Policy Gradient Planning for Environmental Decision Making with Existing Simulators

Mark Crowley and David Poole

University of British Columbia

crowley@cs.ubc.ca poole@cs.ubc.ca

google+ http://gplus.to/crowley

The Problem

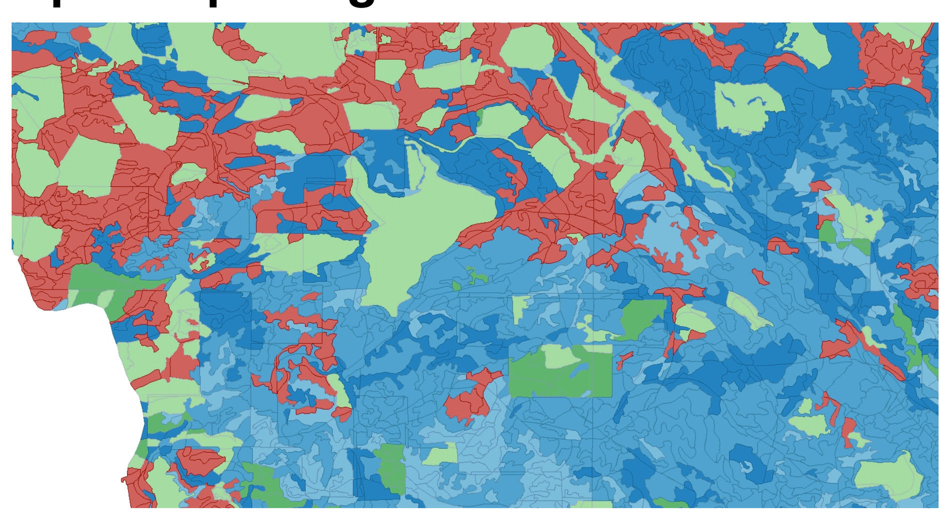
Automated planning in large scale spatiotemporal environmental domains such as forestry.
Actions need to be taken at multiple locations at each moment in time.

Why is this hard?

I - Cannot enumerate states or actions

locations/cells (C) : 1000-100,000
actions (A) : cut, nocut, ...
features $(\mathcal{F})$: discrete or continuous, 1-30 features

Example Map of Age Feature



Age of trees in cell.

0-25	26-50	51-75	76-100	101-150	150-

Scale for 10 Binary Features and Binary Actions

Number of ...	at each cell	entire landscape
actions	2	$2^{1000} \approx 10^{300}$
states	2^{10}	$(2^{10})^{1000} \approx 10^{3000}$

II - Cannot treat locations as independent

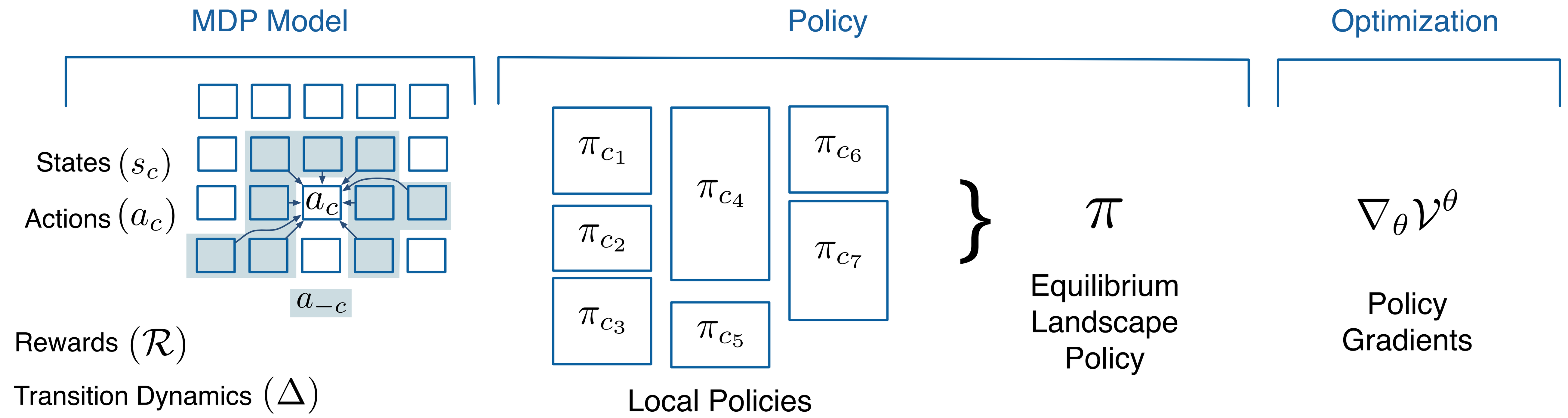
- **non-local rewards** - constraint on total harvest per year, *constraint on irregular harvest flow year to year*
- **spatial constraints** - *no cutting of adjacent cells*, maintaining an age distribution
- **spatial dynamics** - Mountain Pine Beetle spread

