Policy Gradient Planning for Environmental Decision Making with Existing Simulators

Mark Crowley and David Poole

crowley@cs.ubc.ca

University of British Columbia

The Problem
Automated planning in large scale spatiotemporal environmental domains such as forestry. Actions need to be taken at multiple locations at each moment in time.

Why is this hard?

I - Cannot enumerate states or actions
locations/cells (C) = 1000-100,000 actions (A): cut, ncut, ...
features (F): discrete or continuous, 1-30 features

II - Cannot treat locations as independent
- non-local rewards - constraint on total harvest per year, constraint on irregular harvest flow year to year
- spatial constraints - no cutting of adjacent cells, maintaining an age distribution
- spatial dynamics - Mountain Pine Beetle spread

III - Cannot analyse dynamics directly
- FSSAM (Forest Service Spatial Analysis Model) - developed for BC Forest Service to simulate effects of different harvest quotas on forest development.

Equilibrium Policy Gradient Framework

Policy Gradient Planning
Gradient of value function does not require dynamics, only gradient of log policy.
\[ V = \sum_{k \in K} R(k) V_k \]
\[ \nabla_{\theta} V^\theta = \frac{1}{|H|} \sum_{k \in H} R(k) V_k \log \pi(a_k^k | \theta) \]

Policy Gradients

Policy Parameters
\[ \theta_f(a) : \mathcal{F} \times \mathcal{A} \rightarrow \mathbb{R} \]
Features - \( f(a_{a,s}) \)

Equilibrium Landscape Policy
Locations are not independent, so landscape policy has a cyclic structure.

Policy Representation

Policy Parameters
\[ \theta_f(a) : \mathcal{F} \times \mathcal{A} \rightarrow \mathbb{R} \]
Action - Age Max Avail AnyAdj Volume
Cut - 1.65 1.98 0.74 0.29
NoCut - 7.82 6.97 3.85 4.79

Experiments
- 1880 cells
- binary actions (cut, ncut)
- 4 features
- 100 year planning horizon
- 10 policy updates
- 500 MCMC steps after burn-in

Three Reward Models
Provide positive reward for volume cut minus...
- HVR = penalty for irregular harvest volumes over time
- AVR = penalty for irregular available volume of the forest over time
- HAVR = HVR + AVR

Local Policies

Single Generated Chain
Simulated Longer Chains

Standard Deviation from mean Volume for Policy Under Each Value Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Harvest Volume</th>
<th>Available Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVR</td>
<td>18,085</td>
<td>411,085</td>
</tr>
<tr>
<td>HVR</td>
<td>14,422</td>
<td>248,920</td>
</tr>
<tr>
<td>HAVR</td>
<td>20,309</td>
<td>224,212</td>
</tr>
<tr>
<td>HAVR High</td>
<td>50,089</td>
<td>417,278</td>
</tr>
<tr>
<td>AVR Collapsed</td>
<td>72,859</td>
<td>1,125,138</td>
</tr>
</tbody>
</table>

Gibbs Sampling

Typical results from HAVR reward model. Sustainable, low cut plan.

Common local minima from another run of HAVR. More aggressive plan, still sustainable over 100 years.

Unsustainable plan coming from an AVR run. Forest population collapses completely.