Tutorial on Safe Reinforcement Learning

Felix Berkenkamp, Andreas Krause

@EWRL, October 1 2018
Reinforcement Learning (RL)

Reward $r_t$

State $s_{t+1}$

Agent

Environment

Action $a_t$

Need to trade exploration & exploitation

Reinforcement Learning: An Introduction
How can we learn to act safely in unknown environments?
Therapeutic Spinal Cord Stimulation

Safe Exploration for Optimization with Gaussian Processes
Y. Sui, A. Gotovos, J. W. Burdick, A. Krause

Stagewise Safe Bayesian Optimization with Gaussian Processes
Y. Sui, V. Zhuang, J. W. Burdick, Y. Yue
Safe Controller Tuning

Safe Controller Optimization for Quadrotors with Gaussian Processes
Outline

Specifying safety requirements and quantify risk

Acting safely in *known* environments

Acting safely in *unknown* environments

Safe exploration (model-free and model-based)
Specifying safe behavior

Is this trajectory safe?

\[ g(\{s_t, a_t\}_{t=0}^{N}) = g(\tau) > 0 \]

e.g. \[ g(\tau) = \min_{t=1:N} \Delta(s_t, a_t) \]

**Monitoring temporal properties of continuous signals**
O. Maler, D. Nickovic, FT, 2004

**Safe Control under Uncertainty**
D. Sadigh, A. Kapoor, RSS, 2016
What does it mean to be safe?

Fix a policy $a_t = \pi(s_t, \theta)$

Safety $\cong$ avoid bad trajectories (states/actions) $g(\{s_t, a_t\}_{t=0}^N) > 0$

How do I quantify uncertainty and risk?
Stochastic environment / policy

Safety function \( g(\tau) \geq 0 \) is now a random variable \( G \)

\[
P(G = g)
\]

Expected safety \( E[G] \)
Expected safety can be misleading.
Expected safety and variance

\[ P(G = g) \]

\[ E[G] \]
Even at low variance, a significant amount of trajectories may still be unsafe.
Value at Risk

Use confidence lower-bound instead!

\[ \text{VaR}_{0.1}[G] \]

\[ P(G = g) \]

\[ \text{VaR}_\delta[G] = \inf\{\epsilon \in \mathbb{R} : P(G \leq \epsilon) \geq \delta \} \]
Conditional Value at Risk

\[ \text{CVaR}_{0.1}[G] \]

\[ \text{VaR}_{0.1}[G] \]

\[ P(G = g) \]

\[ g \]

\[ \text{VaR}_\delta[G] = \inf \{ \epsilon \in \mathbb{R} : P(G \leq \epsilon) \geq \delta \} \]

\[ \text{CVaR}_\delta[G] = \frac{1}{\delta} \int_0^\delta \text{VaR}_\alpha[G] d\alpha \]
Worst-case

\[ P(G = g) \]

\[ P(G > 0) = 1 \]

or \( g(\tau) > 0 \ \forall \tau \in \Gamma \)
Notions of safety

- **Stochastic**
  - Expected risk: $E[G]
  - Moment penalized: $E[e^{\tau G}]$
  - Value at risk: $\text{VaR}_\delta[G] = \inf\{\epsilon \in \mathbb{R} : P(G \leq \epsilon)\} \geq \delta$
  - Conditional value at risk: $\text{CVaR}_\delta[G] = \frac{1}{\delta} \int_0^\delta \text{VaR}_\alpha[G] d\alpha$

- **Worst-case**
  - $g(\tau) > 0 \ \forall \tau \in \Gamma$
  - $\rightarrow$ Robust Control
  - $\rightarrow$ Formal verification
Acting in known model with safety constraints

Constrained Markov decision processes
Eitan Altman, CRC Press, 1999

Essentials of robust control
Kemin Zhou, John C. Doyle, PH, 1998

Robust control of Markov decision processes with uncertain transition matrices
Arnab Nilim, Laurent El Ghaoui, OR, 2005
Reinforcement Learning

Key challenge: Don’t know the consequences of actions!
How to start acting safely?

