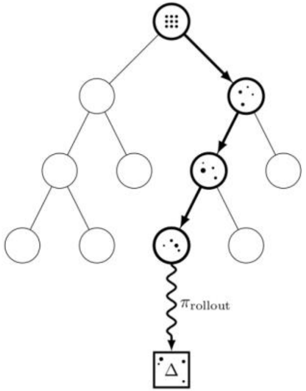
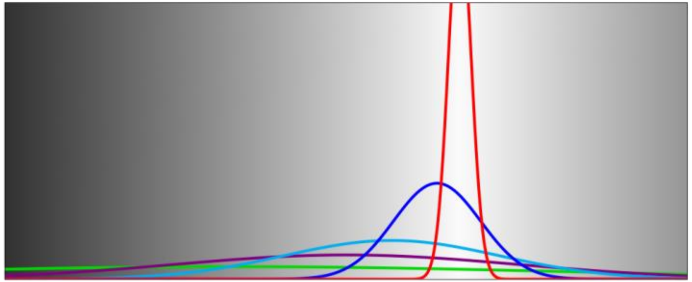
Information Particle Filter Tree: An Online Algorithm for POMDPs with Belief-Based Rewards on Continuous Domains

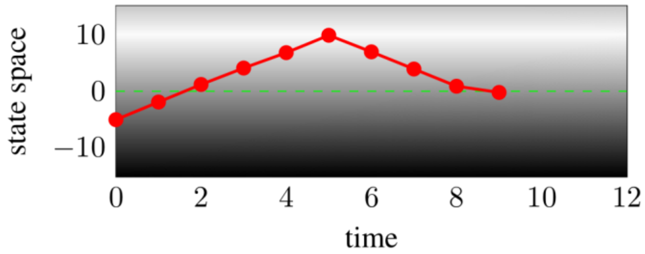
Johannes Fischer * and Ömer Sahin Tas *

*Equal contribution

International Conference on Machine Learning 2020







POMDPs

- Model decision problems under uncertainty

POMDPs

- Model decision problems under uncertainty
- Cover uncertainties in
  - Models
  - Environment
  - Future behavior of others

POMDPs

- Model decision problems under uncertainty
- Cover uncertainties in
  - Models
  - Environment
  - Future behavior of others

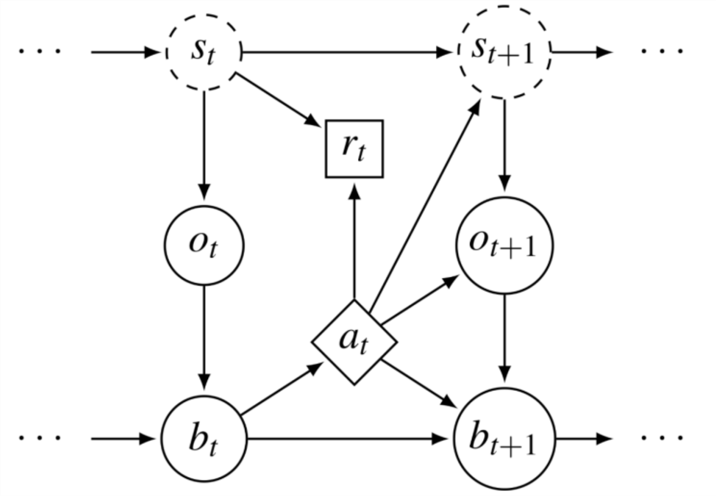


Figure: Probabilistic graphical model of a POMDP.

POMDPs

- Model decision problems under uncertainty
- Cover uncertainties in
  - Models
  - Environment
  - Future behavior of others
- Reasoning in high dimensional belief space

  →*Difficult to solve!*

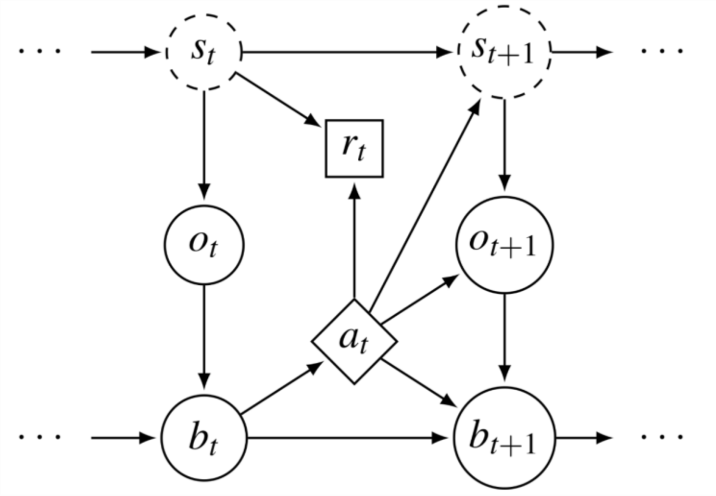


Figure: Probabilistic graphical model of a POMDP.

POMDPs

- Model decision problems under uncertainty
- Cover uncertainties in
  - Models
  - Environment
  - Future behavior of others
- Reasoning in high dimensional belief space

  →*Difficult to solve!*

Can POMDP solvers be improved by considering information?

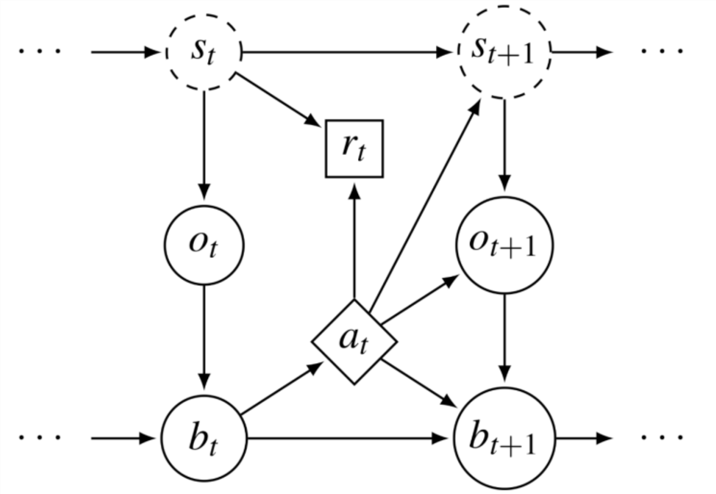


Figure: Probabilistic graphical model of a POMDP.

