Mathematical Analysis of Optimality in MCTS-Generated Paths

Introduction

We aim to determine whether an n-period best path constructed using Monte Carlo Tree Search (MCTS), which is a sequence of individual General Equilibrium (GE) optima, satisfies the properties of the Bellman equation. Specifically, we want to verify if each subsection of the path is optimal in the context of the policy maker's dynamic optimization problem.

Problem Setup

Policy Maker's Objective

The policy maker maximizes expected social welfare over an infinite horizon:

$$\max_{\{a_t\}} \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \left(u(S_t) - D(E_t, \theta_t) Y_t\right)\right],$$

where:

- $\beta \in (0, 1)$ is the discount factor.
- S_t is the state at time t.
- a_t is the action (policy decision) at time t.
- $u(S_t)$ is the utility derived from the state S_t.
- $D(E_t, \theta_t) Y_t$ represents damages, with E_t being cumulative emissions and θ_t the damage parameters.
- Y_t is aggregate output at time t.

State Dynamics

The state evolves according to:

$$S_{t+1} = f(S_t, a_t, \epsilon_t),$$

where:

- ϵ_t captures stochastic elements such as new information on damages and technology growth.

- The function f encapsulates the transition dynamics, including the GE model and learning processes.

Markov Decision Process (MDP) Formulation

The problem can be framed as an MDP with:

- **States**: $S_t \in \mathcal{S}$.
- **Actions**: $a_t \in \mathcal{A}(S_t)$.
- **Transition Function**: $S_{t+1} = f(S_t, a_t, \epsilon_t)$.
- **Reward Function**: $R(S_t, a_t) = u(S_t) - D(E_t, \theta_t)Y_t$.

The Markov property holds since S_{t+1} depends only on S_t, a_t, and ϵ_t.

The Bellman Equation

The Bellman equation characterizes the optimal value function $V^*(S_t)$:

$$V^*(S_t) = \max_{a_t \in \mathcal{A}(S_t)} \{R(S_t, a_t) + \beta \mathbb{E}_{\epsilon_t}[V^*(S_{t+1})|S_t, a_t]\}.$$

Bellman Optimality Principle: An optimal policy has the property that, regardless of the initial state and decision, the remaining decisions constitute an optimal policy starting from the state resulting from the first decision.

Monte Carlo Tree Search (MCTS)

Overview

MCTS is a simulation-based algorithm that builds a search tree to find optimal policies:

- **Tree Nodes**: Represent states S_t.
- **Edges**: Represent actions a_t.
- **Simulations**: Used to estimate the value of actions by sampling possible future trajectories.

Properties Relevant to Our Analysis

- **Asymptotic Optimality**: As the number of simulations $N \to \infty$, MCTS converges to the optimal policy.
- **Value Estimation**: Estimates the expected cumulative reward from each state-action pair.
- **Policy Improvement**: Selects actions that maximize the estimated value function.

Analysis of Optimality

We aim to show that the path generated by MCTS satisfies the Bellman equation, ensuring that each subsection is optimal.

Assumptions

1. **Finite Action and State Spaces**: For mathematical tractability, we assume that $\mathcal{A}(S_t)$ and $\mathcal{S}$ are finite.
2. **Bounded Rewards**: The reward function $R(S_t, a_t)$ is bounded.
3. **Discount Factor**: $\beta \in (0, 1)$.
4. **Perfect Simulation**: MCTS can accurately simulate transitions and rewards.

Proof Outline

Step 1: MCTS Approximates the Optimal Value Function

As $N \to \infty$, MCTS estimates converge to the true value function:

$$V_{\text{MCTS}}(S_t) \xrightarrow{N \to \infty} V^*(S_t).$$

Step 2: MCTS Action Selection Mimics the Bellman Equation

At each state S_t, MCTS selects the action:

$$a_t^* = \arg\max_{a_t \in \mathcal{A}(S_t)} \left\{ R(S_t, a_t) + \beta \mathbb{E}_{\epsilon_t}[V_{\text{MCTS}}(S_{t+1}) | S_t, a_t] \right\}.$$

As $V_{\text{MCTS}}(S_{t+1}) \to V^*(S_{t+1})$, this selection satisfies the Bellman equation.

Step 3: Optimality of Subpaths

By the Bellman Optimality Principle, the optimality of a_t^* ensures that the subsequent path from S_{t+1} onward is also optimal. Therefore, any subsection of the path is optimal.