No knowledge!
Now what?

Felix Berkenkamp, Andreas Krause
Imitation learning

\[ a_t = \pi^*(s_t) \]

\[ a_t = \pi(s_t, \theta) \]

Dataset \( \mathcal{D} = \{s_t, \pi^*(s_t)\}_{t=0}^{N} \)

\[ \theta = \arg\min_{\theta} \sum_{s,a \in \mathcal{D}} \|\pi(s, \theta) - a_t\| \]

No experience gained here!
Imitation learning algorithms

Data aggregation

Generate state sequence with policy $\pi(s_t, \theta)$

$$D \leftarrow D \cup \{ (s_t, \pi^*(s_t)) \}_{t=1}^T$$

$$\theta = \arg\min_{\theta} \sum_{s,a \in \mathcal{D}} ||\pi(s, \theta) - a_t||$$

Policy aggregation

$$\pi_0 = \pi^*$$

Generate state sequence $\mathcal{D}$ with policy $\pi_i$

$$\theta_{i+1} = \arg\min_{\theta} \sum_{s,a \in \mathcal{D}} ||\pi(s, \theta) - a_t||$$

$$\pi_{i+1}(s) = \alpha_0 \pi^*(s) + \sum_{j=1}^{i+1} \alpha_j \pi(s, \theta_j)$$

Search-based Structured Prediction
Hal Daume III, John Langford, Daniel Marcu, ML, 2009

Efficient Reductions for Imitation Learning
Stephan Ross, Drew Bagnell, AISTATS 2010

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
Stephane Ross, Geoffrey J. Gordon, J. Andrew Bagnell, 2011
Only apply action from learned policy when

\[ \| \pi(s_t, \theta) - \pi^*(s_t) \| \leq \epsilon \]

Query-Efficient Imitation Learning for End-to-End Autonomous Driving
Jiakai Zhang, Kyunghyun Cho, AAAI, 2017

\[ \text{Var}[\pi(s_t, \theta)] \leq \gamma \]

EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning
Kunal Menda, Katherine Driggs-Campbell, Mykel J. Kochenderfer, arXiv2018
What to do with this initial policy?

Can find an initial, safe policy based on domain knowledge.

How to improve?

Agent

Environment

Reward $r_t$

State $s_{t+1}$

Action $a_t$
Prior knowledge as backup for learning

Safety controller takes over

Learner is seen as a disturbance

Know what is safe

Provably safe and robust learning-based model predictive control
A. Aswani, H. Gonzalez, S.S. Satry, C. Tomlin, Automatica, 2013

Safe Reinforcement Learning via Shielding
M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Nickum, U. Topcu, AAAI, 2018

Safe Exploration of State and Action Spaces in Reinforcement Learning
J. Garcia, F. Fernandez, JAIR, 2012

Safe Exploration in Continuous Action Spaces
G. Dalai, K. Dvijotham, M. Veccerik, T. Hester, C. Paduraru, Y. Tassa, arXiv, 2018

Linear Model Predictive Safety Certification for Learning-based Control
K.P. Wabersich, M.N. Zeilinger, CDC, 2018
Prior knowledge as backup for learning

Safety controller takes over

Learner is seen as a disturbance

Know what is safe

Need to know what is unsafe in advance.

Without learning, need significant prior knowledge.

The learner does not know what’s happening!
Safety as improvement in performance (Expected safety)

Initial, stochastic policy \( \pi(s, \theta_b) \)

Performance

\[
J(\theta) = \mathbb{E}_{s_t \sim \rho(\theta)} \left[ \sum_{t=1}^{T} \gamma^t r_t(s_t) \right] = \mathbb{E}_{\tau \sim \rho(\theta)} \left[ g(\tau) \right]
\]

Safety constraint

\[
\Pr( J(\theta) \geq J(\theta_b) ) \geq 1 - \delta
\]

Need to estimate \( J(\theta) \) based only on data from \( \pi(s, \theta_b) \)
Off-Policy Policy Evaluation

$$a_t = \pi(s_t, \theta_b)$$

What does this tell me about a different policy $$\pi(s, \theta)$$?

