Safe Exploration with Bayesian Reinforcement Learning

Eric Langlois

CSC 2125

April 29, 2019

Safe Reinforcement Learning

Reinforcement learning with safe behaviour

Safe Reinforcement Learning

Reinforcement learning with safe behaviour

Safe Policy Behave safely given known environment dynamics

- RL is OK; can encode safety in the reward.

Safe Exploration Behave safely while learning about the environment

- RL is not OK; ϵ-greedy takes random actions.

Safe RL Background

Many different definitions of safety

Optimization Objective

- Explicit safety constraints on states
- Worst-case reward
- Risk-averse reward
- Preserve ergodicity

Modified exploration

- Follow demonstrations
- Avoid risk

Safe RL

Limitations

- Can be hard to specify explicit safety constraints
- Structural constraints may be inappropriate
- How to choose risk aversion?
- Many approaches do not apply to exploration.

Objective: Reward-based safety

Define safety as maximizing expected total reward

- Causing the experiment to halt is automatically penalized via opportunity cost.
- Moves part of safety specification into dynamics model.
  - Actions are unsafe if they may produce negative long-term reward.

Objective: Reward-based safety

Define safety as maximizing expected total reward

- Causing the experiment to halt is automatically penalized via opportunity cost.
- Moves part of safety specification into dynamics model.
  - Actions are unsafe if they may produce negative long-term reward.

Hypothesis

It is easier to specify dynamics than safety constraints.

- Dynamics are objective properties of the environment

Objective: Reward-based safety

Define safety as maximizing expected total reward

- Causing the experiment to halt is automatically penalized via opportunity cost.
- Moves part of safety specification into dynamics model.
  - Actions are unsafe if they may produce negative long-term reward.

Hypothesis

It is easier to specify dynamics than safety constraints.

- Dynamics are objective properties of the environment

Selfish Safety?

- Only encourages safe behaviour that affects reward

Objective: Reward-based safety

Define safety as maximizing expected total reward

- Causing the experiment to halt is automatically penalized via opportunity cost.
- Moves part of safety specification into dynamics model.
  - Actions are unsafe if they may produce negative long-term reward.

Hypothesis

It is easier to specify dynamics than safety constraints.

- Dynamics are objective properties of the environment

Selfish Safety?

- Only encourages safe behaviour that affects reward
- Advantage? Terminate episode if human experimenter unhappy
  - Implied safety constraint: satisfy experimenter's notion of safety

Markov Decision Processes

Definition

A tuple (S, A, P, R)

$$\begin{aligned} S &: \text{State space} \\ A &: \text{Action space} \\ P &: \text{Transition probability matrix } P(s_{t+1}|s_t, a) \\ R &: \text{Reward function } R(s, a) \end{aligned}$$

In model-free RL,
P and R are unknown and must be discovered through exploration.

Uncertainty in Reinforcement Learning

No Uncertainty : Frequentist estimation of transition matrix

- ϵ-greedy exploration

Static Uncertainty : Model uncertainty but not updates

- Explore high-uncertainty states; stuck watching noise

Bayesian Uncertainty : Model uncertainty and updates to uncertainty.

Bayesian Reinforcement Learning

Objective

Maximize expected reward over a prior distribution of transitions P.

- Policy can depend on full observation history, incorporates learning.
- Typically motivated by explore / exploit trade-off
- An optimal policy will take deliberate exploration to learn about the environment as effectively as possible.

Objective: Safe Exploration

Want the agent to behave safely during exploration.
Should be possible with Bayesian RL:

- Considers consequences of exploratory actions
- Attempt to infer potential (reward) harm based on prior over models
- Avoid actions that might cause long-term low reward
- Explore where there is high potentially payoff

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

- Transform problem into MDP

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

- Transform problem into MDP
- Include agent's belief as part of the state
  - Where am I?
  - How fast am I moving?

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

- Transform problem into MDP
- Include agent's belief as part of the state
  - Where am I?
  - How fast am I moving?
  - *What do I know about the world?*

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

- Transform problem into MDP
- Include agent's belief as part of the state
  - Where am I?
  - How fast am I moving?
  - *What do I know about the world?*
- Transitions use Bayes Rule: update beliefs given observations
- Can in theory solve with regular RL

Bayes-Adaptive MDP

How does the agent reason about its own learning process?