Information Measures

- Optimal value function V^* and information measures have similar shape
  →"more information = higher value"

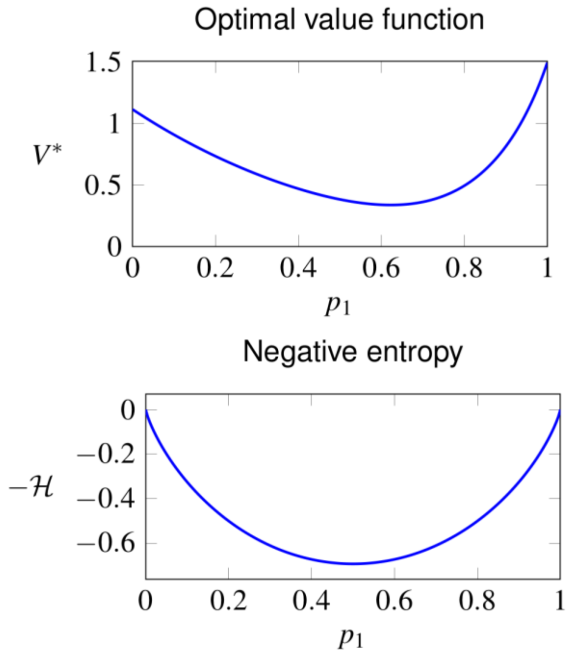


Figure: Shape of optimal value function and negative entropy.

Information Measures

- Optimal value function V^* and information measures have similar shape
  → "more information = higher value"

- Motivation
  - Speed up planning
  - Allow active information gathering

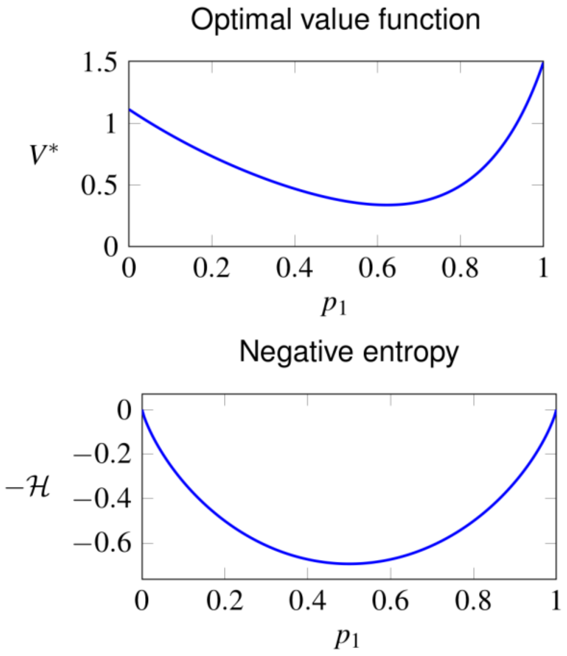


Figure: Shape of optimal value function and negative entropy.

POMDPs

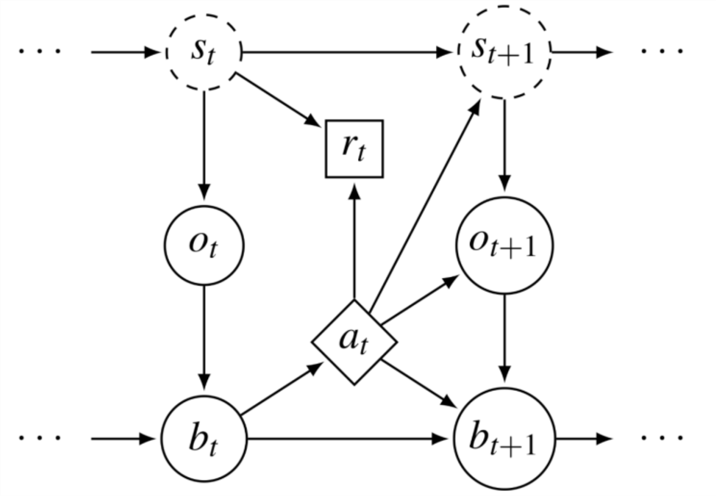


Figure: Probabilistic graphical model of a POMDP.

ρPOMDPs

- Extension of POMDP framework
- Belief-dependent reward model $\rho(b, a)$

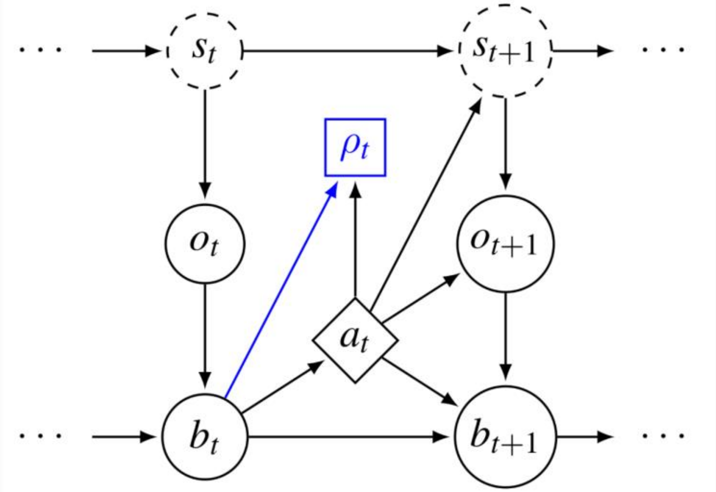


Figure: Probabilistic graphical model of a ρPOMDP.