Formal Proof

Proposition

Given the assumptions, the policy π_{MCTS} generated by MCTS converges to the optimal policy π^* as $N \to \infty$, and the value function $V_{\text{MCTS}}(S_t)$ converges to $V^*(S_t)$.

Proof

1. **Convergence of Value Estimates**:

By the Law of Large Numbers, the estimated rewards and value functions in MCTS converge to their expected values as $N \to \infty$:

$$\hat{R}(S_t, a_t) \to \mathbb{E}[R(S_t, a_t)],$$

$$V_{\text{MCTS}}(S_t) \to V^*(S_t).$$

2. **Optimal Action Selection**:
   Since the estimated value function converges to the true value function, the action selection criterion in MCTS aligns with the Bellman equation:

$$a_t^* = \arg\max_{a_t \in \mathcal{A}(S_t)} \left\{ \hat{R}(S_t, a_t) + \beta \hat{V}(S_{t+1}) \right\} \to \arg\max_{a_t \in \mathcal{A}(S_t)} \left\{ R(S_t, a_t) + \beta \mathbb{E}_{\epsilon_t}[V^*(S_{t+1}) | S_t, a_t] \right\}.$$

3. **Bellman Equation Satisfaction**:
   Thus, $V_{\text{MCTS}}(S_t)$ satisfies:

$$V_{\text{MCTS}}(S_t) = R(S_t, a_t^*) + \beta \mathbb{E}_{\epsilon_t}[V_{\text{MCTS}}(S_{t+1}) | S_t, a_t^*].$$

4. **Optimality of Subpaths**:
   By induction, starting from any state S_t, the policy π_{MCTS} prescribes actions that maximize the expected cumulative reward, ensuring that each subsection is optimal.

Addressing Infinite Horizon with Finite n-Period Path

While the policy maker's problem is over an infinite horizon, MCTS constructs a finite n-period path. We address this by:

- **Truncation with Terminal Value Approximation**:
  We estimate the value beyond period n using a terminal value function $V_{\text{terminal}}(S_n)$.
- **Error Bound**:
  The truncation error decreases with a higher discount factor β and larger n. Formally:

$$\left| V^*(S_0) - V_{\text{MCTS}}^{(n)}(S_0) \right| \leq \frac{\beta^n R_{\max}}{1 - \beta},$$

  where $R_{\max}$ is the maximum possible reward.
- **Sufficiently Large n**:
  By choosing n large enough, the difference becomes negligible, and the finite-horizon solution approximates the infinite-horizon solution.

Conclusion

- **Bellman Equation Satisfaction**:

The path generated by MCTS satisfies the Bellman equation at each state, as the value function and action selection converge to the optimal ones.

- **Optimality of Subsections**:
  By the Bellman Optimality Principle, each subsection of the path is optimal.

Therefore, the n-period best path constructed using MCTS satisfies the properties of the Bellman equation, and each subsection is indeed optimal.

Additional Mathematical Details

Value Function Convergence

As $N \to \infty$:

$$V_{\text{MCTS}}(S_t) = \max_{a_t \in \mathcal{A}(S_t)} \left\{ \hat{R}(S_t, a_t) + \beta \mathbb{E}_{\epsilon_t}[V_{\text{MCTS}}(S_{t+1})] \right\} \to V^*(S_t).$$

Optimal Policy Convergence

The policy π_{MCTS} converges to the optimal policy π^*:

$$\pi_{\text{MCTS}}(S_t) = \arg\max_{a_t \in \mathcal{A}(S_t)} \left\{ \hat{R}(S_t, a_t) + \beta \mathbb{E}_{\epsilon_t}[V_{\text{MCTS}}(S_{t+1})] \right\} \to \pi^*(S_t).$$

Error Bounds

The error in the value function estimation can be bounded:

$$|V_{\text{MCTS}}(S_t) - V^*(S_t)| \leq \frac{C}{\sqrt{N}},$$

where C is a constant depending on the variance of rewards.

Implications for Policy Maker

- **Policy Consistency**: The policy derived from MCTS is consistent with the Bellman equation, ensuring time consistency.
- **Robustness**: Each decision is optimal given the current state, making the policy robust to changes in initial conditions.

Final Remarks

By providing a rigorous mathematical framework, we've demonstrated that the n-period path constructed using MCTS satisfies the properties of the Bellman equation under the given assumptions. Each subsection of the path is optimal because MCTS approximates the optimal value function and policy as the number of simulations increases.

This analysis confirms that using MCTS to chain individual GE optima yields a globally optimal path consistent with dynamic programming principles.