Introduction, Second Edition" by removing I.
Asynchronous Methods for Deep RL

A 2015 paper from Volodymyr Mnih et al., the same group as DQN.

The authors propose an asynchronous framework, where multiple workers share one neural network, each training using either an off-line or on-line RL algorithm.

They compare 1-step Q-learning, 1-step Sarsa, $n$-step Q-learning and A3C (an *asynchronous advantage actor-critic* method). For A3C, they compare a version with and without LSTM.

The authors also introduce *entropy regularization term* $\beta H(\pi(s; \theta))$ to the loss to support exploration and discourage premature convergence.
Algorithm 1: Asynchronous one-step Q-learning - pseudocode for each actor-learner thread.

// Assume global shared $\theta$, $\theta^-$, and counter $T = 0$.  
Initialize thread step counter $t \leftarrow 0$  
Initialize target network weights $\theta^- \leftarrow \theta$  
Initialize network gradients $d\theta \leftarrow 0$  
Get initial state $s$  
repeat  
  Take action $a$ with $\epsilon$-greedy policy based on $Q(s, a; \theta)$  
  Receive new state $s'$ and reward $r$  
  $y = \begin{cases} 
    r & \text{for terminal } s' \\
    r + \gamma \max_{a'} Q(s', a'; \theta^-) & \text{for non-terminal } s'
  \end{cases}$  
  Accumulate gradients wrt $\theta$: $d\theta \leftarrow d\theta + \frac{\partial(y - Q(s, a; \theta))^2}{\partial \theta}$  
  $s = s'$  
  $T \leftarrow T + 1$ and $t \leftarrow t + 1$  
  if $T \mod I_{target} == 0$ then  
    Update the target network $\theta^- \leftarrow \theta$  
  end if  
  if $t \mod I_{AsyncUpdate} == 0$ or $s$ is terminal then  
    Perform asynchronous update of $\theta$ using $d\theta$.  
    Clear gradients $d\theta \leftarrow 0$.  
  end if  
until $T > T_{max}$

Algorithm 1 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.
Algorithm S2 Asynchronous n-step Q-learning - pseudocode for each actor-learner thread.

```plaintext
Algorithm S2 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.
```
Asynchronous Methods for Deep RL

Algorithm S3 Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

// Assume global shared parameter vectors θ and θ_v and global shared counter T = 0
// Assume thread-specific parameter vectors θ' and θ_v'
Initialize thread step counter t ← 1
repeat
    Reset gradients: dθ ← 0 and dθ_v ← 0.
    Synchronize thread-specific parameters θ' = θ and θ'_v = θ_v
    t_start = t
    Get state s_t
    repeat
        Perform a_t according to policy π(a_t | s_t; θ')
        Receive reward r_t and new state s_{t+1}
        t ← t + 1
        T ← T + 1
    until terminal s_t or t − t_start == t_{max}
    R = \{ 0 for terminal s_t
        \}
    V(s_t; θ_v') for non-terminal s_t// Bootstrap from last state
    for i ∈ \{t − 1, ..., t_{start}\} do
        R ← r_i + γR
        Accumulate gradients wrt θ': dθ ← dθ + \nabla_θ \log π(a_i | s_i; θ') (R − V(s_i; θ_v'))
        Accumulate gradients wrt θ_v: dθ_v ← dθ_v + \partial (R − V(s_i; θ_v')) / \partial θ_v' 
    end for
    Perform asynchronous update of θ using dθ and of θ_v using dθ_v.
until T > T_{max}

Algorithm S3 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.
All methods performed updates every 5 actions \( t_{\text{max}} = I_{\text{AsyncUpdate}} = 5 \), updating the target network each 40,000 frames.

The Atari inputs were processed as in DQN, using also action repeat 4.

The network architecture is: 16 filters \( 8 \times 8 \) stride 4, 32 filters \( 4 \times 4 \) stride 2, followed by a fully connected layer with 256 units. All hidden layers apply a ReLU non-linearity. Values and/or action values were then generated from the (same) last hidden layer.

The LSTM methods utilized a 256-unit LSTM cell after the dense hidden layer.

All experiments used a discount factor of \( \gamma = 0.99 \) and used RMSProp with momentum decay factor of 0.99.
Asynchronous Methods for Deep RL

Figure 1 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.

Table 1 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.

Table 2 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Time</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN</td>
<td>8 days on GPU</td>
<td>121.9%</td>
<td>47.5%</td>
</tr>
<tr>
<td>Gorila</td>
<td>4 days, 100 machines</td>
<td>215.2%</td>
<td>71.3%</td>
</tr>
<tr>
<td>D-DQN</td>
<td>8 days on GPU</td>
<td>332.9%</td>
<td>110.9%</td>
</tr>
<tr>
<td>Dueling D-DQN</td>
<td>8 days on GPU</td>
<td>343.8%</td>
<td>117.1%</td>
</tr>
<tr>
<td>Prioritized DQN</td>
<td>8 days on GPU</td>
<td>463.6%</td>
<td>127.6%</td>
</tr>
<tr>
<td>A3C, FF</td>
<td>1 day on CPU</td>
<td>344.1%</td>
<td>68.2%</td>
</tr>
<tr>
<td>A3C, LSTM</td>
<td>4 days on CPU</td>
<td>496.8%</td>
<td>116.6%</td>
</tr>
<tr>
<td>A3C</td>
<td>4 days on CPU</td>
<td>623.0%</td>
<td>122.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of threads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-step Q</td>
<td>1.0 3.0 6.3 13.3 24.1</td>
</tr>
<tr>
<td>1-step SARSA</td>
<td>1.0 2.8 5.9 13.1 22.1</td>
</tr>
<tr>
<td>n-step Q</td>
<td>1.0 2.7 5.9 10.7 17.2</td>
</tr>
<tr>
<td>A3C</td>
<td>1.0 2.1 3.7 6.9 12.5</td>
</tr>
</tbody>
</table>
Asynchronous Methods for Deep RL

Figure 3 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.
Asynchronous Methods for Deep RL

Figure 4 of the paper "Asynchronous Methods for Deep Reinforcement Learning" by Volodymyr Mnih et al.
Asynchronous Methods for Deep RL

Figure 2 of the paper “Asynchronous Methods for Deep Reinforcement Learning” by Volodymyr Mnih et al.