[1] Araya-López et al., “A POMDP Extension with Belief-dependent Rewards,” (2010)

ρPOMDPs

- Extension of POMDP framework
- Belief-dependent reward model $\rho(b, a)$
- Solvers exist only for
  - Discrete problems
  - Piecewise linear and convex ρ
  - Offline computation

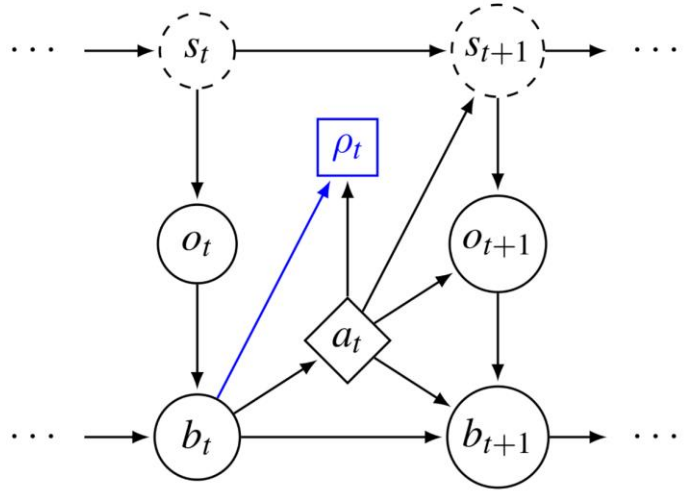


Figure: Probabilistic graphical model of a ρPOMDP.

[1] Araya-López et al., “A POMDP Extension with Belief-dependent Rewards,” (2010)

ρPOMDPs

- Extension of POMDP framework
- Belief-dependent reward model $\rho(b, a)$
- Solvers exist only for
  - Discrete problems
  - Piecewise linear and convex ρ
  - Offline computation

How can ρPOMDPs on continuous domains be solved online?

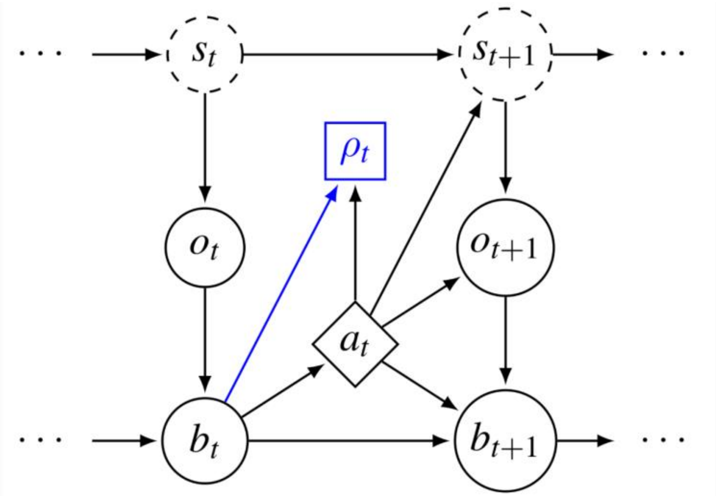


Figure: Probabilistic graphical model of a ρPOMDP.

[1] Araya-López et al., "A POMDP Extension with Belief-dependent Rewards," (2010)

Approach - Information Particle Filter Tree

- Adapt MCTS-based POMDP solver
- Approximate belief by particles
- Evaluate ρ on particle sets

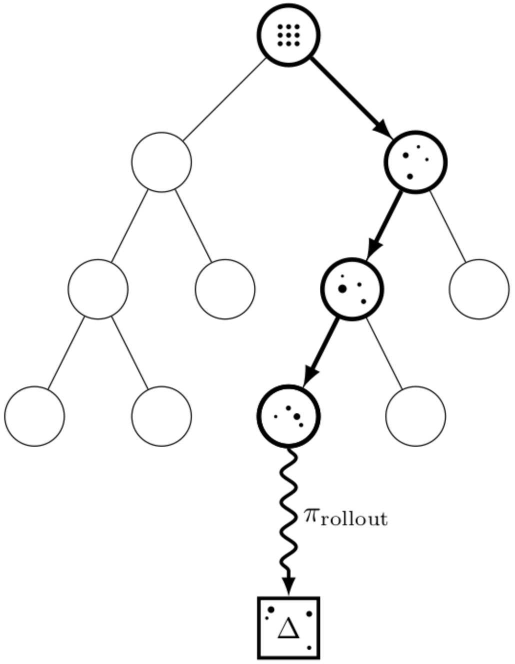


Figure: Simulation phase of IPFT.

Approach - Information Particle Filter Tree

- Adapt MCTS-based POMDP solver
- Approximate belief by particles
- Evaluate ρ on particle sets

→Online anytime algorithm
→Continuous problems

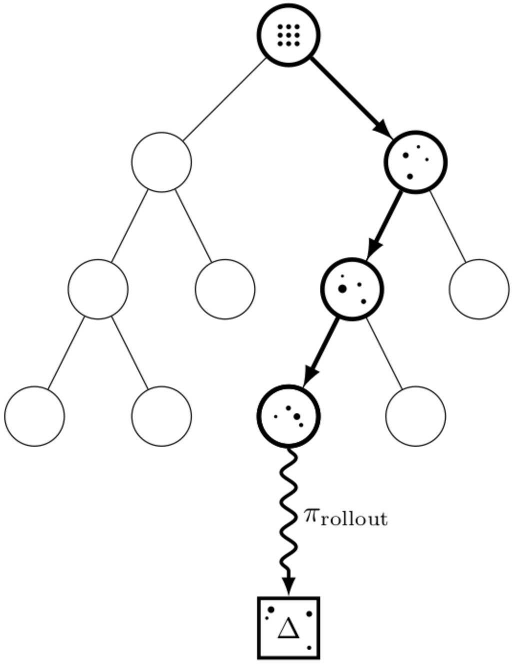


Figure: Simulation phase of IPFT.

Potential-Based Reward Shaping

Karlsruhe Institute of Technology

- Reward shaping changes the optimal policy

$$\tilde{R}(b_t, a_t) = R(b_t, a_t) + F(b_t, a_t, b_{t+1})$$

Potential-Based Reward Shaping

- Reward shaping changes the optimal policy

$$\tilde{R}(b_t, a_t) = R(b_t, a_t) + F(b_t, a_t, b_{t+1})$$

- BUT: Optimal policy is invariant under potential-based reward shaping for infinite horizon [2]

$$F(b_t, a_t, b_{t+1}) = \gamma\phi(b_{t+1}) - \phi(b_t)$$

[2] Eck et. al. “Potential-based reward shaping for finite horizon online POMDP planning.” (2016)

Potential-Based Reward Shaping

- Reward shaping changes the optimal policy

$$\tilde{R}(b_t, a_t) = R(b_t, a_t) + F(b_t, a_t, b_{t+1})$$

- BUT: Optimal policy is invariant under potential-based reward shaping for infinite horizon [2]

$$F(b_t, a_t, b_{t+1}) = \gamma\phi(b_{t+1}) - \phi(b_t)$$

- V^* serves as a particularly effective potential

[2] Eck et. al. “Potential-based reward shaping for finite horizon online POMDP planning.” (2016)