Importance sampling:

$$\mathbb{E}_{\tau \sim \rho(\theta)} \left[ g(\tau) \right] = \mathbb{E}_{\tau \sim \rho(\theta_b)} \left[ \frac{p(\tau|\theta)}{p(\tau|\theta_b)} g(\tau) \right] \prod_{(s_t, a_t) \in \tau} \frac{p(a_t|s_t, \theta)}{p(a_t|s_t, \theta_b)}$$

(there are better ways to do this)

Eligibility Traces for Off-Policy Policy Evaluation
Doina Precup, Richard S. Sutton, S. Singh
Guaranteeing improvement

Unbiased estimate of $J(\theta)$. What about $\Pr\left( J(\theta) \geq J(\theta_b) \right) \geq 1 - \delta$?

Generate trajectories using $\pi(s, \theta_b)$, $\tau \sim \rho(\theta_b)$

Use concentration inequality to obtain confidence intervals

With probability at least $1 - \delta$:

$$J(\theta) = \mathbb{E}_{\tau \sim \rho(\theta)} \left[ g(\tau) \right] \geq \sum_{i=1}^{N} \frac{p(\tau_i|\theta)}{p(\tau_i|\theta_b)} g(\tau_i) - c(N, \delta)$$
Overview of expected safety pipeline

- Trajectory data: $\tau \sim \rho(\theta_b)$
- Training set
- Test set
- Candidate policy
- Evaluate safety
- Use new policy if safe

**High Confidence Policy Improvement**
Philip S. Thomas, Georgios Theocharous, Mohammad Havamzadeh, ICML 2015

**Safe and efficient off-policy reinforcement learning**
Remi Munos, Thomas Stepleton, Anna Harutyunyan, Marc G. Bellemare, NIPS, 2016

**Constrained Policy Optimization**
Joshua Achiam, David Held, Aviv Tamar, Pieter Abbeel, ICML, 2017
Reviewed safety definitions

Saw how to obtain a first, safe policy

Reviewed a first method for safe learning in expectation

Second half: Explicit safe exploration

More model-free safe exploration

Model-based safe exploration without ergodicity

Stochastic
- Expected risk
- Moment penalized
- VaR / CVaR

Worst-case
- Formal verification
- Robust optimization
Reinforcement learning (recap)

\[ \pi_\theta(x) \]

Policy

Exploration

Policy update
Safe reinforcement learning

<table>
<thead>
<tr>
<th>Statistical models to guarantee safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-free</td>
</tr>
</tbody>
</table>

$$a_t = \pi(s_t, \theta)$$  
$$[s_{t+1}, r_t] \sim P(\cdot | s_t, a_t; \theta)$$

Estimate $J(\theta)$ and optimize  
Estimate/identify, then plan/control

Model-free reinforcement learning

\[ a_t = \pi(s_t, \theta) \]

<table>
<thead>
<tr>
<th>Tracking performance</th>
<th>( \max_\theta J(\theta) )</th>
<th>Few, noisy experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety constraint</td>
<td>( g(\theta) \geq 0 )</td>
<td>Safety for all experiments</td>
</tr>
</tbody>
</table>
Safe policy optimization

Goal: \( \max_{\theta} J(\theta) \) s.t. \( g(\theta) \geq 0 \)

Safety: \( g(\theta_t) \geq 0 \) for all \( t \) with probability \( \geq 1-\delta \)
Safe policy optimization illustration

\[ J(\theta) = g(\theta) \]

- Safe seed
- Reachable optimum
- Global optimum
- Safety threshold
Starting Point: Bayesian Optimization

\[
\theta_t \rightarrow y_t = J(\theta_t) + \epsilon_t
\]

**Acquisition function**

**Expected/most prob. improvement** [Močkus et al. ’78,’89]
**Information gain about maximum** [Villemonteix et al. ’09]
**Knowledge gradient** [Powell et al. ’10]
**Predictive Entropy Search** [Hernández-Lobato et al. ’14]
**TruVaR** [Bogunovic et al.’17]
**Max Value Entropy Search** [Wang et al’17]