- Transform problem into MDP
- Include agent's belief as part of the state
  - Where am I?
  - How fast am I moving?
  - *What do I know about the world?*
- Transitions use Bayes Rule: update beliefs given observations
- Can in theory solve with regular RL

Problems

- State space is massive, often impractical
- Bayes rule updates must be estimated, costly

BAMCP

Bayes-Adaptive Monte-Carlo Planning

- Relatively efficient algorithm for Bayesian RL
- Based on a similar algorithm for Partially-Observable MDPs

Details

- Repeat: Sample a transition matrix then plan
  - Avoids costly Bayesian updates in planning
- Uses regular RL in part of the planning (Q learning)
- Aggregate the plans and chose the best action
- Uses UCT for planning and aggregation
  - Efficient tree search algorithm; used in AlphaGo

Objectives Summary

Demonstrate a novel approach to Safe RL in which safe exploration emerges automatically from Bayesian RL on a reward-based objective without explicit safety constraints.

Determine if this is feasible in practice using the BAMCP algorithm.

Investigate whether such exploration is "reasonable"

Compare against existing RL algorithms

Deadly Bandits Environment

MDP Sample

- 2 states: alive and dead (terminal)
- Alive state has N arms
  - Each has a deterministic reward
  - Each has a termination probability; transition to "dead" state

Deadly Bandits Environment

MDP Sample

- 2 states: alive and dead (terminal)
- Alive state has N arms
  - Each has a deterministic reward
  - Each has a termination probability; transition to "dead" state

Prior Distribution

- Arm rewards sampled i.i.d. from $\{0.2, 0.4, 0.6, 0.8, 1\}$
- Termination probs. sampled i.i.d. from $\{0, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$.

Deadly Bandits Environment

MDP Sample

- 2 states: alive and dead (terminal)
- Alive state has N arms
  - Each has a deterministic reward
  - Each has a termination probability; transition to "dead" state

Prior Distribution

- Arm rewards sampled i.i.d. from $\{0.2, 0.4, 0.6, 0.8, 1\}$
- Termination probs. sampled i.i.d. from $\{0, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$.

Posterior Distribution

- Observed rewards are fixed
- Known Risk: Observed termination probs. are fixed
- Unknown Risk: Observed term. probs. sampled given survival count

Experiment

Episode

- Sample a new Known / Unknown Deadly Bandits environment
- Run for 500 steps or until first termination
- Objective: strict evaluation of safe exploration
  - No re-run on same MDP after terminal transition
  - No mistakes allowed
  - No empirical learning about danger other than by posterior

Experiments

- One for each Known and Unknown deadly bandits
- 100,000 independent episodes for baseline agents
- 1,000 episodes for BAMCP agent

Baselines

Uniform Random : Selects actions uniformly at random

Q Learning : Tabular Q learning with ϵ-greedy exploration

Constant : Always chooses action 0

Baselines

Uniform Random : Selects actions uniformly at random

Q Learning : Tabular Q learning with ϵ-greedy exploration

Constant : Always chooses action 0

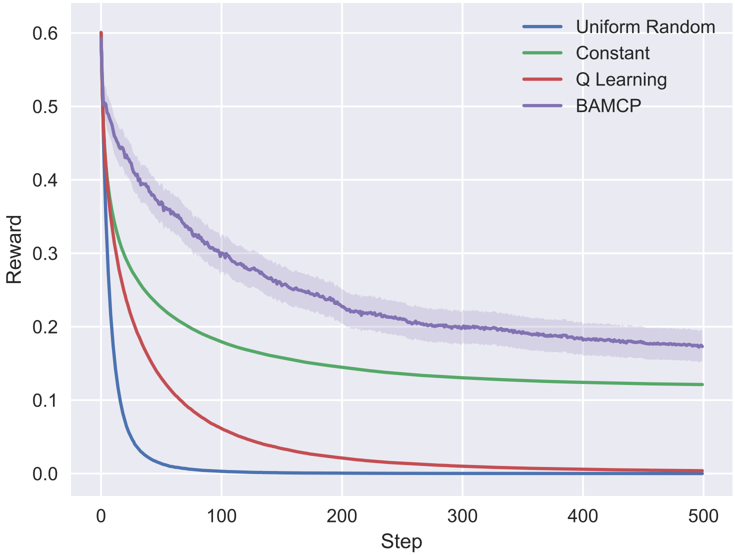
Safe RL

- Unfortunately no baseline Safe RL algorithms implemented
  - Many don't fit this setting; discrete and deterministic history
- Constant as Safe RL Reference:
  - Maximally conservative: never explores
  - As safe as possible without knowledge of prior or posterior.