Information-Theoretic Reward Shaping

- Information measures have similar shape to V^*
  - Convex on belief space
  - →Use as heuristic for V^*

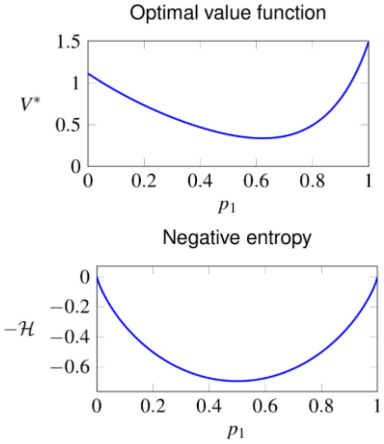


Figure: Shape of optimal value function and negative entropy.

Information-Theoretic Reward Shaping

- Information measures have similar shape to V^*
  - Convex on belief space
  - →Use as heuristic for V^*

- Two potential-based shaping functions
  - Discounted information gain $\Delta\mathcal{I}_\gamma(b, b') = \gamma\mathcal{I}(b') - \mathcal{I}(b)$
  - Undiscounted information gain $\Delta\mathcal{I}_1(b, b') = \mathcal{I}(b') - \mathcal{I}(b)$

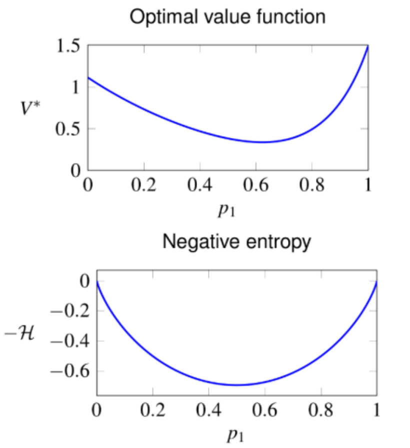


Figure: Shape of optimal value function and negative entropy.

Information-Theoretic Reward Shaping

- Information measures have similar shape to V^*
  - Convex on belief space
  - →Use as heuristic for V^*

- Two potential-based shaping functions
  - Discounted information gain $\Delta\mathcal{I}_\gamma(b, b') = \gamma\mathcal{I}(b') - \mathcal{I}(b)$
  - Undiscounted information gain $\Delta\mathcal{I}_1(b, b') = \mathcal{I}(b') - \mathcal{I}(b)$

$$\rho(b, a, b') = \int_S R(s, a)b(s)ds + \lambda\Delta\mathcal{I}(b, b')$$

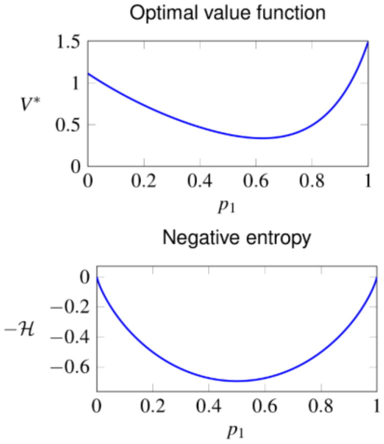


Figure: Shape of optimal value function and negative entropy.

Solving ρPOMDPs in Continuous Domains

- Based on Particle Filter Tree (PFT) Algorithm [3]
  - MCTS → continuous states
  - Double Progressive Widening (DPW)
    → continuous actions & observations

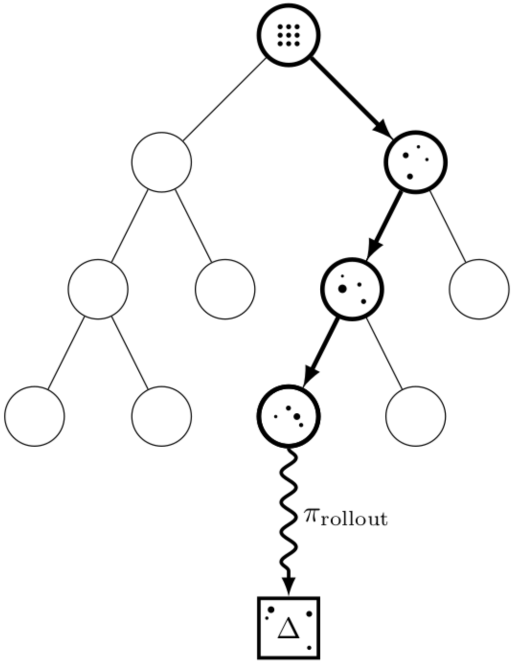


Figure: Simulation phase of PFT.

[3] Sunberg and Kochenderfer, "Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces," (2018)

Solving ρPOMDPs in Continuous Domains

- Based on Particle Filter Tree (PFT) Algorithm [3]
  - MCTS → continuous states
  - Double Progressive Widening (DPW)
    → continuous actions & observations
  - Solves belief MDP
  - Small weighted particle sets $(m = 20)$
  - Update with mean particle return

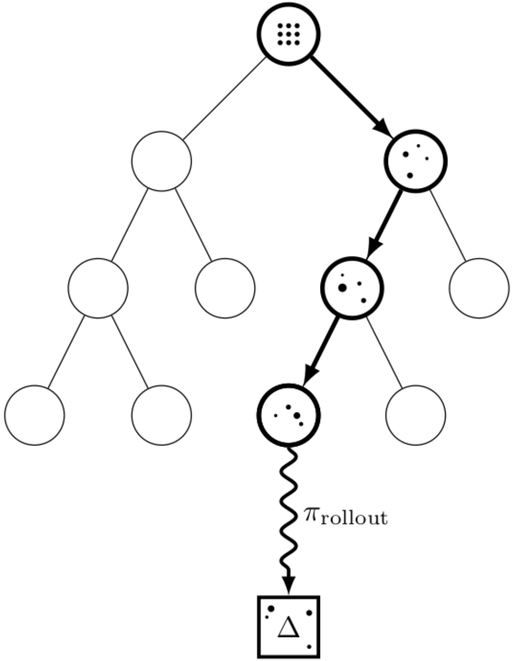


Figure: Simulation phase of PFT.