**Constraints/Multiple Objectives**
[Snoek et al. ‘13, Gelbart et al. ’14, Gardner et al. ‘14, Zuluaga et al. ‘16]
Gaussian process
Gaussian process

Output

Input
Gaussian process
Gaussian process
Gaussian process

\[ \text{Cov}[J(\theta), J(\theta')] = k(\theta, \theta') \]
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
SafeOPT: Constrained Bayesian optimization

Performance $J(\theta)$

Safety $g(\theta)$

Parameters $\theta$
Theorem (informal):

Under suitable conditions on the kernel and on \( J, g \), there exists a function \( T(\epsilon, \delta) \) such that for any \( \epsilon > 0 \) and \( \delta > 0 \), it holds with probability at least \( 1 - \delta \) that

1) \text{SAFEOPT never makes an unsafe decision}

2) After at most \( T(\epsilon, \delta) \) iterations, it found an \( \epsilon \)-optimal reachable point

\[
T(\epsilon, \delta) \in \mathcal{O}\left(\left(\|J\|_k + \|g\|_k\right) \frac{\log^3 1/\delta}{\epsilon^2}\right)
\]
Modelling context

\[ \text{Cov}[J(\theta), J(\theta')] = k(\theta, \theta') \]

Additional parameters

\[ \text{Cov}[J(\theta, z), J(\theta', z')] = k(\theta, \theta') \ast k(z, z') \]
Therapeutic Spinal Cord Stimulation

Stagewise Safe Bayesian Optimization with Gaussian Processes

Y. Sui, V. Zhuang, J. W. Burdick, Y. Yue
Virtual vs. Real: Trading Off Simulations and Physical Experiments in Reinforcement Learning with Bayesian Optimization
A. Marco, F. Berkenkamp, P. Hennig, A. Schöllig, A. Krause, S. Schaal, S. Trimpe, ICRA'17
Modeling this in a Gaussian process

\[ J(\theta) = J_{\text{sim}}(\theta) + \Delta(\theta) \]

- simulation
- experiment
Performance improvement

Starting controller

Learned controller

Video at https://youtu.be/oq9Qgq1lpp8
Safe reinforcement learning

<table>
<thead>
<tr>
<th>Statistical models to guarantee safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-free</td>
</tr>
</tbody>
</table>

- \( a_t = \pi(s_t, \theta) \)
- \([s_{t+1}, r_t] \sim P(\cdot | s_t, a_t; \theta)\)

Estimate \( J(\theta) \)
and optimize

Estimate/identify,
then plan/control
From bandits to Markov decision processes

Can use the same Bayesian model to determine safety of states
Challenges with long-term action dependencies

Non-ergodic MDP
Rendering exploration safe

Exploration:
 Reduce model uncertainty
 Only visit states from which the agent can recover safely

Safe Exploration in Markov Decision Processes
T.M. Moldovan, P. Abbeel, ICML, 2012

Safe Exploration in Finite Markov Decision Processes with Gaussian Processes

Safe Control under Uncertainty
D. Sadigh, A. Kapoor, RSS, 2016

Safe Exploration and Optimization of Constrained MDPs using Gaussian Processes
Akifumi Wachi, Yanan Sui, Yisong Yue, Masahiro Ono, AAAI, 2018
On a real robot
From bandits to Markov decision processes

Next: model-based reinforcement learning
Reinforcement learning (recap)

Policy

$\pi_\theta(x)$

Exploration

Policy update
Model-based reinforcement learning

Policy

\[ \pi_\theta(x) \]

Exploration

Model learning

Safe model-based reinforcement learning

Policy

\[ \pi_\theta(x) \]

Safe exploration

Statistical model learning
A Bayesian dynamics model

\[ s_{t+1} = f(s_t, a_t) + h(s_t, a_t) \]

- \( f(s_t, a_t) \): a priori model
- \( h(s_t, a_t) \): unknown model

\[ \mu(s, a) = f(s, a) \]
Region of attraction

Baseline policy is safe

\[ S \]