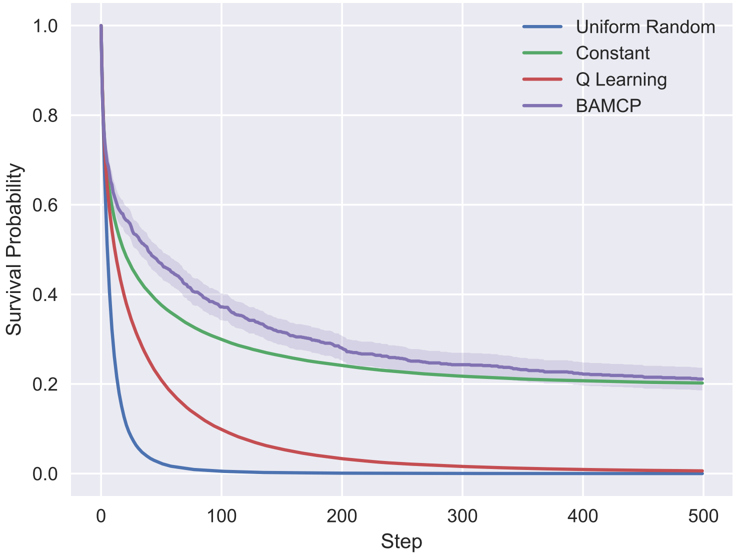
BAMCP Agent

- Perform 20,000 BAMCP search iterations before first action
- 5000 search iterations before other actions
- Discount factor 0.999, discount threshold 0.01
    - => Horizon of 2300 steps
- Internal Q agent
    - Same parameters as baseline Q

