Safe Exploration Using Bayesian World Models and Log-Barrier Optimization

Yarden As
yarden.as@inf.ethz.ch
Learning and Adaptive Systems
ETH Zurich

Bhavya Sukhija
bhavya.sukhija@inf.ethz.ch
Learning and Adaptive Systems
ETH Zurich

Andreas Krause
krausea@ethz.ch
Learning and Adaptive Systems
ETH Zurich

Abstract

A major challenge in deploying reinforcement learning in online tasks is ensuring that safety is maintained *throughout* the learning process. In this work, we propose CERL, a new method for solving constrained Markov decision processes while keeping the policy safe during learning. Our method leverages Bayesian world models and suggests policies that are pessimistic w.r.t. the model's epistemic uncertainty. This makes CERL robust towards model inaccuracies and leads to safe exploration during learning. In our experiments, we demonstrate that CERL outperforms the current state-of-the-art in terms of safety and optimality in solving CMDPs from image observations.

1 Introduction

Despite notable progress in reinforcement learning (RL), its application outside of simulators remains largely limited. This is primarily because exploration in RL often requires an abundance of samples and is inherently unsafe. Furthermore, while RL methods assume full observability of the environment state, in many cases this assumption is not very realistic. For example, even in a simple navigation task, it is not realistic to have direct access to the positions of all obstacles. Therefore, the goal of this work is to design a method that can efficiently learn while also ensuring the safety of themselves and their surroundings, even in light of partial-observability.

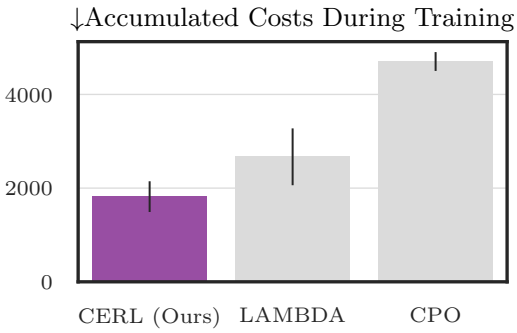

Figure 1: We average the accumulated costs for each training run. Error bars represent the standard deviation across all training runs. As shown, CERL outperforms the baseline algorithms with respect to the accumulated costs during training.

Safety in RL is typically modeled via constrained Markov decision processes (CMDP) (Altman, 1999). CMDPs extend MDPs by incorporating additional cost functions to indicate unsafe behavior. There are several model-free algorithms (Chow et al., 2015; Achiam et al., 2017; Ray et al., 2019; Chow et al., 2019) which show asymptotic convergence to a safe policy. However, these methods are mostly sample inefficient and unsafe during learning, making them ill-suited for online learning in real-world applications. Model-based RL (Deisenroth & Rasmussen, 2011; Chua et al., 2018; Hafner et al., 2019a; Janner et al., 2019) is a more promising alternative to improve sample efficiency.

In this work, we address this precise gap and propose Cost-Efficient Reinforcement Learning (CERL). CERL learns an uncertainty-aware transition model of the underlying MDP and suggests a policy that is pessimistic with respect to the model's epistemic uncertainty. We build upon recent advances in black-box constrained optimization (Usmanova et al., 2022) to propose an RL algorithm that is efficient and safe during learning. We leverage the recurrent state space model (RSSM) (Hafner et al., 2019a) and scale CERL to a high-dimensional real-world setting where the agent only has access to image observations. In our experiments on the Safety-Gym benchmark suite (Ray et al., 2019), we show that CERL learns the optimal policy considerably faster than state-of-the-art RL algorithms for CMDPs while also being *safe during exploration*.

Our contributions

- We empirically show that CERL successfully solves complex navigation tasks with image observations from Safety-Gym while maintaining safety during learning. To the best of our knowledge, this is the first work that achieves milestone towards safe online RL. We summarize this result in Figure 1.
- We demonstrate that using CERL for safe exploration does not degrade the performance *at the end of training* and is on par with previous state-of-the-art methods for this problem.

