Planning and Optimization

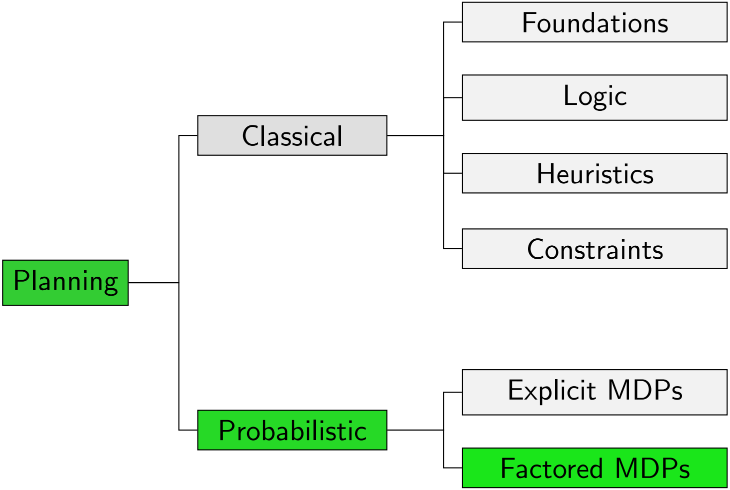
G5. Asymptotically Suboptimal Monte-Carlo Methods

Malte Helmert and Thomas Keller

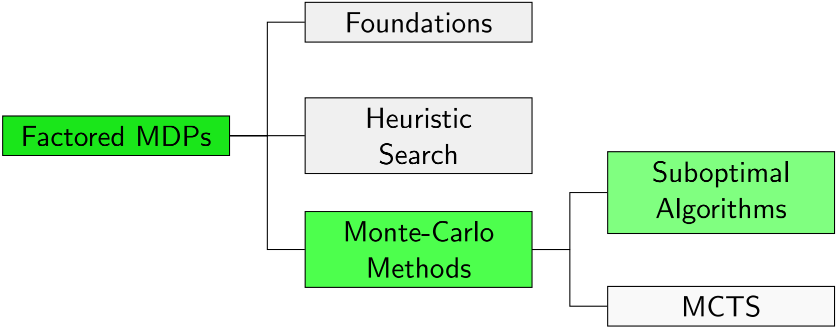
Universität Basel

December 11, 2019

Content of this Course



Content of this Course: Factored MDPs



Motivation

Monte-Carlo Methods: Brief History

- 1930s: first researchers experiment with Monte-Carlo methods
- 1998: Ginsberg's GIB player competes with Bridge experts
- 2002: Kearns et al. propose Sparse Sampling
- 2002: Auer et al. present UCB1 action selection for multi-armed bandits
- 2006: Coulom coins term Monte-Carlo Tree Search (MCTS)
- 2006: Kocsis and Szepesvári combine UCB1 and MCTS to the famous MCTS variant, UCT
- 2007–2016: Constant progress of MCTS in Go culminates in AlphaGo's historical defeat of dan 9 player Lee Sedol

Monte-Carlo Methods

Monte-Carlo Methods: Idea

- Summarize a broad family of algorithms
- Decisions are based on random samples (Monte-Carlo sampling)
- Results of samples are aggregated by computing the average (Monte-Carlo backups)
- Apart from that, algorithms can differ significantly

Careful: Many different definitions of MC methods in the literature

Types of Random Samples

Random samples only have in common that something is drawn from a given probability distribution. Some examples:

- a determinization is sampled (Hindsight Optimization)
- runs under a fixed policy are simulated (Policy Simulation)
- considered outcomes are sampled (Sparse Sampling)
- runs under an evolving policy are simulated (Monte-Carlo Tree Search)

Reminder: Bellman Backups

Algorithms like Value Iteration, (L)AO* or (L)RTDP use the Bellman equation as an update procedure.

The i-th state-value estimate of state s, $\hat{V}^i(s)$, is computed with Bellman backups as

$$\hat{V}^i(s) := \min_{\ell \in L(s)} \left(c(\ell) + \sum_{s' \in S} T(s, \ell, s') \cdot \hat{V}^{i-1}(s') \right).$$

(Some algorithms use a heuristic if the state-value estimate on the right hand side of the Bellman backup is undefined.)

