Evolutionary Policy Iteration under a Sampling Regime for Stochastic Combinatorial Optimization

Lauren A. Hannah, and Warren B. Powell, Member, IEEE

Abstract

This article modifies the evolutionary policy selection algorithm of Chang et al. (2005, 2007) [1], [2], which was designed for use in infinite horizon Markov decision processes (MDPs) with a large action space to a discrete stochastic optimization problem, in an algorithm called Evolutionary Policy Iteration-Monte Carlo (EPI-MC). EPI-MC allows EPI to be used in a stochastic combinatorial optimization setting with a finite action space and a noisy cost (value) function by introducing a sampling schedule. Convergence of EPI-MC to the optimal action is proven and experimental results are given.

Index Terms

Evolutionary Policy Iteration, Monte Carlo, stochastic optimization, combinatorial optimization.

I. INTRODUCTION

This article proposes a convergent algorithm for one-stage stochastic combinatorial optimization problems, based on a modification of Evolutionary Policy Iteration (EPI) of Chang et al. (2005) [1] and Chang et al. (2007) [2]. Evolutionary Policy Iteration is an algorithm designed to solve infinite horizon Markov decision process problems with large action spaces and small

E-mail: lhannah@princeton.edu

W.B. Powell is with the Department of Operations Research and Financial Engineering, Princeton University.

This research was supported in part by grant AFOSR-F49620-93-1-0098 from the Air Force Office of Scientific Research.
state spaces. It accommodates large action spaces by using policy switching, a novel method that generates an elite policy every iteration based on a small number of candidate policies, and mutation, which allows for a global search of the policy space. EPI was originally designed for infinite horizon Markov decision processes, but we have taken the algorithm structure and applied it to the situation where we start in a fixed state at time 0, have a single period time horizon, and transition to a state in a set with potentially infinite dimensionality. EPI also assumes the ability to quickly compute expected costs, a strong assumption that our modification relaxes. This article presents an algorithm, Evolutionary Policy Iteration–Monte Carlo (EPI-MC) that carries the same basic structure as EPI but is used for stochastic combinatorial optimization and allows Monte Carlo estimation of expectations.

Our problem setting is as follows: after beginning at a fixed state $x_0 \in X_0$, an action $a \in A$ is taken, where $A$ is a finite set of actions. Based on $x_0$ and $a$, exogenous information $W$ occurs, where $W$ is a random variable. It is used to transition the state from $x_0$ to $x_1 \in X_1$. Note that $|X_0| = 1$ and $|X_1|$ is much larger, possibly infinite. The objective is to minimize the expected value of the cost function $R$ at time 1. That is,

$$\min_{a \in A} \mathbb{E} \, R(x_1(a, W)).$$

The difficulty in this problem arises in two areas: first, traversing the set of actions $A$, which can be quite large, and computing the expectation $\mathbb{E} \, R(x_1(a, W))$ for a given $a$.

There are a number of provably convergent algorithms for stochastic combinatorial optimization, including simulated annealing [3], the stochastic ruler [4], nested partitions [5] and ant colony optimization [6]. To the best of the authors’ knowledge, evolutionary algorithms have not been applied to stochastic combinatorial optimization. We modified EPI for this situation because EPI is a genetic algorithm, has a proof of convergence in the non-stochastic case and has a tunable local search.

Section II of this paper will extend EPI to the case outlined in Equation (1) with an algorithm called Evolutionary Policy Iteration–Monte Carlo (EPI-MC). It is a finite-horizon MDP where $T = 1$. Furthermore, the action space has been limited to a finite set and the outcome space is allowed to be infinite, given a fixed starting state. Section III gives a proof of convergence for EPI-MC. Section IV gives numerical results for a solid oxide fuel cell R&D problem, and
Section V is a discussion.