[3] Sunberg and Kochenderfer, "Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces," (2018)

Solving ρPOMDPs in Continuous Domains - Information Particle Filter Tree (IPFT)

- Particle set approximates belief

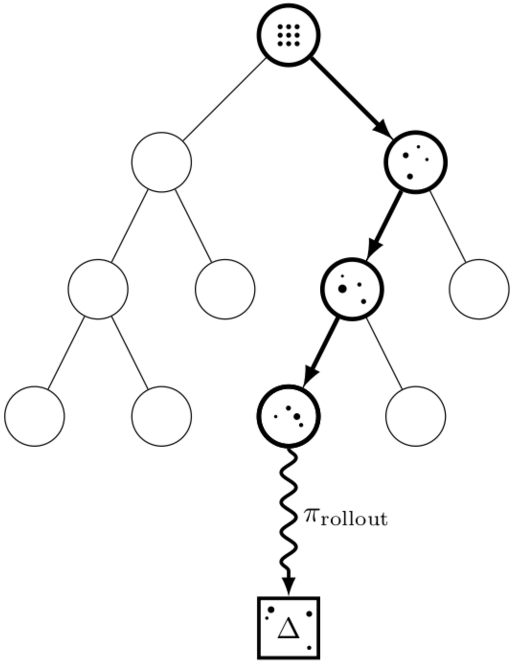


Figure: Simulation phase of IPFT.

Solving ρPOMDPs in Continuous Domains - Information Particle Filter Tree (IPFT)

- Particle set approximates belief
- Evaluate ρ on weighted particle sets, e.g.
  - $-\mathcal{H}(b) = \int_S b(s) \log b(s) \, \mathrm{d}s \approx \sum_i w_i \log b(s_i)$

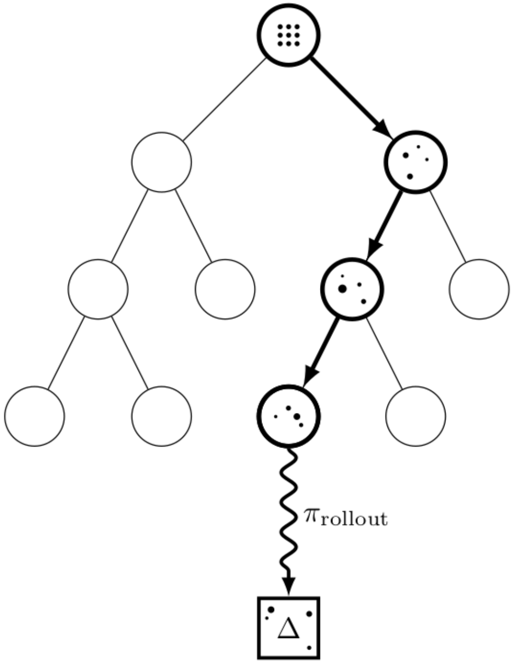


Figure: Simulation phase of IPFT.

Solving ρPOMDPs in Continuous Domains - Information Particle Filter Tree (IPFT)

- Particle set approximates belief
- Evaluate ρ on weighted particle sets, e.g.
  - $-\mathcal{H}(b) = \int_S b(s) \log b(s) \, \mathrm{d}s \approx \sum_i w_i \log \hat{b}(s_i)$
  - Particle-based kernel density estimate $\hat{b}$

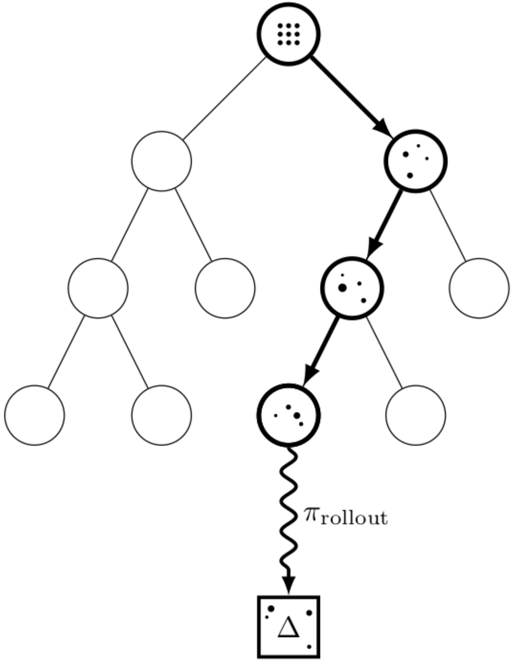


Figure: Simulation phase of IPFT.

Solving ρPOMDPs in Continuous Domains - Information Particle Filter Tree (IPFT)

- Particle set approximates belief
- Evaluate ρ on weighted particle sets, e.g.
  - $-\mathcal{H}(b) = \int_S b(s) \log b(s) \, \mathrm{d}s \approx \sum_i w_i \log \hat{b}(s_i)$
  - Particle-based kernel density estimate $\hat{b}$
- Averaging over many particle sets leads to better entropy estimate

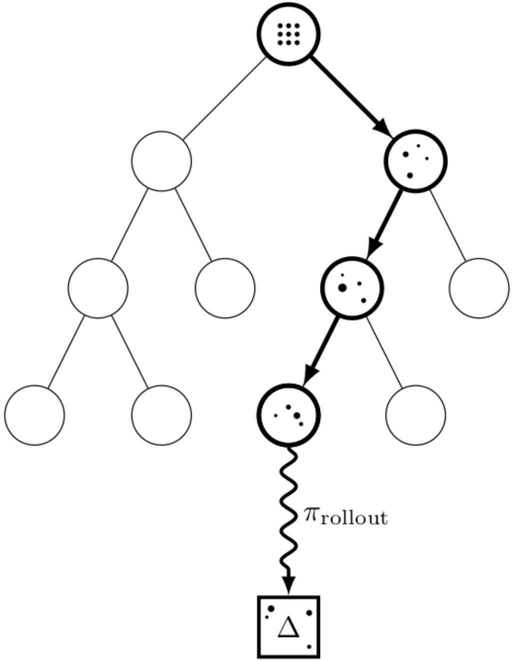


Figure: Simulation phase of IPFT.