2 Related Works

Safe RL in continuous domains Multiple methods have been previously proposed to solve CMPDs in continuous domains (Achiam et al., 2017; Chow et al., 2019; Tessler et al., 2019; Liu et al., 2022). Notably, Dalal et al. (2018) propose a safety filter approach to ensure safe exploration with state-wise constraints. While Dalal et al. (2018) demonstrate strong empirical results, their safety filter lacks guarantees on optimality and safety. This work relies on a constrained optimizer that, under moderate assumptions, has guarantees on (local) optimality and constraint satisfaction. Berkenkamp et al. (2017) interpret safety with Lyapunov stability, to derive a method that is theoretically guaranteed to be safe and can (empirically) safely learn control policies in small-scale continuous domains such as an inverted pendulum. This work takes one step further by performing safe exploration in continuous domains and under partial-observability. Similarly to this work, As et al. (2022) propose a Bayesian model-based approach that solves CMDPs from high-dimensional inputs such as image observations. This work builds on the same ideas from As et al. (2022), though significantly improves safety performance during learning thanks to improved uncertainty estimation and a novel solver for stochastic and constrained optimization problems which ensures feasibility of optimization iterates (Usmanova et al., 2022).

Other works on safe exploration While the main contribution of this work is of empirical nature, we note a few other works that focus on the theoretical challenges of this problem. Berkenkamp et al. (2021) and extensions thereof (Turchetta et al., 2019; Baumann et al., 2021; Sukhija et al., 2023; Hübotter et al., 2024) propose a general-purpose safe RL algorithm and apply it for tuning controllers for robotic systems such as quadrupeds (Widmer et al., 2023). The proposed methods come with strong theoretical guarantees on the safety and optimality of the algorithm and also demonstrate empirical safety and sample efficiency when evaluated on hardware. Despite explicitly addressing the primary challenges that arise in safe exploration, which we outline in Section 3, these methods focus primarily on Gaussian Processes (GP) and are limited to low-dimensional policies, making them difficult to scale. Lastly, (Efroni et al., 2020) analyze the exploration-exploitation dilemma in tabular CMDPs. Efroni et al. (2020) do not treat the safe exploration problem as a hard requirement, but derive (sublinear) regret bounds for constraint violation during learning.

3 Problem Setting

(Partially-observable) Markov decision processes We study an episodic, discrete-time, Markov decision process (MDP). The environment's state at time t is defined as $\boldsymbol{s}_t \in \mathbb{R}^n$, the

agent can take an action $\boldsymbol{a}_t \in \mathbb{R}^m$. Each episode starts by sampling from the initial-state distribution $\boldsymbol{s}_0 \sim \rho(\boldsymbol{s}_0)$. At each time step t, the agent observes an observation $\boldsymbol{o}_t \sim p(\boldsymbol{o}_t|\boldsymbol{s}_t)$ and takes an action by sampling from a policy distribution $\boldsymbol{a}_t \sim \pi(\cdot|\boldsymbol{o}_{:t})$. The next state is then sampled from an unknown transition distribution $\boldsymbol{s}_{t+1} \sim p(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$ and a reward $r_t \sim p(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$ is obtained. To learn, the agent collects data by drawing trajectories $\tau \sim p(\tau) = \prod_{t=0}^{T} \pi(\boldsymbol{a}_t|\boldsymbol{o}_{:t}) p(\boldsymbol{o}_t|\boldsymbol{s}_t) p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \rho(\boldsymbol{s}_0)$. The goal is to efficiently collect data to learn a policy that maximizes the sum of rewards over a horizon T, that is

$$J(\pi, p) = \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} r_t \right]. \tag{1}$$