\( s_0 \)

unsafe
Linear case

\[ s_{t+1} = A \, s_t + B \, a_t \]

Uncertainty about entries

Designing safe controllers for quadratic costs is a convex optimization problem

Safe and Robust Learning Control with Gaussian Processes
F. Berkenkamp, A.P. Schoellig, ECC, 2015

Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator
S. Dean, H. Mania, N. Matni, B. Recht, S. Tu, arXiv, 2018
Outer approximation contains true dynamics for all time steps with probability at least $1 - \delta$

**Learning-based Model Predictive Control for Safe Exploration**
T. Koller, F. Berkenkamp, M. Turchetta, A. Krause, CDC, 2018
Region of attraction

Theorem (informally): Under suitable conditions can always guarantee that we are able to return to the safe set.
Learning-based Model Predictive Control for Safe Exploration
T. Koller, F. Berkenkamp, M. Turchetta, A. Krause, CDC, 2018

Reachability-Based Safe Learning with Gaussian Processes

Robust constrained learning-based NMPC enabling reliable mobile robot path tracking

Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control
S. Kamthe, M.P. Deisenroth, AISTATS, 2018

Chance Constrained Model Predictive Control
A.T. Schwarm, M. Nikolaou, AIChE, 1999
Example

Robust constrained learning-based NMPC enabling reliable mobile robot path tracking

Video at https://youtu.be/3xRNmNv5Efk
Region of attraction

$\mathcal{S}$

first step same

exploration trajectory

$\mathcal{V}(c_0)$

unsafe

safety trajectory

Exploration limited by size of the safe set!

Felix Berkenkamp, Andreas Krause
Theorem (informally):
Under suitable conditions can identify (near-)maximal subset of $X$ on which $\pi$ is stable, while never leaving the safe set.
Lyapunov functions

\[ s_{t+1} = f(s_t, \pi(s, \theta)) \]

Lyapunov Design for Safe Reinforcement Learning
T.J. Perkings, A.G. Barto, JMLR, 2002

\[ V(s_{t+1}) < V(s_t) \]
\[ \forall s_t \in \mathcal{V}(c) \setminus \mathcal{V}(c_0) \]

[A.M. Lyapunov 1892]
Lyapunov functions

\[ s_{t+1} = f(s_t, \pi(s, \theta)) + g(s_t, \pi(s, \theta)) \]

\[ \Pr \left\{ V(s_{t+1}) < V(s_t); \forall s_t \in \mathcal{V}(c) \setminus \mathcal{V}(c_0) \right\} \geq 1 - \delta \]
Need to **safely** explore!
Illustration of safe learning

Safe Model-based Reinforcement Learning with Stability Guarantees
Finding the right Lyapunov function is difficult! 

\[ V(s) = \phi_\theta(s)^T \phi_\theta(s) \]

Weights - positive-definite 
Nonlinearities - trivial nullspace

Decision boundary 
\[ V(s) = 1 \]
\[ V(s_{t+1}) < V(s_t) \]
\[ \forall s_t \in V(c) \setminus V(c_0) \]

The Lyapunov Neural Network: Adaptive Stability Certification for Safe Learning of Dynamic Systems 
S.M. Richards, F. Berkenkamp, A. Krause
Towards safe reinforcement learning

How to verify safety of a given policy

\( S \)

first step same

exploration trajectory

unsafe

safety trajectory

How to verify safety of a given policy

\( V(c_0) \)
Summary

Reviewed safety definitions

Saw how to obtain a first, safe policy

Reviewed a first method for safe learning in expectation

Safe Bayesian optimization for safe exploration

How to transfer this intuition to the safe exploration in MDPs

Model-based methods (reachability=safety, certification, exploration)

Stochastic
  - Expected risk
  - Moment penalized
  - VaR / CVaR

Worst-case
  - Formal verification
  - Robust optimization
Where to go from here?

- Machine learning
- Control theory
- Safe Reinforcement Learning
- Format methods
- Statistics
- Decision theory

Scalability (computational & statistical)
Safe imitation learning
Tradeoff safety and performance (theory & practice)
Lower bounds; define function classes that are safely learnable