Solving ρPOMDPs in Continuous Domains - Information Particle Filter Tree (IPFT)

- Particle set approximates belief
- Evaluate ρ on weighted particle sets, e.g.
  - $-\mathcal{H}(b) = \int_S b(s) \log b(s)\,\mathrm{d}s \approx \sum_i w_i \log \hat{b}(s_i)$
  - Particle-based kernel density estimate $\hat{b}$
- Averaging over many particle sets leads to better entropy estimate

→IPFT can solve arbitrary ρPOMDPs on continuous domains

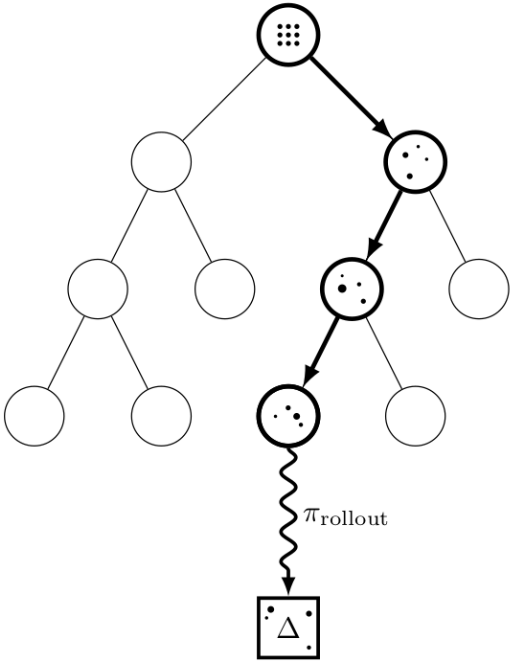
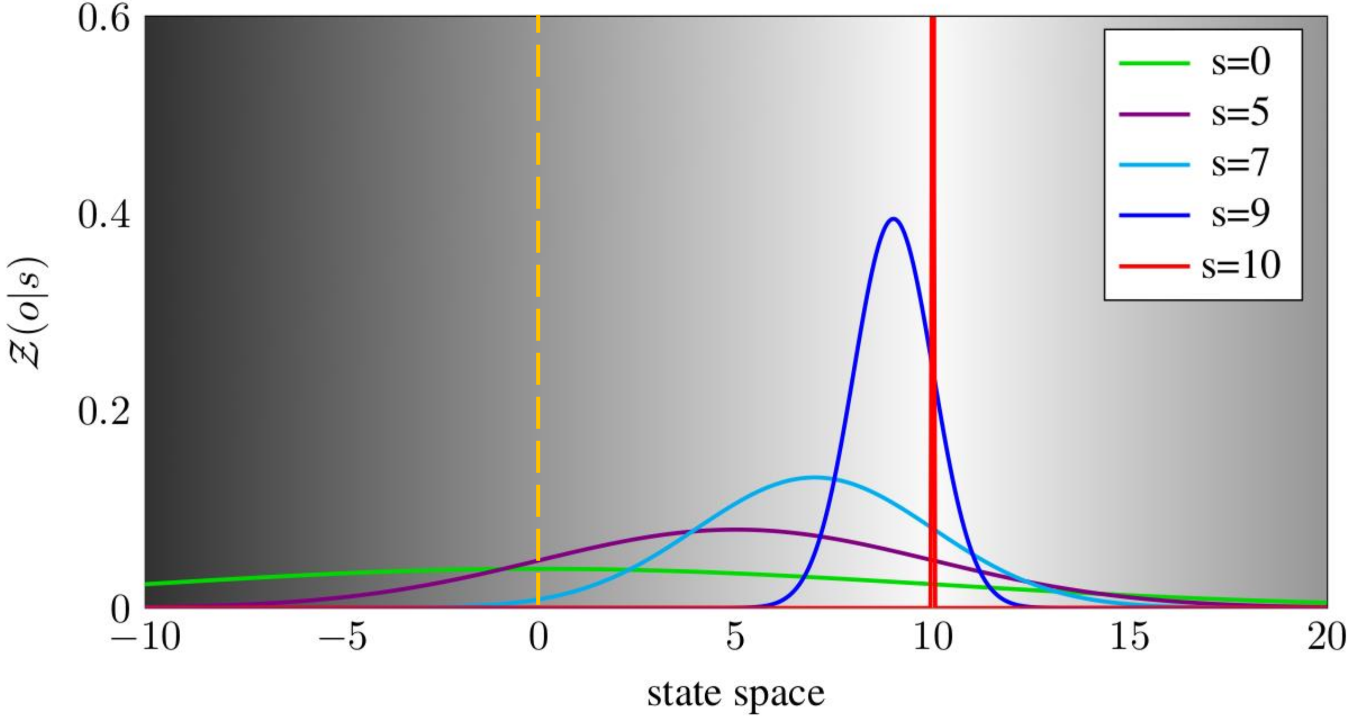


Figure: Simulation phase of IPFT.

Experiments – Light Dark



Experiments – Light Dark

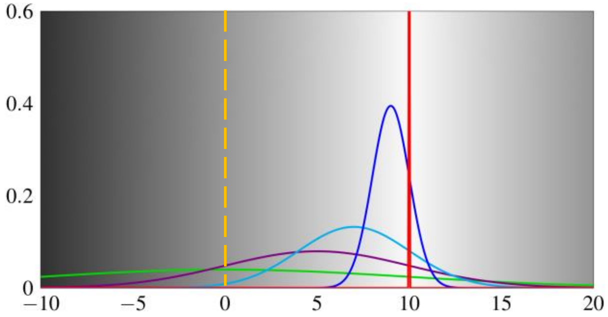


Figure: Light Dark environment.

- Goal: execute $a = 0$ at $s = 0$
- Consider action spaces

$$\mathbb{A}_{10} = \{-10, -1, 0, 1, 10\}$$
$$\mathbb{A}_{3} = \{-3, -1, 0, 1, 3\}$$

Experiments – Light Dark

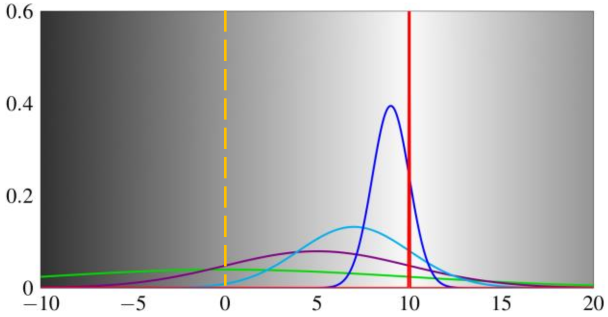


Figure: Light Dark environment.