Constrained Markov decision processes (CMDP) CMDPs (Altman, 1999) extend general MDP formulation to the constrained setting. In CMDPs, the agent observes a cost signal $c_t \sim p(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$ alongside the reward. While in the general case CMDPs consider multiple cost functions, in this work we focus on the single-constraint setting for conciseness, highlighting that our results can be easily extended to the multi-constraint setting. Given c_t, we define the constraints over the horizon T as

$$J^c(\pi, p) = \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} c_t \right] \leq 0. \tag{2}$$

For instance, a common cost function is $c(\boldsymbol{s}_t) = \mathbf{1}_{\mathrm{s_t} \in \mathcal{H}}$, where $\mathcal{H}$ is the set of harmful states. In the CMDP setting, the goal is to find a policy π for the true unknown dynamics $p^\star$ that solves the following problem

$$\max_{\pi \in \Pi} J(\pi, p^\star) \quad \text{s.t.} \quad J^c(\pi, p^\star) \leq 0. \tag{3}$$

Model-based reinforcement learning In model-based reinforcement learning (MBRL), at each iteration, the agent collects a dataset $\mathcal{D}$ of observed trajectories $\{\tau_1, \ldots, \tau_M\}$ to fit a statistical model $p_\theta(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$ that approximates the true transition distribution $p^\star$. We focus on parametric models that use parameters θ to learn the dynamics.[1] The agent uses the estimated model for planning, either within an online MPC scheme (Chua et al., 2018) or via policy optimization (Janner et al., 2019). Model-based RL, as opposed to its model-free counterpart, is known to be more sample-efficient (Deisenroth & Rasmussen, 2011; Chua et al., 2018; Hafner et al., 2019b; Curi et al., 2020) making it better suited for learning online.

Safe exploration While Equation (3) only requires the policy to satisfy the constraint in Equation (2) *at the end of learning*, learning entails *exploration* of the CMDP, which, without special care, may cause the agent *to violate the constraints*, as we also show in Section 5. Concretely, for each learning iteration $n \in \mathbb{N}$ the agent must satisfy $J_n^c(\pi_n, p^\star) \leq 0$. To overcome this challenge, the agent must explore only within areas that are deemed to be safe with high probability. This involves three algorithmic challenges: **(1)** estimating a pessimistic set of safe policies; **(2)** improving the policy only within this safe set, and finally, **(3)** expanding the safe set. See Sui et al. (2015) for the general black-box optimization setting and for a more thorough discussion. The contributions of this paper mainly focus on the first and second challenges. The third challenge generally requires some form of pure exploration (Amani et al., 2019; Hübotter et al., 2024).

4 Cost-Efficient Reinforcement Learning (CERL)

In the following, we propose our algorithm, which learns an uncertainty-aware transition distribution and uses it to *maintain safety during learning*.

[1] Non-parametric models can be successfully used in this setting as well, albeit harder to scale (Berkenkamp et al., 2017).

Leveraging Bayesian world models To handle partial-observability, we choose to base our world model on the Recurrent State Space Model (RSSM) introduced in Hafner et al. (2019a). The RSSM can be thought of as a sequential variational auto-encoder that learns the (latent) dynamics $p_\theta(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$. To quantify the uncertainty over the RSSM's parameters, we take a Bayesian approach, where we adopt a prior on the model parameters and estimate the posterior using approximate Bayesian inference techniques, in particular probabilistic ensembles (Lakshminarayanan et al., 2017). A posterior distribution over model parameters allows the agent reason about what is *unknown* during learning (Ghavamzadeh et al., 2015). Such Bayesian reasoning forms the basis of many MBRL algorithms (Deisenroth & Rasmussen, 2011; Chua et al., 2018; Curi et al., 2020; Sukhija et al., 2024) and is commonly used to drive (provably-efficient) exploration (Auer & Ortner, 2007).

