Extending Deep Model Predictive Control with Safety Augmented Value Estimation from Demonstrations

Brijen Thananjeyan*, Ashwin Balakrishna*, Ugo Rosolia, Felix Li, Rowan McAllister, Joseph E. Gonzalez, Sergey Levine, Francesco Borrelli, Ken Goldberg
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{bthananjeyan, ashwin_balakrishna}@berkeley.edu
* equal contribution

Abstract

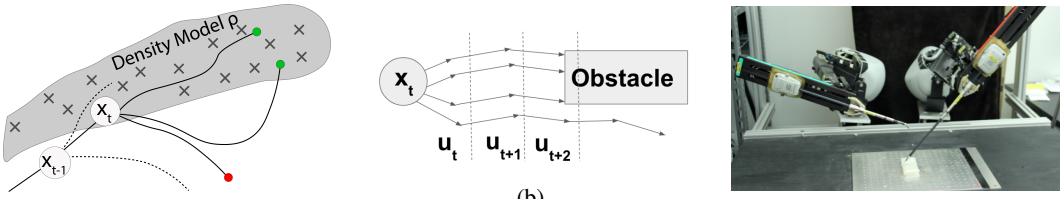
Reinforcement learning (RL) for robotics is challenging due to the difficulty in hand-engineering a dense cost function, which can lead to unintended behavior, and dynamical uncertainty, which makes it hard to enforce constraints during learning. We address these issues with a new model-based reinforcement learning algorithm, safety augmented value estimation from demonstrations (SAVED), which uses supervision that only identifies task completion and a modest set of suboptimal demonstrations to constrain exploration and learn efficiently while handling complex constraints. We derive iterative improvement guarantees for SAVED under known stochastic nonlinear systems. We then compare SAVED with 3 state-of-the-art model-based and model-free RL algorithms on 6 standard simulation benchmarks involving navigation and manipulation and 2 real-world tasks on the da Vinci surgical robot. Results suggest that SAVED outperforms prior methods in terms of success rate, constraint satisfaction, and sample efficiency, making it feasible to safely learn complex maneuvers directly on a real robot in less than an hour. For tasks on the robot, baselines succeed less than 5% of the time while SAVED has a success rate of over 75% in the first 50 training iterations.

1 Introduction

To use RL in the real world, algorithms need to be efficient, easy to use, and safe, motivating methods which are reliable even with significant uncertainty. Deep model-based reinforcement learning (deep MBRL) is an area of current interest because of its sample efficiency advantages over model-free methods in a variety of tasks, such as assembly, locomotion, and manipulation [13, 15, 16, 26, 28, 36, 43]. However, past work in deep MBRL typically requires dense hand-engineered cost functions, which are hard to design and can lead to unintended behavior [2]. It would be easier to simply specify task completion in the cost function, but this setting is challenging due to the lack of expressive supervision. This motivates using demonstrations, which allow the user to roughly specify desired behavior without extensive engineering effort. Furthermore, in many robotic tasks, specifically in domains such as surgery, safe exploration is critical to ensure that the robot does not damage itself or cause harm to its surroundings. To enable this, deep MBRL algorithms also need the ability to satisfy complex constraints.

We develop a method to efficiently use deep MBRL in dynamically uncertain environments with both sparse costs and complex constraints. We address the difficulty of hand-engineering cost functions by using a small number of suboptimal demonstrations to provide a signal about delayed costs in sparse cost environments, which is updated based on agent experience. Then, to enable stable policy improvement and constraint satisfaction, we impose two probabilistic constraints to (1) constrain exploration by ensuring that the agent can plan back to regions in which it is confident in task completion with high probability and (2) leverage uncertainty estimates in the learned dynamics to implement chance constraints [38] during learning. The probabilistic implementation of constraints

Preprint. Under review.



(a) (b) (c)

Figure 1: **SAVED: (a) Task completion driven exploration:** A density model is used to represent the region in state space where the agent has high confidence in task completion, where trajectory samples over the learned dynamics that do not have sufficient density at the end of the planning horizon are discarded; **(b) Chance constraint enforcement:** Implemented by sampling imagined rollouts over the learned dynamics for the same sequence of actions multiple times and estimating the probability of constraint violation by the percentage of rollouts that violate a constraint. In the above example, 75% of the rollouts are constraint violating, making $u_{t:t+H-1}$ a poor control choice; **(c) Setup for physical experiments:** SAVED is able to safely learn complex surgical maneuvers on the da Vinci surgical robot, which is difficult to precisely control [49].

makes this approach broadly applicable, since it can handle settings with significant dynamical uncertainty, where enforcing constraints exactly is difficult.