II. EPI-MC

The algorithm that we propose relies on the basic structure of Evolutionary Policy Iteration. EPI has two main features, policy switching and mutations. Policy switching is used to traverse a large action space, generating an elite policy at every iteration based on a small group of candidate policies. This is done at each iteration $n$ by selecting a group of $\Lambda(n)$ policies and calculating the value of each policy $\pi_i^n \in \Lambda(n)$, for each state $x \in \mathcal{X}$, denoted by $V^n_i(x, \pi^n_i(x))$. Note that the state from which we take actions, $x_0$, is fixed and that there is only one time period, so actions are equivalent to policies in this case. In EPI, a new elite policy $\bar{\pi}^n$ is chosen by solving

$$
\bar{\pi}^n(x) = \arg \min_{\pi_i^n \in \Lambda(n)} V^n_i(x, \pi^n_i(x)) \quad \forall x \in \mathcal{X}.
$$

Since $x_0$ is fixed and there is a single time period, we can find an elite action $\bar{a}^n$ by using policy switching,

$$
\bar{a}^n = \arg \min_{a_i^n \in \Lambda(n)} V^n_i(s_0, a_i^n).
$$

In this case, policy switching is equivalent to the more general notion of elitism.

Along with choosing an elite policy by policy switching, EPI chooses a set of candidate policies, $\Lambda(n)$, every iteration by a mixture of local and global mutations of the elite policy. $\Lambda(n)$ includes the elite policy of the previous iteration, $\bar{\pi}^{n-1}$ and $k$ mutations of that policy. This article assumes that local mutations for a policy $\pi$ are drawn randomly from $\mathcal{N}(\pi)$, a predetermined neighborhood around $\pi$. Global mutations are a random choice from $\mathcal{A}$, since actions are equivalent to policies in this case. Each policy $\pi_i^n$ is a local mutation with probability $p$ and a global mutation with probability $1 - p$. Since comparing multiple versions of the same policy is computationally inefficient, steps are taken to ensure that the same policy is not included in the population twice. Again, since our modification is limited to a case where policies are equivalent to actions, actions will replace policies within the algorithm. EPI for the one-stage stochastic combinatorial optimization problem is outlined in Algorithm I.

Both proposals of EPI [1], [2] assume that $\mathbb{E} R(x_1(a, W))$ can be computed without error for every $a \in \mathcal{A}$. [2] note that $\mathbb{E} R(x_1(a, W))$ could be computed via simulation, but does not offer
Algorithm 1: Evolutionary Policy Iteration [1] Applied to One-Stage Stochastic Combinatorial Optimization

1: Initialize $\bar{a}^0$, specify $p$
2: for $n = 1$ to $N$ do
3: Choose $\Lambda(n)$, a population of policies:
4: for $i = 1$ to $k$ do
5: if $u \leq p$, where $u \sim U[0, 1]$ then
6: Choose $a^n_i$ randomly in $\mathcal{N}(\bar{a}^{n-1})$
7: else
8: Choose $a^n_i$ randomly in $\mathcal{A}$
9: end if
10: end for
11: Set $a^{n+1}_k = \bar{a}^{n-1}$
12: for $i = 1$ to $k + 1$ do
13: Set $v^n_i = \mathbb{E} R(x_1(a^n_i, W))$
14: end for
15: Set $\bar{a}^n = a^n_i$, $i = \arg \min_j v^n_j$
16: end for

a sampling routine or convergence results.

Algorithm 2: EPI-MC: Sampling routine to replace lines 12 – 14 of Algorithm [1]

1: Set $v^n_1 = v^n_2 = \cdots = v^n_{k+1} = 0$
2: for $j = 1$ to $\xi_n$ do
3: Get an outcome $\omega^n_j$
4: for $i = 1$ to $k + 1$ do
5: $v^n_i = \frac{i-1}{j} v^n_i + \frac{1}{j} R(x_1(a^n_i, \omega^n_j))$
6: end for
7: end for

EPI-MC modifies EPI to incorporate Monte Carlo estimates by the inclusion of a sampling schedule, $(\xi_n)_{n \geq 0}$. To make the algorithm converge, the number of samples for each action in iteration $n$ must be $O(n^\gamma)$, where $\gamma > 2$. Lines 12 – 14 of Algorithm [1] are replaced by those described in Algorithm [2]