Estimating the pessimistic safe set Extending these ideas to safety, we define a set of plausible dynamics $\mathcal{P}$ and let $p_\theta \in \mathcal{P}$ be a particular transition density in this set. We assume that the true model $p^\star$ is within the support of $\mathcal{P}$. We approximate $\mathcal{P}$ by sampling $\theta \sim p(\theta|\mathcal{D})$ and taking the union over the different samples i.e., $\mathcal{P} = \bigcup_{i=0}^{N-1}\{p_{\theta_i}\}$, where N is the number of samples. Since $p^\star \in \mathcal{P}$, we can ensure constraint satisfaction for $p^\star$ by picking a policy that satisfies the constraints for all transition distributions in $\mathcal{P}$. This motivates the following constrained optimization problem

$$\max_{\pi_n \in \Pi} J(\pi_n, p_\theta) \quad \text{s.t.} \quad \max_{p_{\theta^i} \in \mathcal{P}} J^c(\pi_n, p_{\theta^i}) \leq 0 \; \forall n \tag{4}$$

Equation (4) picks a policy that satisfies the constraints for the worst-case model in $\mathcal{P}$, i.e., is pessimistic with respect to the constraints and dynamics in $\mathcal{P}$. While for many real-world settings, it is challenging to verify if $p^\star \in \mathcal{P}$, Equation (4) can still be viewed as being robust to model inaccuracies. In practice, we evaluate the policy independently using each of the models p_{θ_i} and pick the most pessimistic evaluation.

Policy improvement within the safe set To solve the constrained optimization in Equation (4), we use Log-Barriers SGD (LBSGD), a constrained black-box optimizer proposed by Usmanova et al. (2022). LBSGD is an interior-point method that guarantees all iterates to be feasible, that is, to remain within the safe set. To achieve that, LBSGD finds a (noisy) estimate of the log-barrier function

$$B_\eta(\pi_n) = J(\pi_n, p_\theta) - \eta \log\left(-J_P^c(\pi_n)\right) \tag{5}$$

$$\nabla B_\eta(\pi_n) = \nabla J(\pi_n, p_\theta) + \eta \frac{\nabla J_P^c(\pi_n)}{-J_P^c(\pi_n)} \tag{6}$$

whereby $J_P^c(\pi_n) = \max_{p_{\theta_i} \in \mathcal{P}} J^c(\pi_n, p_{\theta_i})$. Estimating $\nabla B_\eta(\pi_n)$ is done by drawing mini-batches of states, planning with the model and backpropagating gradients through the (worst-case) model akin to Hafner et al. (2021) and As et al. (2022). LBSGD ensures that distance is always kept from the boundaries of the safe set *from its interior* by adaptively changing SGD's step size based on the gradient direction of $J_P^c(\pi_n)$ and smoothness assumptions $J(\pi_n, p_\theta)$ and $J_P^c(\pi_n)$. Overall, LBSGD can give feasibility guarantees under stricter assumptions such as smoothness of $p^\star$, safe initialization of π_0 and an unbiased gradient estimator of $\nabla B_\eta(\pi_n)$. While these assumptions are hard to validate in practice, our experiments show LBSGD's utility for safe exploration even without formal guarantees.

5 Experiments

Setup We study CERL's performance on the Safety-Gym benchmark suite for safe learning in CMDPs. We repeat the same experimental setup in Ray et al. (2019) and As et al. (2022). In particular, each episode has a length of $T = 1000$ steps. We set the cost budget for each episode to $d = 25$ as described by Ray et al. (2019). We measure CERL's performance on the three tasks of Safety-Gym with the Point robot. We deviate from Safety-Gym by increasing the number of obstacles, more details can be found in our open-source implementation `https://anonymous.4open.science/r/safe-opax-F5FF/`. After each training epoch we estimate $J(\pi_n, p^\star)$ and $J^c(\pi_n, p^\star)$ by fixing the policy and sampling 10 episodes (denoting the estimates with $\hat{J}(\pi_n, p^\star)$ and $\hat{J}^c(\pi_n, p^\star)$).