Monte-Carlo Backups

Monte-Carlo methods estimate state-values by averaging over all samples instead.

Let $N^i(s)$ be the number of samples for state s in the i first algorithm iterations and let $cost^k(s)$ be the cost for s in the k-th sample ($cost^k(s) = 0$ if k-th sample has no estimate for s).

The i-th state-value estimate of state s, $\hat{V}^i(s)$, is computed with Monte-Carlo backups as

$$\hat{V}^i(s) := \frac{1}{N^i(s)} \cdot \sum_{k=1}^{i} cost^k(s).$$

Monte-Carlo Backups: Properties

- no need to store $cost^k(s)$ for $k = 1, \dots, i$:
  it is possible to compute Monte-Carlo backups iteratively as

$$\hat{V}^i(s) := \hat{V}^{i-1}(s) + \frac{1}{N^i(s)}(cost^i(s) - \hat{V}^{i-1}(s))$$

- no need to know SSP model for backups
- if s is a random variable, $\hat{V}^i(s)$ converges to $\mathbb{E}[s]$
  due to the strong law of large numbers
- if s is not a random variable, this is not always the case

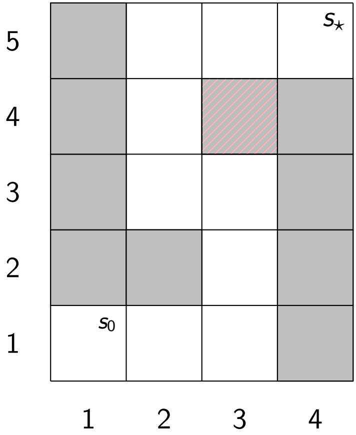
Hindsight Optimization

Hindsight Optimization: Idea

Repeat as long as resources (deliberation time, memory) allow:

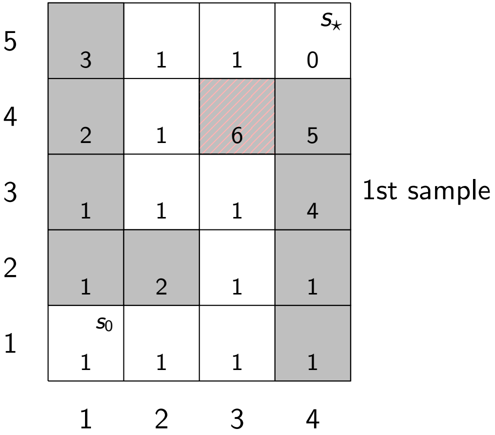
- Sample outcomes of all actions
  $\Rightarrow$ deterministic (classical) planning problem
- For each applicable action $\ell \in L(s_0)$,
  compute plan in the sample that starts with ℓ
- Execute the action with the lowest average plan cost

Hindsight Optimization: Example



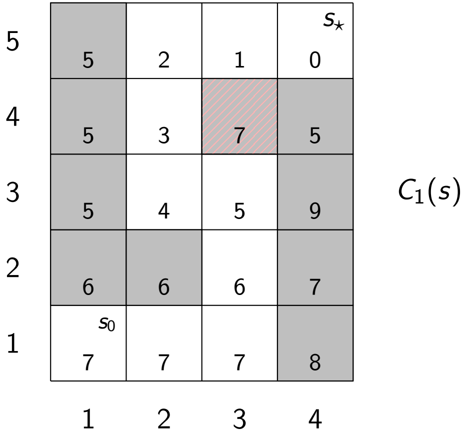
- cost of 1 for all actions except for moving away from (3,4) where cost is 3
- get stuck when moving away from gray cells with prob. 0.6