III - Cannot analyse dynamics directly

External Simulators

- **black box** - best models are simulators built by researchers in forestry. Designed to explore scenarios by manually adjusting parameters.
- **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.

$$\mathcal{V}^\theta = \sum_{\mathbf{k}\in\mathcal{K}} p(\mathbf{k}|\theta)\mathcal{R}(\mathbf{k})$$

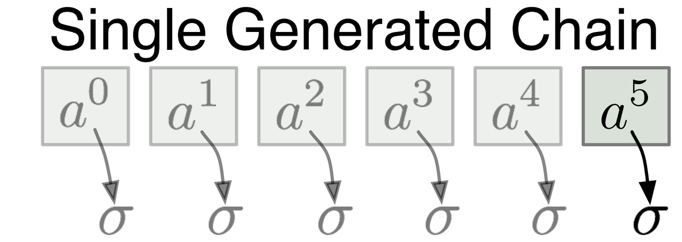
$$\nabla_\theta \mathcal{V}^\theta \approx \frac{1}{|\mathrm{H}|}\sum_{\mathbf{k}\in\mathrm{H}} \mathcal{R}(\mathbf{k})\nabla_\theta \log \pi(\mathbf{a^k}|\mathbf{s^k},\theta)$$

where H is the set the trajectories sampled so far. The PG algorithm updates the policy parameters by following the gradient of the value function.

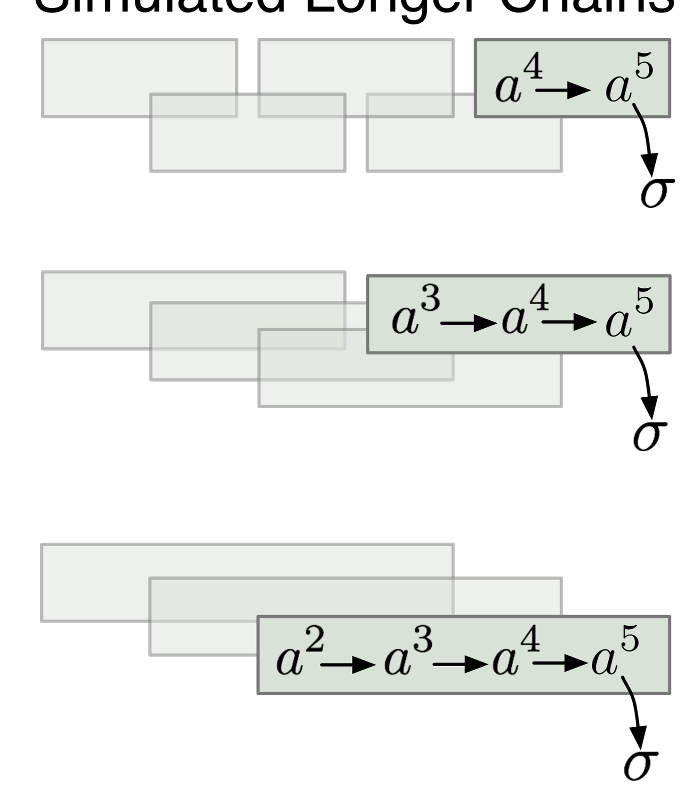
$$\theta' = \theta + \lambda\nabla_\theta\mathcal{V}^\theta$$

Gradient of Landscape Policy
Approximated by generating a Markov chain.