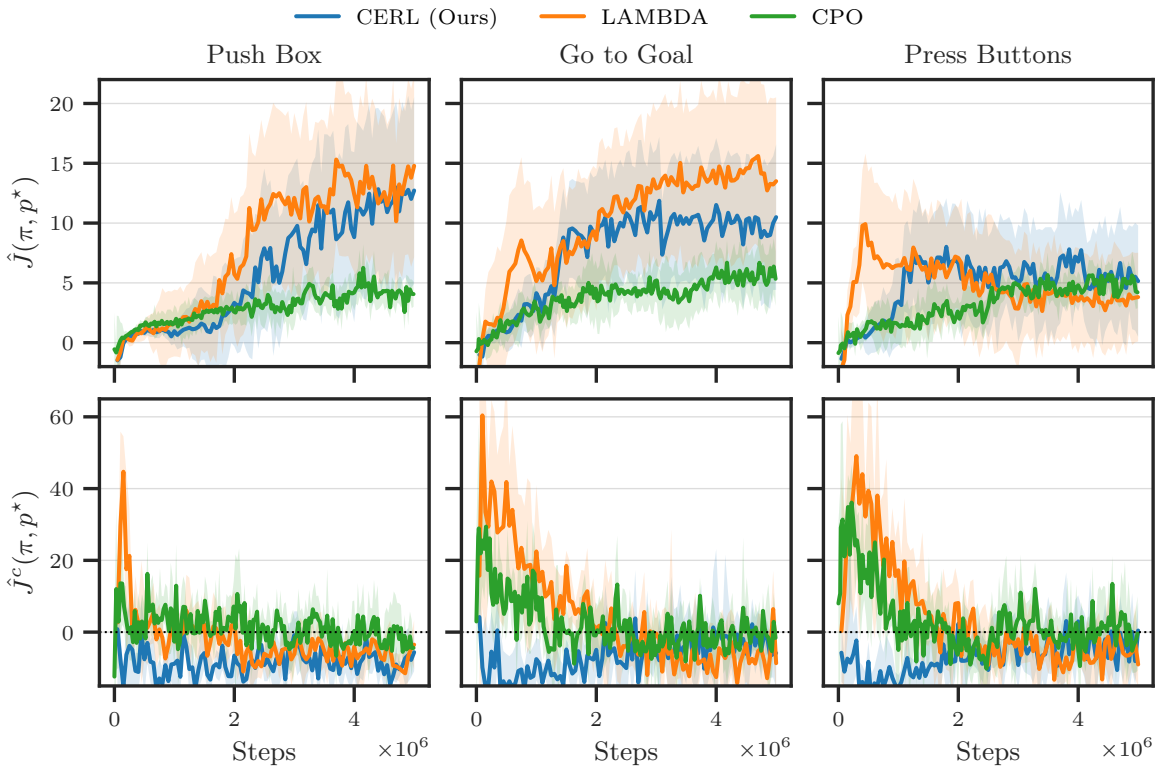

Figure 2: Learning curves for the objective and constraint for CERL and the baseline algorithms.

In all our experiments we use 5 random seeds and report the median and standard deviation across these seeds. Finally, we use a budget of 5M training steps for each training run.

Baselines We compare CERL with two strong baselines. The first baseline is LAMBDA (As et al., 2022). LAMBDA uses an "Augmented Lagrangian" approach to solve Equation (4). As in this work, LAMBDA uses images as state observations. The second baseline we compare with is Constrained Policy Optimization (CPO) (Achiam et al., 2017). CPO is considered a standard baseline to solving CMDPs due to its consistent performance, akin to PPO (Schulman et al., 2017) for standard RL. Unlike LAMBDA and CERL, CPO is an on-policy, *model-free* algorithm and is generally considered significantly less sample-efficient.