Hindsight Optimization: Example



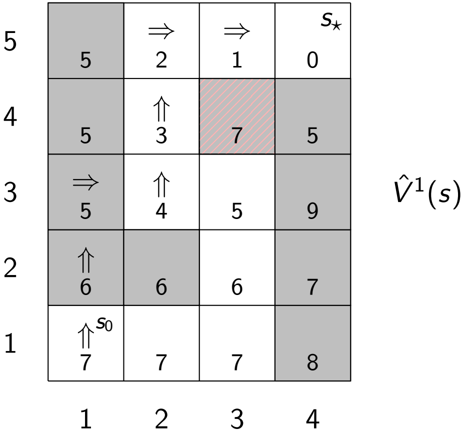
- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



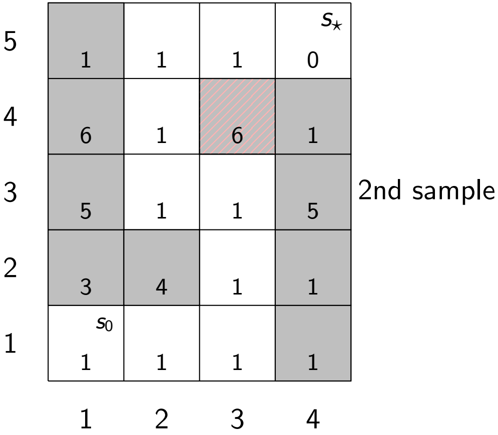
- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



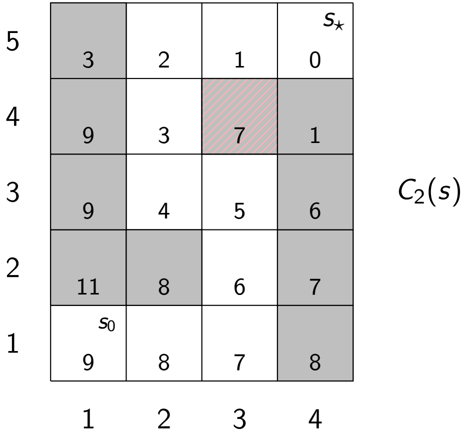
- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



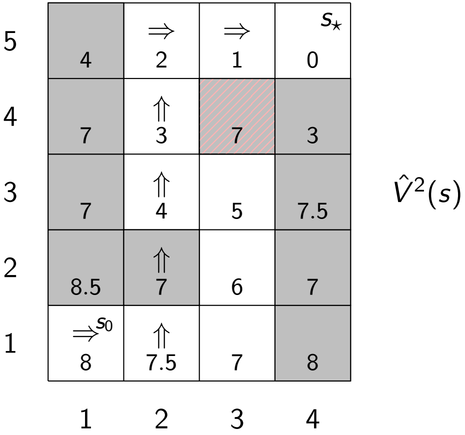
- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



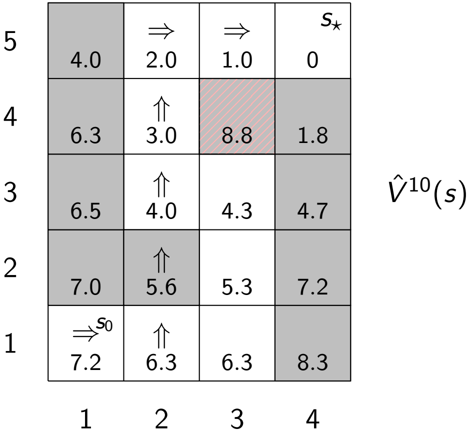
- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example



- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example

	1	2	3	4
5	4.55	$\Rightarrow$ 2.0	$\Rightarrow$ 1.0	$s_\star$ 0
4	5.43	$\Uparrow$ 3.0	8.50	2.40
3	6.57	$\Uparrow$ 4.0	$\Leftarrow$ 4.51	4.99
2	8.22	6.69	$\Uparrow$ 5.51	7.16
1	$\Rightarrow$ s_0 7.69	$\Rightarrow$ 6.89	$\Uparrow$ 6.51	8.48

$\hat{V}^{100}(s)$

- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

Hindsight Optimization: Example

5	4.58	$\Rightarrow$ 2.0	$\Rightarrow$ 1.0	$s_\star$ 0
4	5.56	$\Uparrow$ 3.0	8.33	2.44
3	6.54	$\Uparrow$ 4.0	4.49	4.84
2	7.88	$\Uparrow$ 6.48	5.49	6.80
1	$\Rightarrow$ s_0 7.60	$\Uparrow$ 6.75	6.49	8.44
	1	2	3	4