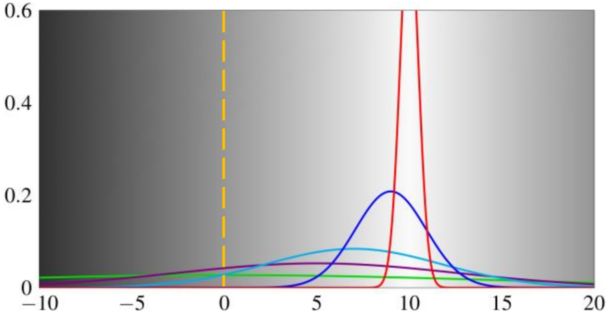


Figure: Continuous Light Dark environment.

- Goal: execute $a = 0$ at $s = 0$
- Consider action spaces

$$\mathbb{A}_{10} = \{-10, -1, 0, 1, 10\}$$

$$\mathbb{A}_{3} = \{-3, -1, 0, 1, 3\}$$

- Continuous state space
- Transition noise
- Increased observation noise

Results – Light Dark

Karlsruhe Institute of Technology

Algorithm	Light Dark problem: action space $\mathbb{A}_{10}$		Light Dark problem: action space $\mathbb{A}_3$	
IPFT($\Delta\mathcal{I}_1$)	58.2 ± 0.4		34.8 ± 0.7	
IPFT($\Delta\mathcal{I}_\gamma$)	55.4 ± 0.5		27.8 ± 0.8	
POMCPOW	58.6 ± 0.5		-2.6 ± 0.9	
PFT-DPW	57.4 ± 0.5		33.9 ± 0.8	

Table: Mean reward and standard deviation of 1000 simulations.

Results – Light Dark

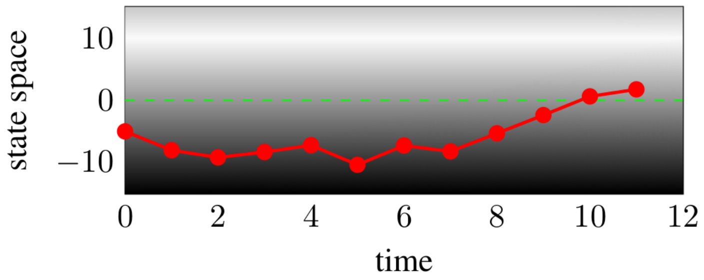
Algorithm	Light Dark problem: action space $\mathbb{A}_{10}$	Light Dark problem: action space $\mathbb{A}_3$	Continuous Light Dark problem: action space $\mathbb{A}_{10}$	Continuous Light Dark problem: action space $\mathbb{A}_3$
IPFT($\Delta\mathcal{I}_1$)	58.2 ± 0.4	34.8 ± 0.7	35.7 ± 1.8	35.9 ± 1.0
IPFT($\Delta\mathcal{I}_\gamma$)	55.4 ± 0.5	27.8 ± 0.8	38.4 ± 1.7	32.3 ± 1.4
POMCPOW	58.6 ± 0.5	-2.6 ± 0.9	-8.5 ± 2.3	-2.9 ± 2.1
PFT-DPW	57.4 ± 0.5	33.9 ± 0.8	-33.1 ± 2.4	-19.6 ± 2.3

Table: Mean reward and standard deviation of 1000 simulations.

Results – Light Dark

Algorithm	Light Dark problem: action space $\mathbb{A}_{10}$		Light Dark problem: action space $\mathbb{A}_3$		Continuous Light Dark problem: action space $\mathbb{A}_{10}$		Continuous Light Dark problem: action space $\mathbb{A}_3$	
IPFT($\Delta\mathcal{I}_1$)	58.2 ± 0.4		34.8 ± 0.7		35.7 ± 1.8		35.9 ± 1.0	
IPFT($\Delta\mathcal{I}_\gamma$)	55.4 ± 0.5		27.8 ± 0.8		38.4 ± 1.7		32.3 ± 1.4	
POMCPOW	58.6 ± 0.5		-2.6 ± 0.9		-8.5 ± 2.3		-2.9 ± 2.1	
PFT-DPW	57.4 ± 0.5		33.9 ± 0.8		-33.1 ± 2.4		-19.6 ± 2.3	

Table: Mean reward and standard deviation of 1000 simulations.



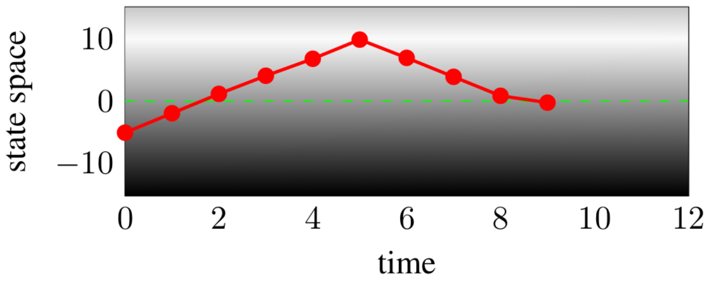


Figure: Exemplary trajectories of POMCPOW (left) and IPFT (right) in Continuous Light Dark problem.

Laser Tag

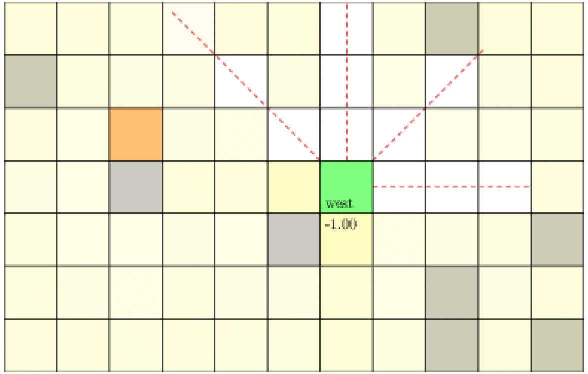


Figure: Laser Tag problem.