III. Convergence

This section gives the sampling routine and a few conditions on $C$ needed for convergence. We then prove that EPI-MC converges to the optimal solution. Before that can be shown, we
need a few lemmas. To show convergence, we can view EPI-MC as a Markov chain, where the state of the process $Z^n$ is the elite action, $\bar{a}^{n-1}$, and show that at some point $Z^n$ enters $\mathcal{A}^*$, the set of optimal actions, and does not leave. In general, there may be multiple optimal actions. We next prove two technical lemmas that establish a bound on the probability that we choose the optimal solution and a bound on the cumulative density function of the normal distribution. These allow us to then prove the main convergence theorem.

**Lemma 1.** Define $\tau$ as the first time after $n$ where an optimal action, $a^* \in \mathcal{A}^*$ is chosen as the elite action given that $Z^n = z$,

$$\tau = \min_{t \in \{0,1,2,\ldots\}} \{ Z^{n+t} \in \mathcal{A}^* | Z^n = z \}.$$  

Then, for every $n \in \mathbb{N}$ and $z \in \mathcal{A}$, $\tau$ is an almost surely finite stopping time.

**Proof:** At each iteration $n$, $k$ unique actions are chosen, $a^n_1, \ldots, a^n_k$, along with the elite action of the previous iteration, $\bar{a}^{n-1}$. We denote this collection of possible actions during iteration $n$ as $\Lambda(n)$. Let $a^* \in \mathcal{A}^*$ be an optimal action. Note that $\mathbb{P}\{a^* \in \Lambda(n)\} \geq \mathbb{P}\{a^* \in \{a^n_1, \ldots, a^n_k\}|\bar{a}^{n-1} \notin \mathcal{A}^*\}$. The right hand side is bounded from below,

$$\mathbb{P}\{a^* \in \{a^n_1, \ldots, a^n_k\}|\bar{a}^{n-1} \notin \mathcal{A}^*\} \geq \frac{(1 - p)k}{|\mathcal{A}|},$$

where $1 - p$ is the probability of a global mutation and $|\mathcal{A}|$ is the size of the action space. Given that $a^*$ is in $\Lambda(n)$, it is not guaranteed that it is selected as $\bar{a}^n$ (although in regular EPI it is guaranteed). Selection of $a^*$ is not guaranteed in EPI-MC because the value of each choice is estimated by Monte Carlo estimates. Therefore,

$$\mathbb{P}\{a^n \in \mathcal{A}^*|a^* \in \Lambda(n)\} \geq \frac{1}{k + 1},$$

and

$$\mathbb{P}\{\bar{a}^n \in \mathcal{A}^*\} \geq \frac{(1 - p)k}{|\mathcal{A}|(k + 1)}.$$
For every state \( z \in \mathcal{A} \) and \( n \in \mathbb{N} \),
\[
P\{Z^{n+1} \in \mathcal{A}^*|Z^n = z\} \geq \frac{(1-p)k}{|\mathcal{A}|(k+1)} > 0.
\]

We then see that for every \( t = 1, 2, \ldots \)
\[
P\{\tau > t\} = P\{\tau > t, \tau > t-1, \ldots, \tau > 0\}
= P\{\tau > t|\tau > t-1\}
\times P\{\tau > t-1|\tau > t-2\} \ldots P\{\tau > 0\}
\leq \left(1 - \frac{(1-p)k}{|\mathcal{A}|(k+1)}\right)^t
= 0 \text{ as } t \to \infty
\]

The next lemma provides a bound on the cumulative density function of the normal distribution.

**Lemma 2.** There exists \( Y > 0 \) such that for every \( y \geq Y \),
\[
\Phi(y) \geq \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,
\]
where
\[
\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx
\]
is the Gaussian cumulative density function.