Single Generated Chain



Simulated Longer Chains



Policy Representation

Policy Parameters

$\theta_f(a) : \mathcal{F} \times A \rightarrow \Re$

	Features : $f_c(a_{-c}, \mathbf{s})$			
Action	Age	Max Avail	AnyAdj	Volume
Cut	-1.55	-1.98	0.71	0.29
NoCut	7.82	6.97	3.85	4.79

Local Policy

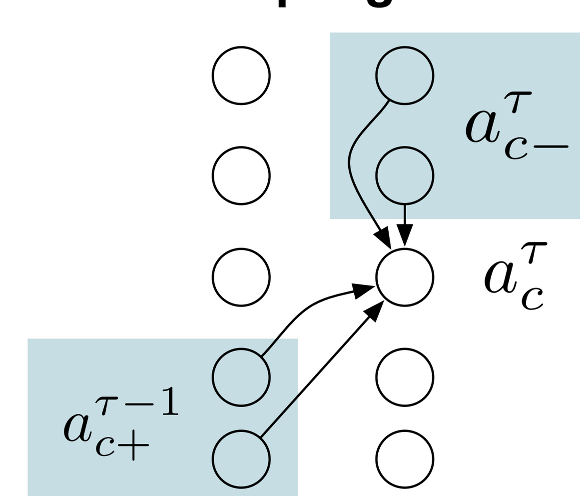
$$\pi_c(a_c|a_{-c},\mathbf{s},\theta) = \frac{\exp(\sum_f \theta_f(a_c)f_c(a_{-c},\mathbf{s}))}{\sum_{b_c\in A}\exp(\sum_f \theta_f(a_c)f_c(a_{-c},\mathbf{s}))}$$

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

The distribution over landscape actions $\mathbf{A}$ is the ***equilibrium of a Markov chain*** where the transitions are defined by the local policy where:

$$a_{-c} = a^{\tau-1}_{c+} \cup a^{\tau}_{c-}$$

Gibbs Sampling



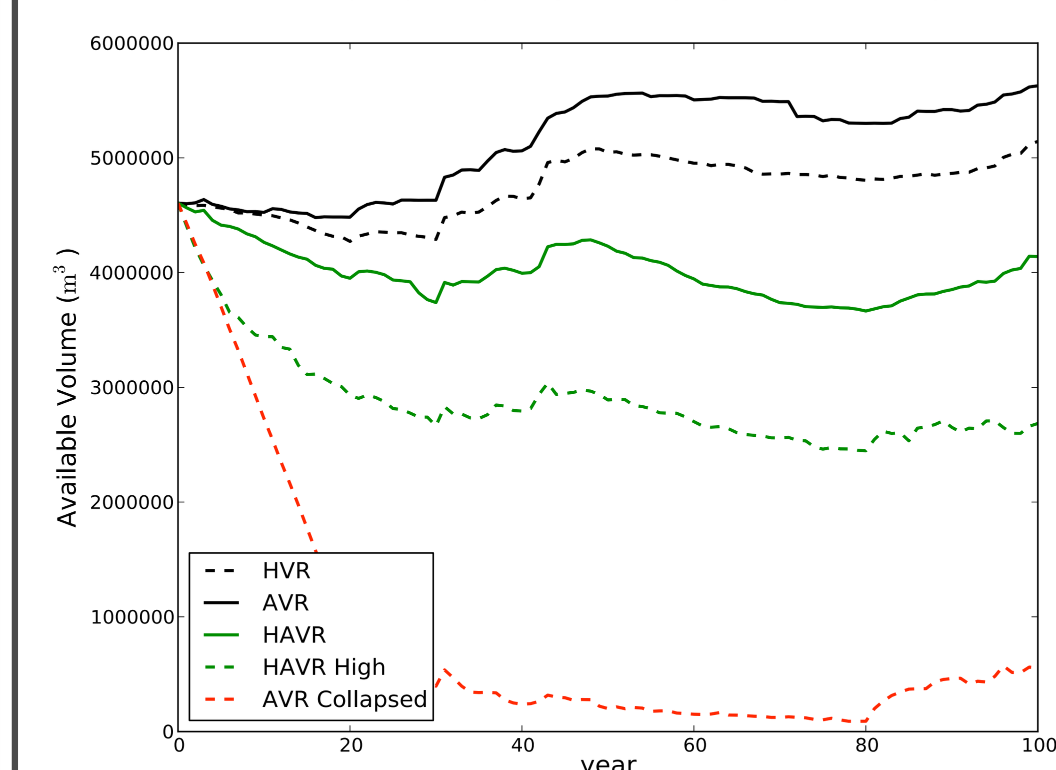
Experiments

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

Three Reward Models

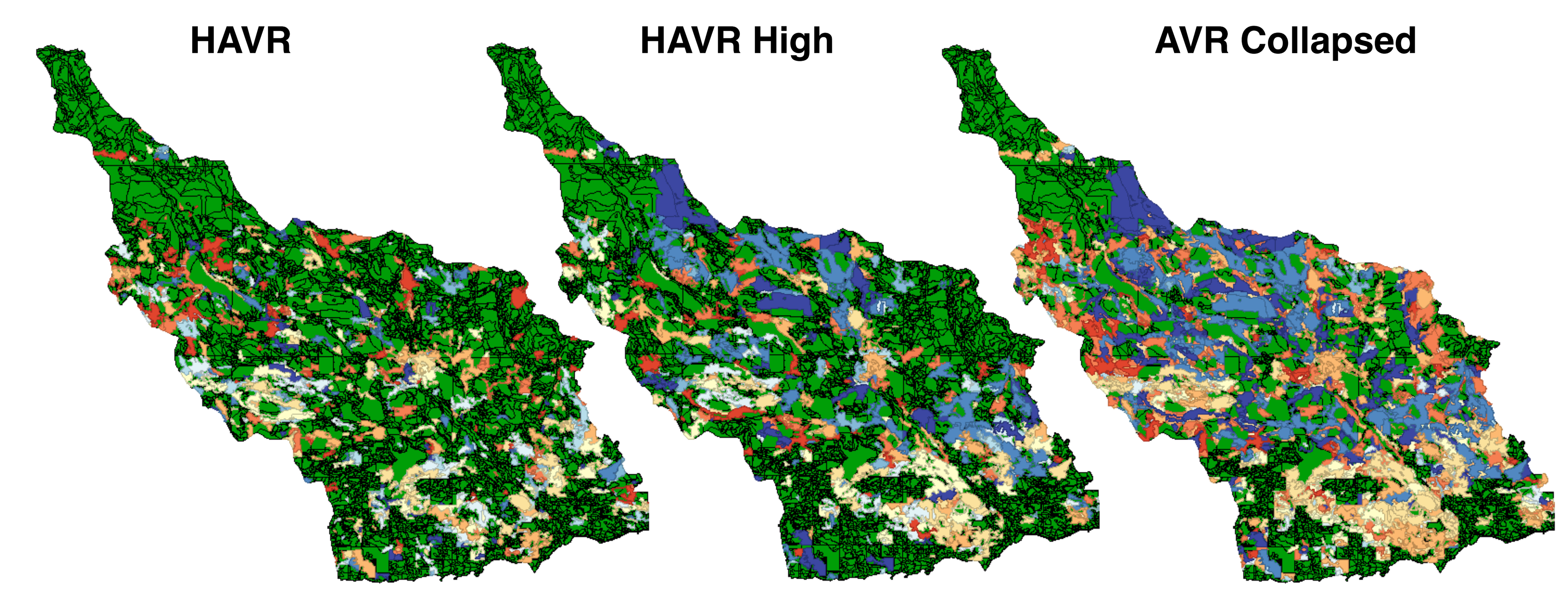
Provide positive reward for volume cut minus...
- **HVR =** penalty for irregular harvest volumes over time
- **AVR =** penalty for irregular available volume of the forest over time
- **HAVR =** HVR + AVR

Available Volume Over Time



Standard Deviation from mean Volume for Policy Under Each Value Model

	Harvest Volume	Available Volume
AVR	18,065	411,085
HVR	**14,422**	248,920
HAVR	20,309	**224,212**
HAVR High	50,059	417,278
AVR Collapsed	72,859	1,125,138



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.

Decade in which cell was harvested

0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100