We introduce a new algorithm motivated by deep model predictive control (MPC) and robust control, safety augmented value estimation from demonstrations (SAVED), which enables efficient learning for sparse cost tasks given a small number of suboptimal demonstrations while satisfying provided constraints. We specifically consider tasks with a tight start state distribution and fixed, known goal set, and only use supervision that indicates task completion. We then show that under mild assumptions and given known stochastic nonlinear dynamics, SAVED has guaranteed iterative improvement in expected performance, extending prior analysis of similar methods for linear dynamics with additive noise [45, 46]. The contributions of this work are (1) a novel method for constrained exploration driven by confidence in task completion, (2) a technique for leveraging model uncertainty to probabilistically enforce complex constraints, enabling obstacle avoidance or optimizing demonstration trajectories while maintaining desired properties, (3) analysis of SAVED which provides iterative improvement guarantees in expected performance for known stochastic nonlinear systems, and (4) experimental evaluation against 3 state-of-the-art model-free and model-based RL baselines on 8 different environments, including simulated experiments and challenging physical maneuvers on the da Vinci surgical robot. Results suggest that SAVED achieves superior sample efficiency, success rate, and constraint satisfaction rate across all domains considered and can be applied efficiently and safely for learning directly on a real robot.

2 Related work

Model-based reinforcement learning: There is significant interest in deep MBRL [13, 15, 16, 26, 27, 32, 36, 55, 57] due to the improvements in sample efficiency when planning over learned dynamics compared to model-free methods for continuous control [18, 20, 30, 47, 48]. However, most prior deep MBRL algorithms use hand-engineered dense cost functions to guide exploration, which we avoid by using demonstrations to provide signal about delayed costs. Additionally, in contrast to prior work, we enable probabilistic enforcement of complex constraints during learning, allowing constrained exploration from successful demonstrations while avoiding unsafe regions. **Reinforcement learning from demonstrations:** Demonstrations have been leveraged to accelerate learning for a variety of model-free RL algorithms, such as Deep Q Learning [4, 21, 33] and DDPG [30, 37, 54]. However, these techniques are applied to model-free RL algorithms and may be inefficient compared to model-based methods. Furthermore, they cannot anticipate constraint violations since they use a reactive policy [14, 52]. Fu et al. [16] use a neural network prior from previous tasks and online adaptation to a new task using iLQR and a dense cost, distinct from the task completion based costs we consider. Finally, Brown et al. [10] use inverse reinforcement learning to significantly outperform suboptimal demonstrations, but do not explicitly optimize for constraint satisfaction or consistent task completion during learning. **Iterative learning control:** In Iterative learning control (ILC), the controller tracks a predefined reference trajectory and data from each iteration is used to improve closed-loop performance [9]. ILC has seen significant success in tasks such as quadrotor and vehicle control [23, 42], and can have provable robustness to external disturbances and model mismatch [11, 25, 31, 51]. Rosolia et al. [44–46] provide a reference-free algorithm to iteratively improve the

performance of an initial trajectory by using a safe set and terminal cost to ensure recursive feasibility, stability, and local optimality given a known, deterministic nonlinear system or stochastic linear dynamics with additive disturbances under mild assumptions. We extend this analysis, and show that given task completion based costs, similar guarantees hold for stochastic nonlinear systems with bounded disturbances satisfying similar assumptions. **Safe reinforcement learning:** There has been significant interest in safe RL [19], typically focusing on exploration while satisfying a set of explicit constraints [1, 29, 35], satisfying specific stability criteria [6], or formulating planning via a risk sensitive Markov decision process [34, 39]. Distinct from prior work in safe RL and control, SAVED can be successfully applied in settings with both highly uncertain dynamics and sparse costs while probabilistically enforcing constraint satisfaction and task completion during learning.

3 Assumptions and preliminaries

In this work, we consider stochastic, unknown dynamical systems with a cost function that only identifies task completion. We assume that (1) tasks are iterative in nature, and thus have a fixed low-variance start state distribution and fixed, known goal set $\mathcal{G}$. This is common in a variety of repetitive tasks, such as assembly, surgical knot tying, and suturing. Additionally, we assume that (2) a modest set of suboptimal but successful demonstration trajectories are available, for example from imprecise human teleoperation or from a hand-tuned PID controller. This enables rough specification of desired behavior without having to design a dense cost function.