**Proof:** There exists a \( Y > 0 \) such that
\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{Y^2}{2}} = \frac{1}{2} e^{-\frac{Y}{2}}.
\]
Since the Laplacian density has heavier tails than the Gaussian density, for every \( y > Y \),
\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \leq \frac{1}{2} e^{-\frac{y}{2}}.
\]
Therefore, for every \( y \geq Y \),

\[
\Phi(y) = 1 - \int_y^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx
\geq 1 - \int_y^\infty \frac{1}{2} e^{-\frac{|x|}{2}} \, dx
= \int_{-\infty}^y \frac{1}{2} e^{-\frac{|x|}{2}} \, dx.
\]

Now the main theorem can be stated.

**Theorem 1.** Suppose that the cost function \( R \) is bounded and the sampling sequence \((\xi_n)_{n \geq 0}\) increases with \( \mathcal{O}(n^\gamma) \), with \( \gamma > 2 \). Then, Algorithm 1 when modified by Algorithm 2 converges to the optimal action set \( \mathcal{A}^* \) such that

\[
\mathbb{P}\{\bar{a}_n \notin \mathcal{A}^*\} = 0
\]

as \( n \to \infty \).

**Proof:** Since \( \mathcal{A} \) is finite, enumerate its elements as \( a_1, a_2, \ldots \). Let \( Z^n \) record the elite action of EPI-MC at iteration \( n \), denoted by \( Z^n = z \) for some \( z \in \mathcal{A} \). Define the random variable

\[
Y^i_j = \frac{1}{\xi_i} \sum_{k=1}^{\xi_i} R(x_1(a_j, W_k))
\]

where the samples \( W_k \) are i.i.d. Let \( \mu_j \) and \( \sigma^2_j \) be the mean and variance, respectively, of the random variable \( R(x_1(a_j, W)) \). Since \( R \) is bounded by assumption, all means, variances and third moments are finite. Define

\[
\mu_{\min} = \min_{j \in \mathcal{A} \backslash a^*} \mu_j, \quad \sigma^2_{\min} = \max_{j \in \mathcal{A}} \sigma^2_j,
\]

and let

\[
Y^i_{\min} = \frac{1}{\xi_i} \sum_{k=1}^{\xi_i} Q_{\min}(W_k),
\]

where \( Q_{\min} \) is a random variable such that \( \mathbb{E} Q_{\min}(W) = \mu_{\min} \) and \( \text{Var}(Q_{\min}(W)) = \sigma^2_{\min} \). Let \( Y^i_{\ast} = \frac{1}{\xi_i} \sum_{k=1}^{\xi_i} R(x_1(a^*, W_k)) \) for \( a^* \in \mathcal{A}^* \) be the random variable associated with actions in
the optimal set. Let $E(R(x_1(a^*, W)) = \mu_*$ and $Var(R(x_1(a^*, W))) = \sigma_*^2$.

Since the first three moments are finite, the Central Limit Theorem states that the estimates of the cost for action $j$ in iteration $i$ converge in distribution to a Gaussian distribution with mean $\mu_j$ and variance $\sigma_j^2$. Moreover, the Berry-Esséen Theorem (BET) [7] states that the deviations from the Gaussian cumulative distribution function are also bounded for every $x \in \mathbb{R}$,

$$\left| \mathbb{P} \left\{ \frac{\sqrt{\xi_i} (Y^i_j - \mu_j)}{\sigma_j} \leq x \right\} - \Phi(x) \right| \leq \frac{B}{\sqrt{\xi_i}},$$

where $B$ is a constant. Since $\mathcal{A}$ has a finite number of elements, $B$ can be chosen such that this holds for all $z \in \mathcal{A}$.

To show Equation (2) holds, it is sufficient to show that for any initial elite decision $z \in \mathcal{A}$, for every $\epsilon > 0$ there exists $N$ such that for every $n \geq N$ and for every $m > n$,

$$\mathbb{P}\{Z^m \in \mathcal{A}^*|Z^n = z\} \geq 1 - \epsilon.$$

Let $N$ satisfy Lemma 2. Fix $\epsilon > 0$ and suppose $Z^n = z$. Using Lemma 1 for every $k \geq N$ and every $z \in \mathcal{A}$:

$$\mathbb{P}\{Z^m \in \mathcal{A}^*|Z^n = z\} \geq \prod_{i=n+\tau}^{\infty} \mathbb{P}\{Z^{i+1} \in \mathcal{A}^*|Z^i \in \mathcal{A}^*\} \geq \prod_{i=N}^{\infty} \mathbb{P}\{Z^{i+1} \in \mathcal{A}^*|Z^i \in \mathcal{A}^*\} \geq \prod_{i=N}^{\infty} \int_{\mathbb{R}} \mathbb{P}\{Y^i_{z+1} \in dx\} \mathbb{P}\{Y^i_{z+1} > x\} \cdots \times \mathbb{P}\{Y^i_{z+k} > x\} \geq \prod_{i=N}^{\infty} \int_{\mathbb{R}} \mathbb{P}\{Y^i_{z+1} \in dx\} (\mathbb{P}\{Y^i_{min} > x\})^k,$$
by the BET \cite{7},

$$\geq \prod_{i=N}^{\infty} \int_{\mathbb{R}} \mathbb{P}\{Y_{i+1}^* \in dx\}$$

$$\times \left(1 - \Phi \left(\frac{\sqrt{\xi_i}(x - \mu_{\text{min}})}{\sigma_{\text{min}}}\right) - \frac{B}{\sqrt{\xi_i}}\right)^k$$

$$\geq \prod_{i=N}^{\infty} \int_{-\infty}^{\mu_++\mu_{\text{min}}} \mathbb{P}\{Y_{i+1}^* \in dx\}$$

$$\times \left(\Phi \left(\frac{\sqrt{\xi_i}(\mu_{\text{min}} - \mu_*)}{2\sigma_{\text{min}}}\right) - \frac{B}{\sqrt{\xi_i}}\right)^k,$$

again, by the BET \cite{7},

$$\geq \prod_{i=N}^{\infty} \left(\Phi \left(\frac{\sqrt{\xi_i}(\mu_{\text{min}} - \mu_*)}{2\sigma_{\text{min}}}\right) - \frac{B}{\sqrt{\xi_i}}\right)^{k+1},$$

by Lemma \cite{2}

$$\geq \prod_{i=N}^{\infty} \left(\int_{-\infty}^{\sqrt{\xi_i}(\mu_{\text{min}} - \mu_*)} \frac{1}{2} e^{-|x|} \, dx - \frac{B}{\sqrt{\xi_i}}\right)^{k+1}$$

$$= \prod_{i=N}^{\infty} \left(1 - \frac{1}{2} e^{-\frac{\sqrt{\xi_i}(\mu_{\text{min}} - \mu_*)}{2\sigma_{\text{min}}}} - \frac{B}{\sqrt{\xi_i}}\right)^{k+1}$$

$$\geq \prod_{i=N}^{\infty} \exp \left(-\frac{k + 1}{2} e^{-d\sqrt{\xi_i}} - \frac{(k + 1)B}{\sqrt{\xi_i}}\right)$$

$$= \exp \left(-\frac{(k + 1)}{2} \sum_{i=N}^{\infty} \left(\frac{e^{-d\sqrt{\xi_i}}}{2} + \frac{B}{\sqrt{\xi_i}}\right)\right),$$

where $d = \frac{\mu_{\text{min}} - \mu_*}{2\sigma_{\text{min}}}$. Since both $\sum_{i=N}^{\infty} e^{-d\sqrt{\xi_i}}$ and $\sum_{i=N}^{\infty} \frac{1}{\sqrt{\xi_i}}$ converge if $(\xi_i)$ increases with $O(i^\gamma)$, $\gamma > 2$, an $N$ can be chosen such that for every $m > n \geq N$

$$\mathbb{P}\{Z^m \in \mathcal{A}^*|Z^n = z\} \geq 1 - \epsilon.$$
### Table I

Mean, standard deviation, minimum and maximum for simulated annealing and EPI-MC under different problem sizes and number of cost function evaluations. EPI-MC consistently outperforms SA.