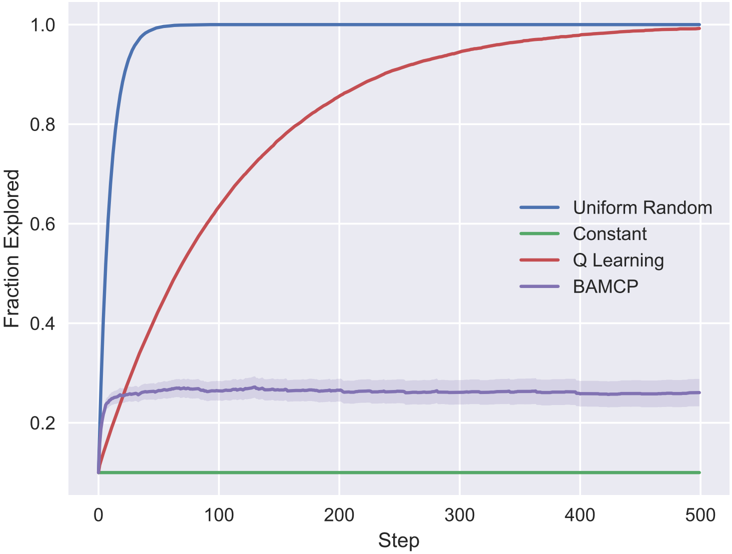
Known Deadly Bandits — Step Reward



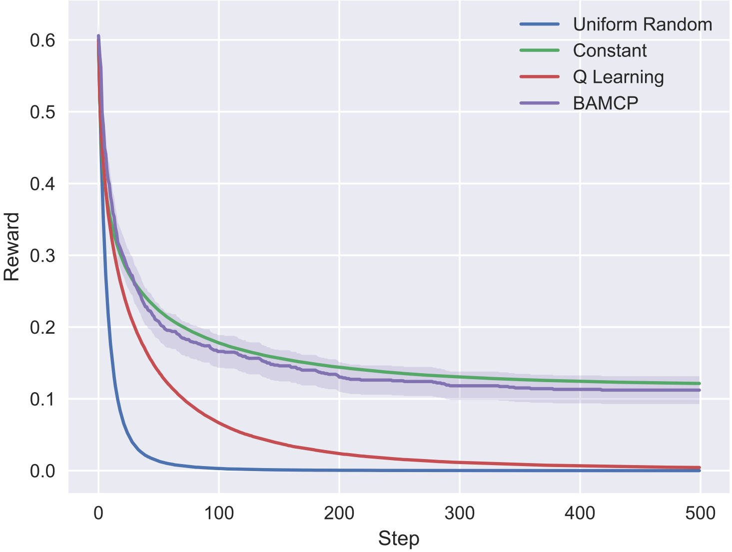
Known Deadly Bandits — Survival Rate



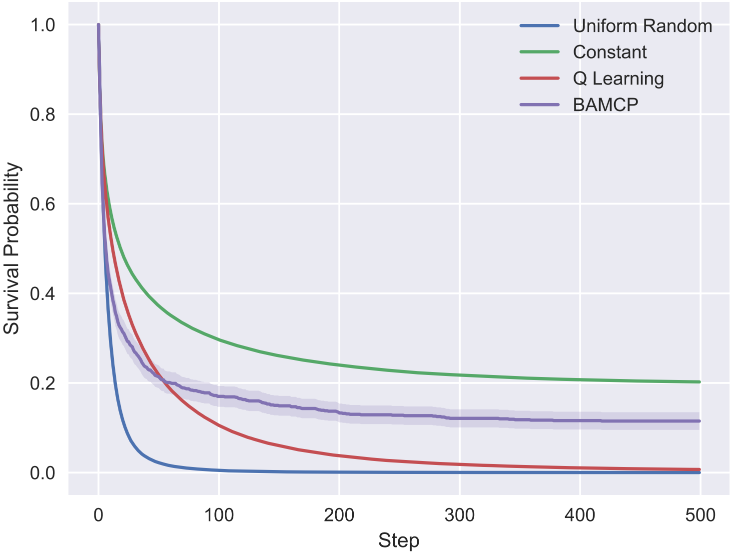
Known Deadly Bandits — Explored Actions



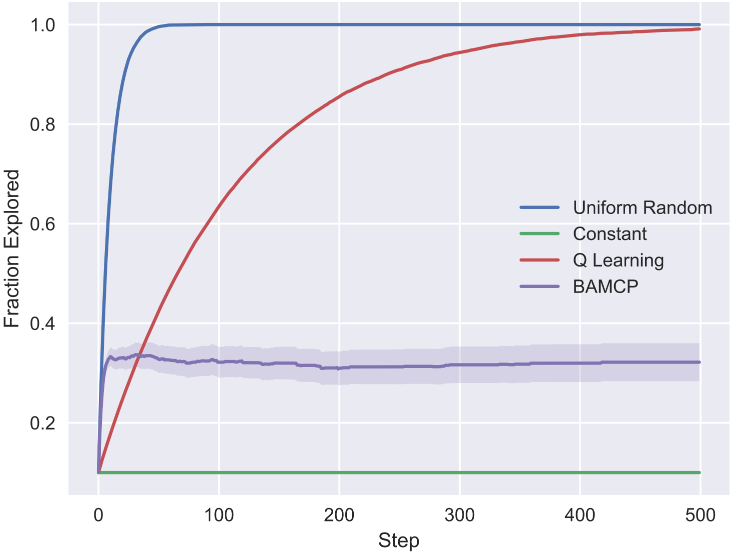
Unknown Deadly Bandits — Step Reward



Unknown Deadly Bandits — Survival Rate



Unknown Deadly Bandits — Explored Actions



Summary Statistics — Known Deadly Bandits

Agent	Episode Reward	Survival Rate	Time Per Step (ms)
Uniform Random	6	0.00006	0.09
Constant	80	0.202 (0.0006)	0.04
Q Learning	23	0.006	0.07
BAMCP	122	0.211 (0.007)	3482.35

Discussion

Demonstrate Emergent Safe Exploration via Bayesian RL on Rewards

Success

BAMCP learns to safely explore in the experiments without an explicit safety objective.

Not perfect: worse than constant on unknown transition probabilities

Discussion

Demonstrate Emergent Safe Exploration via Bayesian RL on Rewards

Success

BAMCP learns to safely explore in the experiments without an explicit safety objective.

Not perfect: worse than constant on unknown transition probabilities

Safe exploration with BAMCP is feasible in practice

Mostly success

BAMCP able to learn safe exploration on a simple environment in practice.

Orders of magnitude slower than baselines but doable.

Discussion

"Reasonable" Safe Exploration

Weak evidence

Explores at the start when exploration is most useful and not afterwards.

Future work: Investigate in more detail.

- Explicit comparison of informative vs. non-informative actions
- High exploration risk vs. moderate long-term risk.

Discussion

"Reasonable" Safe Exploration

Weak evidence

Explores at the start when exploration is most useful and not afterwards.

Future work: Investigate in more detail.

- Explicit comparison of informative vs. non-informative actions
- High exploration risk vs. moderate long-term risk.

Comparison with existing Safe RL

Incomplete

Compared with constant as a baseline but not fully satisfactory.

Existing Safe RL algorithms do not apply easily to the test environment.

Future work: Allow failures during training. Enables more fair comparison with other RL and Safe RL algorithms.

Questions?