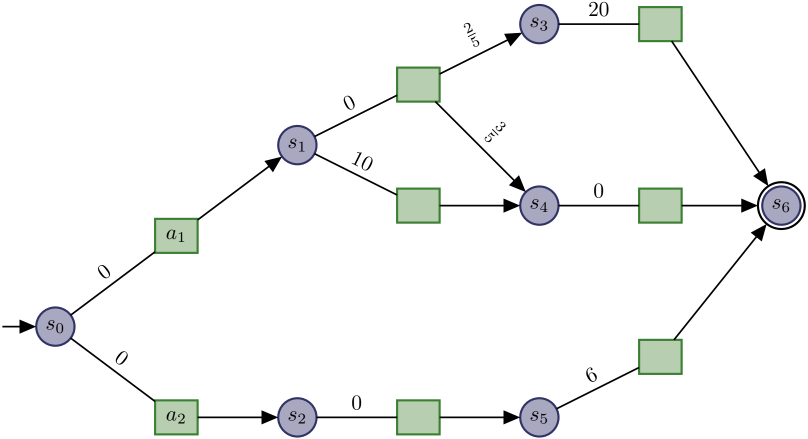
$\hat{V}^{1000}(s)$

- Samples can be described by number of times agent is stuck
- Multiplication with cost to move away from cell gives cost of leaving cell in sample

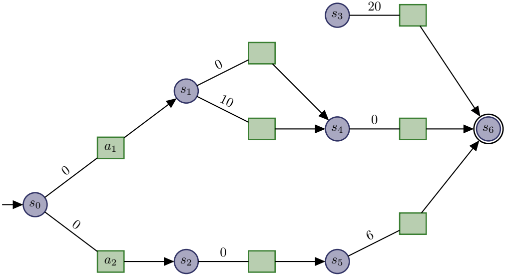
Hindsight Optimization: Evaluation

- HOP well-suited for some problems
- must be possible to solve sampled SSP efficiently:
  - domain-dependent knowledge (e.g., games like Bridge, Skat)
  - classical planner (FF-Hindsight, Yoon et. al, 2008)
- What about optimality in the limit?

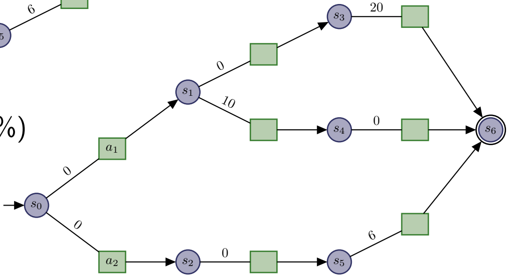
Hindsight Optimization: Optimality in the Limit



Hindsight Optimization: Optimality in the Limit

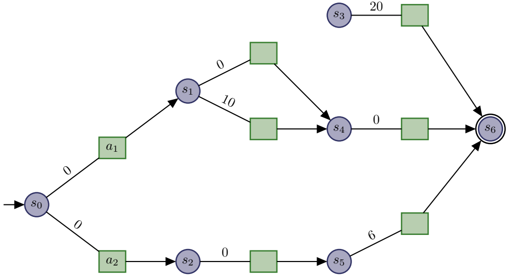


(sample probability: 60%)



(sample probability: 40%)

Hindsight Optimization: Optimality in the Limit

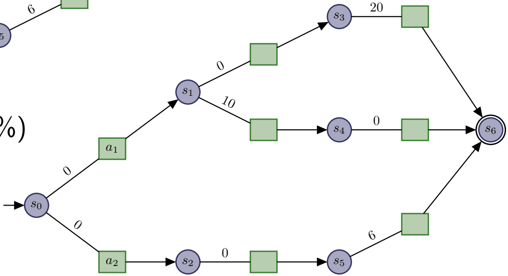


(sample probability: 60%)

with $k \to \infty$:

$\hat{Q}^k(s_0, a_1) \to 4$

$\hat{Q}^k(s_0, a_2) \to 6$



(sample probability: 40%)

Hindsight Optimization: Evaluation

- HOP well-suited for some problems
- must be possible to solve sampled MDP efficiently:
  - domain-dependent knowledge (e.g., games like Bridge, Skat)
  - classical planner (FF-Hindsight, Yoon et. al, 2008)
- What about optimality in the limit?
  $\Rightarrow$ in general not optimal due to assumption of clairvoyance

Policy Simulation

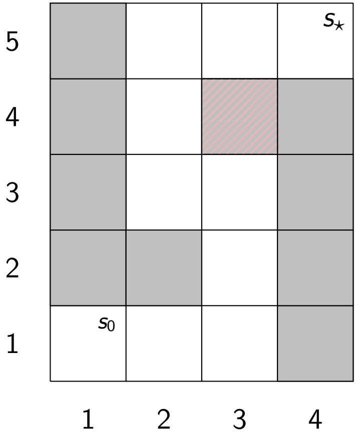
Policy Simulation: Idea

Repeat as long as resources (deliberation time, memory) allow:

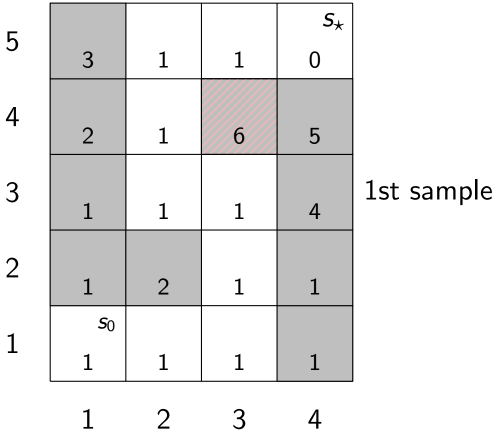
- For each applicable action $\ell \in L(s_0)$,
  start a run from s_0 with ℓ and then follow a given policy π
- Execute the action with the lowest average simulation cost

Avoids clairvoyance by evaluation of policy
through simulation of its execution.

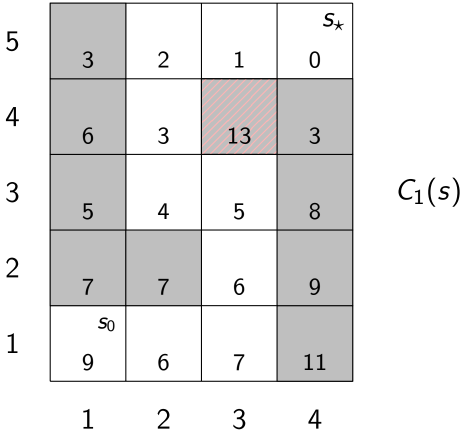
Policy Simulation: Example (following Optimistic Policy)



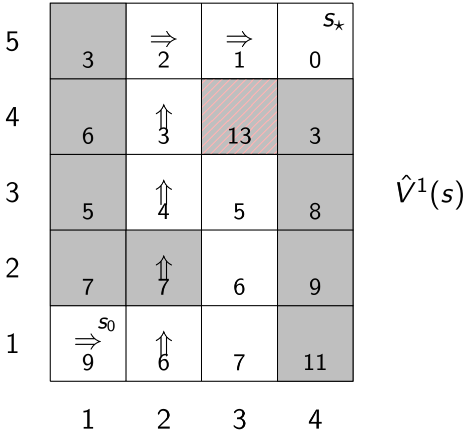
Policy Simulation: Example (following Optimistic Policy)



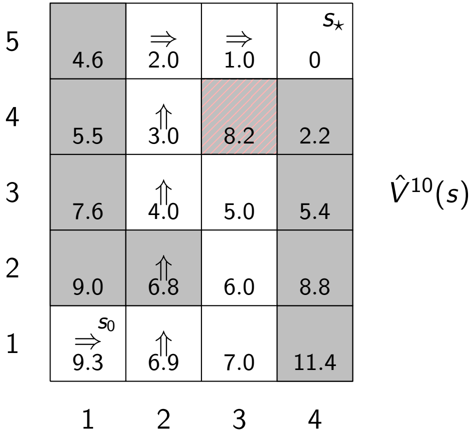
Policy Simulation: Example (following Optimistic Policy)