<table>
<thead>
<tr>
<th>€/kWh</th>
<th>9 of 18, 2,000 It.</th>
<th>9 of 18, 20,000 It.</th>
<th>9 of 18, 10'' It.</th>
<th>20 of 100, 2,000 It.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.66</td>
<td>6.43</td>
<td>1.31</td>
<td>1.16</td>
</tr>
<tr>
<td>Std.</td>
<td>1.03</td>
<td>10.46</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Min</td>
<td>0.57</td>
<td>0.58</td>
<td>0.57</td>
<td>0.53</td>
</tr>
<tr>
<td>Max</td>
<td>7.05</td>
<td>47.70</td>
<td>5.69</td>
<td>6.46</td>
</tr>
</tbody>
</table>

### IV. Numerical Results

EPI-MC was used to solve a one stage R&D portfolio problem applied to solid oxide fuel cells. In the generic one-stage R&D problem, there are $M$ potential research projects that can either be funded or not. Let $a_i \in \{0, 1\}$ be the action to fund project $i$ and let $a = (a_1, \ldots, a_M)$. The action is subject to a budget $b$. That is,

$$
\sum_{i=1}^{M} c_i a_i \leq b,
$$

where $c_i$ is the cost of project $i$. Each project $i$ has associated technology parameters $\rho_{0,i} = (\rho_{0,1}^i, \rho_{0,2}^i, \ldots, \rho_{0,r(i)}^i)$, where $r(i)$ is the total number of technologies associated with project $i$ and the 0 in the subscript denotes the time. If project $i$ is funded, the random vector of technology changes $W_i = (W_{i1}, W_{i2}, \ldots, W_{ir(i)})$ is added to the time 0 technology vector to produce the time 1 technology vector,

$$
\rho_{1,i}^j = \rho_{0,i}^j + a_i W_{ij}.
$$

In this case, $\rho_t$, $t \in \{0, 1\}$ is the state variable. A cost is then calculated based on the entire technology state at time 1, $R(\rho_1)$. The goal is to choose the action which minimizes the expected cost at time 1,

$$
\min \mathbb{E} R(\rho_1(a, W)) \quad (3)
$$

subject to

$$
\sum_{i=1}^{M} c_i a_i \leq b
$$

$$
a_i \in \{0, 1\}, \quad i = 1, \ldots, M.
$$
In a solid oxide fuel cell setting, projects pertain to six fuel cell components: anodes, cathodes, electrolytes, bipolar plates, seals and pressure vessels. Each component has two to three associated technology parameters, and the goal is to produce the cheapest cost per running hour fuel cell within some given running temperature and pressure vessel design specifications. The cost function is non-separable and quite complex.

The possible action space is very large for this problem, and there is no easily exploitable problem structure. EPI-MC was used along with simulated annealing to solve Equation (3). These methods were chosen because both offer theoretical guarantees of convergence, exploit some local structure, and only require minimal memory usage—both store only an elite state and up to k+1 values each iteration.

Empirical quantiles as a function of cost function evaluations of simulated annealing and EPI-MC are given in Figure 1. The number of extremely poor choices seems to diminish much more quickly with EPI-MC than with simulated annealing. A summary of experimental results is given in Table I. Full numerical results are given in [8].

Fig. 1. The 5%, 50%, and 95% quantiles of EPI-MC and simulated annealing estimated values plotted against number of cost function evaluations for a solid oxide fuel cell problem where 7 projects out of 18 are chosen.
V. Discussion

In this paper we used the structure of the EPI algorithm to solve a stochastic combinatorial optimization problem, producing the EPI-MC algorithm. We presented a sampling regime and gave a proof of convergence under that regime along with numerical results in a R&D portfolio optimization test case. Convergence for the stochastic combinatorial optimization problem provides insight into the behavior of EPI in an MDP setting with stochastic rewards. Our analysis suggests that we would need to compute $O(n^\gamma)$, $\gamma > 2$ samples for each state; the computation is not trivial as $|\mathcal{X}|$ gets large. This finding coincides with the original assertion of [1] and [2] that EPI is best in small $|\mathcal{X}|$, large $|\mathcal{A}|$ settings.

References