Here we outline the framework for MBRL using a standard Markov decision process formulation. A finite-horizon Markov decision process (MDP) is a tuple $(\mathcal{X},\mathcal{U},P(\cdot,\cdot),T,C(\cdot,\cdot))$ where $\mathcal{X}$ is the feasible state space and $\mathcal{U}$ is the action space. The stochastic dynamics model P maps a state and action to a probability distribution over states, T is the task horizon, and C is the cost function. A stochastic control policy π maps an input state to a distribution over $\mathcal{U}$. We assume that the cost function only identifies task completion: $C(x,u) = \mathbb{1}_{\mathcal{G}^C}(x)$, where $\mathcal{G} \subset \mathcal{X}$ defines a goal set in the state space. We define task success by convergence to $\mathcal{G}$ at the end of the task horizon without violating constraints.

4 Safety augmented value estimation from demonstrations (SAVED)

This section describes how SAVED uses a set of suboptimal demonstrations to constrain exploration while satisfying user-specified state space constraints. First, we discuss how SAVED learns system dynamics and a value function to guide learning in sparse cost environments. Then, we motivate and discuss the method used to enforce constraints under uncertainty to both ensure task completion during learning and satisfy user-specified state space constraints.

SAVED optimizes agent trajectories by using MPC to optimize costs over a sequence of actions at each state. However, when using MPC, since the current control is computed by solving a finite-horizon approximation to the infinite-horizon control problem, agents may take shortsighted control actions which may make it impossible to complete the task, such as planning the trajectory of a race car over a short horizon without considering an upcoming curve [7]. Thus, to guide exploration in temporally-extended tasks, we solve the problem in equation 4.0.1a subject to 4.0.1b. This corresponds to the standard objective in MPC with an appended value function V^{π}, which provides a terminal cost estimate for the current policy at the end of the planning horizon. While prior work in deep MBRL [13, 36] has primarily focused only on planning over learned dynamics, we introduce a learned value function, which is initialized from demonstrations to provide initial signal, to guide exploration even in sparse cost settings. The learned dynamics model f_θ and value function V^{π}_{ϕ} are each represented with a probabilistic ensemble of 5 neural networks, as is used to represent system dynamics in Chua et al. [13]. These functions are initialized from demonstrations and updated on each training iteration, and collectively define the current policy $\pi_{\theta,\phi}$. See supplementary material for further details on how these networks are trained.

The core novelties of SAVED are the additional probabilistic constraints in 4.0.1c to encourage task completion driven exploration and enforce user-specified chance constraints. First, a non-parametric density model ρ enforces constrained exploration by requiring x_{t+H} to fall in a region with high probability of task completion. This enforces cost-driven constrained exploration, which enables reliable performance even in sparse cost domains. Second, we require all elements of $x_{t:t+H}$ to fall in

the feasible region $\mathcal{X}$ with probability at least β, which enables probabilistic enforcement of state space constraints. In Section 5.1, we discuss the methods used for task completion driven exploration and in Section 5.2, we discuss how probabilistic constraints are enforced during learning.

$$u^*_{t:t+H-1} = \underset{u_{t:t+H-1} \in \mathcal{U}^H}{\arg\min}\ \mathbb{E}_{x_{t:t+H}}\left[\sum_{i=0}^{H-1} C(x_{t+i}, u_{t+i}) + V^{\pi}_{\phi}(x_{t+H})\right] \tag{4.0.1a}$$

$$\text{s.t. } x_{t+i+1} \sim f_\theta(x_{t+i}, u_{t+i})\ \forall i \in \{0, \ldots, H-1\} \tag{4.0.1b}$$

$$\rho_\alpha(x_{t+H}) > \delta, \mathbb{P}\left(x_{t:t+H} \in \mathcal{X}^{H+1}\right) \geq \beta \tag{4.0.1c}$$

We summarize SAVED in Algorithm 1. At each iteration, we sample a start state and then controls are generated by solving equation 4.0.1 using the cross-entropy method (CEM) [8] at each timestep. Transitions are collected in a replay buffer to update the dynamics, value function, and density model.