Policy Simulation: Example (following Optimistic Policy)



Policy Simulation: Example (following Optimistic Policy)



Policy Simulation: Example (following Optimistic Policy)

	1	2	3	4
5	4.55	$\Rightarrow$ 2.0	$\Rightarrow$ 1.0	$s_\star$ 0
4	5.54	$\Uparrow$ 3.0	8.42	2.37
3	6.52	$\Uparrow$ 4.0	5.0	5.13
2	9.2	$\Uparrow$ 6.69	6.0	8.43
1	s_0 $\Rightarrow$ 10.06	$\Uparrow$ 7.63	7.0	10.66

$\hat{V}^{100}(s)$

Policy Simulation: Example (following Optimistic Policy)

5	4.53	$\Rightarrow$ 2.0	$\Rightarrow$ 1.0	$s_\star$ 0
4	5.46	$\Uparrow$ 3.0	8.24	2.53
3	6.52	$\Uparrow$ 4.0	5.0	5.11
2	8.99	$\Uparrow$ 6.42	6.0	8.56
1	s_0 $\Rightarrow$ 10.11	$\Uparrow$ 7.78	7.0	11.09
	1	2	3	4

$\hat{V}^{1000}(s)$

Policy Simulation: Evaluation

- Base policy is static
- No mechansim to overcome weaknesses of base policy
  (if there are no weaknesses, we don't need policy simulation)
- Suboptimal decisions in simulation affect policy quality
- What about optimality in the limit?
  $\Rightarrow$ in general not optimal due to inability of policy to improve

Sparse Sampling

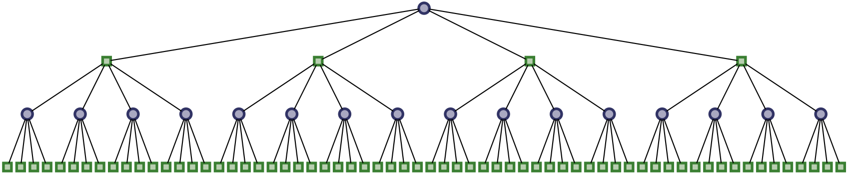
Sparse Sampling: Idea

Sparse Sampling (Kearns et al., 2002) approaches problem that number of reachable states under a policy can be too large

- Creates search tree up to a given lookahead horizon
- A constant number of outcomes is sampled for each state-action pair
- Outcomes that were not sampled are ignored
- Near-optimal: expected cost of resulting policy close to expected cost of optimal policy
- Runtime independent from the number of states

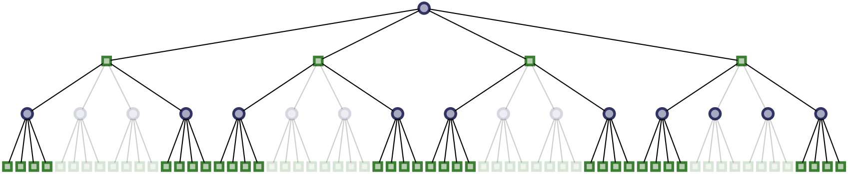
Sparse Sampling: Search Tree

Without Sparse Sampling



Sparse Sampling: Search Tree

With Sparse Sampling



Sparse Sampling: Problems

- Independent from number of states, but still exponential in lookahead horizon
- Constants that give number of outcomes and lookahead horizon large for good bounds on near-optimality
- Search time difficult to predict
- Search tree is symmetric
  $\Rightarrow$ resources are wasted in non-promising parts of the tree

Summary

Summary

- Monte-Carlo methods have a long history but no successful applications until 1990s
- Monte-Carlo methods use sampling and backups that average over sample results
- Hindsight optimization averages over plan cost in sampled determinization
- Policy simulation simulates the exection of a policy
- Sparse sampling considers only a fixed amount of outcomes
- All three methods are not optimal in the limit