Results and discussion We present our results in Figure 2. First, observe that CERL is the only algorithm that maintains safety *throughout* the learning process. Specifically, it takes CPO and LAMBDA roughly 1.5M training steps to satisfy the constraints. As opposed to LAMBDA, CERL uses a stronger black-box optimizer from Usmanova et al. (2022). We believe this plays a crucial role in obtaining empirical safety. Moreover, CERL is safer on all tasks and being only slightly outperformed by LAMBDA at the end of training in the "Go to Goal" task. Generally, it is known that safe exploration comes at a price for optimality (Berkenkamp et al., 2021) and this task highlights the natural trade-off between better performance and safety. In all cases, after a budget of 5M steps, both LAMBDA and CERL outperform CPO. Our results indicate that CERL can be used to learn safe policy online in real-world settings, as it is more sample efficient than CPO, safe during learning as opposed to LAMBDA, and operates directly in the observation space.

6 Outlook

In this paper we introduce CERL. Our experiments demonstrate that CERL improves on previous work by maintaining safety *during learning*. CERL suffers from two limitations. First, it is hard to realistically satisfy LBSGD's assumptions, and thus practically impossible to theoretically guarantee safe exploration in general. Secondly, even though CERL satisfies the constraints in the classical CMDP setting, where we bound the *expected cost return*, in many real applications we must enforce state-wise safety. Still, this empirical result shows that *safe exploration in high dimensions is possible*, giving hope for more theoretically-grounded methods as well as bridging the gap between practice and theory.

Broader Impact Statement

We design a new method for solving CMDPs while ensuring safety during learning. We believe that one of the greatest current challenges in applying online reinforcement learning "in the wild" is making sure that safety requirements are kept at all times. Addressing this challenge is an important step towards deploying reinforcement learning agents on real robotic systems, allowing them to continually improve while maintaining safety.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization, 2017.

E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints, 2019.

Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models, 2022.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. URL https://proceedings.neurips.cc/paper/2006/file/c1b70d965ca504aa751ddb62ad69c63f-Paper.pdf.

Dominik Baumann, Alonso Marco, Matteo Turchetta, and Sebastian Trimpe. Gosafe: Globally optimal safe robot learning. In *ICRA*, 2021.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees, 2017.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pp. 1–35, 2021.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *CoRR*, abs/1512.01629, 2015. URL http://arxiv.org/abs/1512.01629.

Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control, 2019.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *CoRR*, abs/1805.12114, 2018. URL http://arxiv.org/abs/1805.12114.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning, 2020.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces, 2018.

Marc Peter Deisenroth and Carl Edward Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 465–472, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps, 2020.

Mohammed Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5–6):359–483, 2015. ISSN 1935-8245. doi: 10.1561/2200000049. URL http://dx.doi.org/10.1561/2200000049.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565, 2019a.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *CoRR*, abs/1912.01603, 2019b. URL http://arxiv.org/abs/1912.01603.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2021.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Information-based transductive active learning, 2024.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Zhiwei Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning, 2022.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 997–1005, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/sui15.html.

Bhavya Sukhija, Matteo Turchetta, David Lindner, Andreas Krause, Sebastian Trimpe, and Dominik Baumann. Gosafeopt: Scalable safe exploration for global optimization of dynamical systems. *Artificial Intelligence*, 2023.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. *NeurIPS*, 2024.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control, 2019.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration for interactive machine learning. *NeurIPS*, 32, 2019.

Ilnura Usmanova, Yarden As, Maryam Kamgarpour, and Andreas Krause. Log barriers for safe black-box optimization with application to safe reinforcement learning. *arXiv preprint arXiv:2207.10415*, 2022.

Daniel Widmer, Dongho Kang, Bhavya Sukhija, Jonas Hübotter, Andreas Krause, and Stelian Coros. Tuning legged locomotion controllers via safe bayesian optimization. In *Conference on Robot Learning*, 2023.