Laser Tag

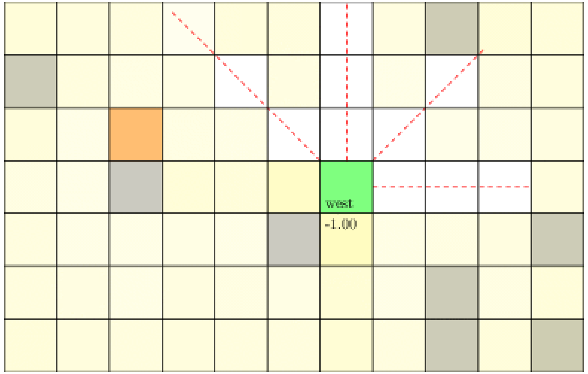


Figure: Laser Tag problem.

	Laser Tag	
IPFT($\Delta\mathcal{I}_1$)	-9.0 ± 0.2	
IPFT($\Delta\mathcal{I}_\gamma$)	-8.9 ± 0.2	
POMCPOW	-9.9 ± 0.2	
PFT-DPW	-12.0 ± 0.2	

Table: Mean reward and standard deviation of 1000 simulations.

Hyperparameter Sensitivity Analysis

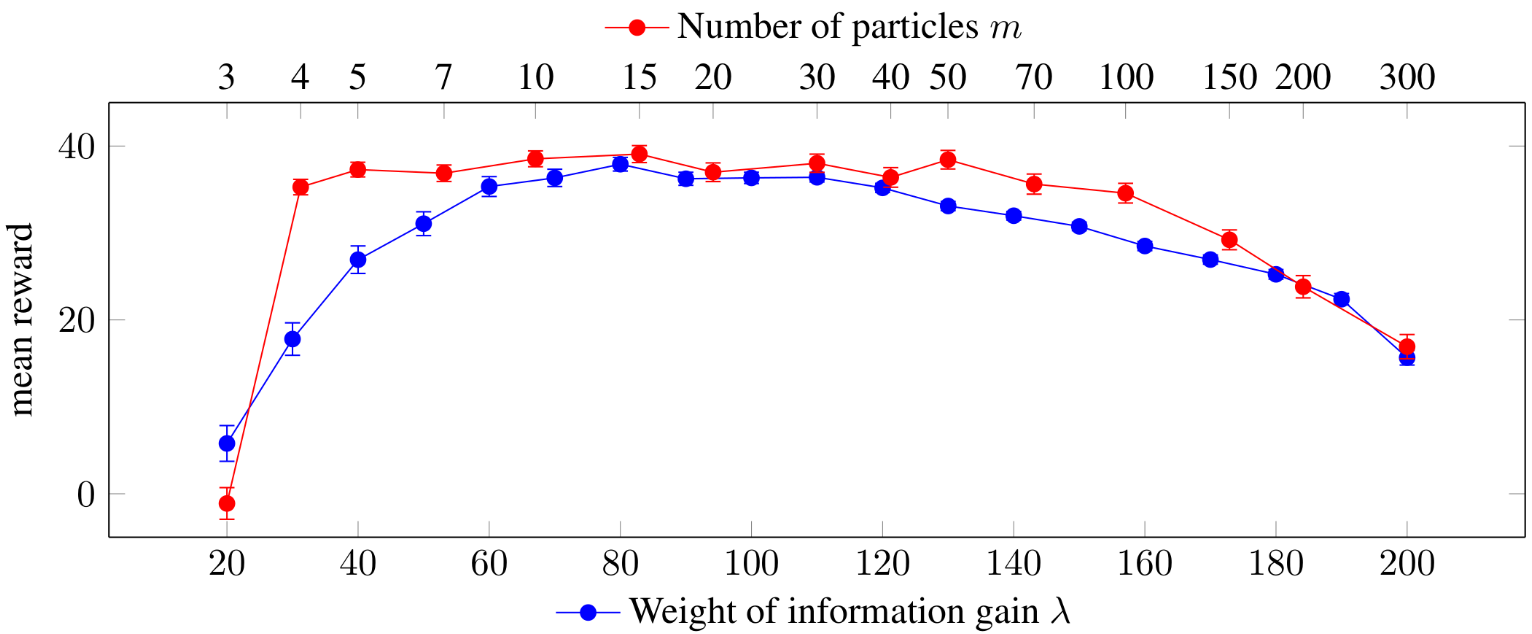


Figure: Mean reward and standard deviation of 1000 simulations of the Continuous Light Dark problem for different parameters.

Conclusion

Can POMDP solvers be improved by considering information?

Conclusion

Can POMDP solvers be improved by considering information?

- Information-theoretic reward shaping

→Helps by guiding agent to informative beliefs

Conclusion

Can POMDP solvers be improved by considering information?

- Information-theoretic reward shaping

→Helps by guiding agent to informative beliefs

How can ρPOMDPs on continuous domains be solved online?

Conclusion

Can POMDP solvers be improved by considering information?

- Information-theoretic reward shaping

→Helps by guiding agent to informative beliefs

How can ρPOMDPs on continuous domains be solved online?

- IPFT combines PFT algorithm with ρPOMDPs

→General online solver for continuous ρPOMDPs

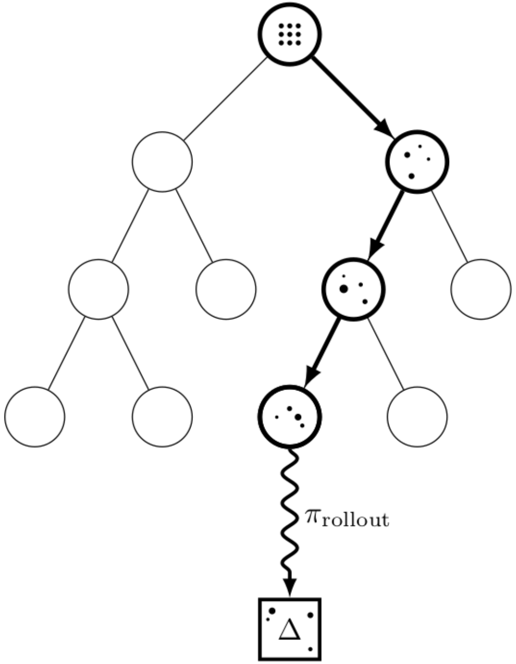


Figure: Simulation phase of IPFT.