Algorithm 1 Safety augmented value estimation from demonstrations (SAVED)

Require: Replay Buffer $\mathcal{R}$; value function $V^{\pi}_{\phi}(x)$, dynamics model $\hat{f}_\theta(x'|x,u)$, and density model $\rho_\alpha(x)$ all seeded with demos; kernel and chance constraint parameters α and β.
for $i \in \{1, \ldots, N\}$ **do**
  Sample x_0 from start state distribution
  for $t \in \{1, \ldots, T-1\}$ **do**
    Pick $u^*_{t:t+H-1}$ by solving equation 4.0.1 using CEM, execute u^*_t, and observe x_{t+1}
    $\mathcal{R} = \mathcal{R} \cup \{(x_t, u^*_t, C(x_t, u^*_t), x_{t+1})\}$
  end for
  if $x_T \in \mathcal{G}$ **then**
    Update density model ρ_α with $x_{0:T}$
  end if
  Optimize θ and ϕ with $\mathcal{R}$
end for

5 Constrained exploration

5.1 Task completion driven exploration

Recent MPC literature [44] motivates constraining exploration to regions in which the agent is confident in task completion, which gives rise to desirable theoretical properties. For a trajectory at iteration k, given by x^k, we define the *sampled safe set* as $\mathcal{SS}^j = \left\{\bigcup_{k \in \mathcal{M}^j} \bigcup_{i=0}^{T} x^k_i\right\}$ where $\mathcal{M}^j$ is the set of indices of all successful trajectories before iteration j as in Rosolia et al. [44]. Thus, $\mathcal{SS}^j$ contains the states from all iterations before j from which the agent controlled the system to $\mathcal{G}$ and is initialized from demonstrations. Under mild assumptions, if states at the end of the MPC planning horizon are constrained to fall in $\mathcal{SS}^j$, iterative improvement, controller feasibility, and convergence are guaranteed given known linear dynamics subject to additive disturbances. [45, 46]. In Section 6, we extend these results to show that, under the same assumptions, we can obtain similar guarantees in expectation for stochastic nonlinear systems if task completion based costs are used. The way we constrain exploration in SAVED builds off of this prior work, but we note that unlike Rosolia et al. [44–46], SAVED is designed for settings in which dynamics are completely unknown.

We develop a method to approximately implement the above constraint with a continuous approximation to $\mathcal{SS}^j$ using non-parametric density estimation, allowing SAVED to scale to more complex settings than prior work using similar cost-driven exploration techniques [44–46]. Since $\mathcal{SS}^j$ is a discrete set, we introduce a new continuous approximation by fitting a density model ρ to $\mathcal{SS}^j$ and constraining $\rho_\alpha(x_{t+H}) > \delta$, where α is a kernel width parameter (constraint 4.0.1c). Since the tasks considered in this work have sufficiently low (< 17) state space dimension, we find kernel density estimation provides a reasonable approximation. We implement a tophat kernel density model using a nearest neighbors classifier with a tuned kernel width α and use $\delta = 0$ for all experiments. Thus, all states within Euclidean distance α from the closest state in $\mathcal{SS}^j$ are considered safe under ρ_α and represent states in which the agent has high confidence in task completion. As the policy improves, it may forget how to complete the task from very old states in $\mathcal{SS}^j$, so very old states are evicted from

$\mathcal{SS}^j$ to reflect the current policy when fitting ρ_α. We discuss how these constraints are implemented in Section 5.2, with further details in the supplementary material. In future work, we will investigate implicit density estimation techniques such as [5, 17, 50] to scale to high-dimensional settings.

5.2 Probabilistic constraint enforcement

SAVED leverages uncertainty estimates in the learned dynamics to enforce probabilistic constraints on its trajectories. This allows SAVED to handle complex, user-specified state space constraints to avoid obstacles or maintain certain properties of demonstrations without a user-shaped or time-varying cost function. We do this by sampling sequences of actions from a truncated Gaussian distribution that is iteratively updated using the cross-entropy method (CEM) [13]. Each action sequence is simulated multiple times over the stochastic dynamics model as in [13] and the average return of the simulations is used to score the sequence. However, unlike Chua et al. [13], we implement chance constraints by discarding actions sequences if more than $100 \cdot (1-\beta)\%$ of the simulations violate constraints (constraint 4.0.1c), where β is a user-specified tolerance. This is illustrated in Figure 1b. The task completion constraint (Section 5.1) is implemented similarly, with action sequences discarded if any of the simulated rollouts do not terminate in a state with sufficient density under ρ_α.

6 Theoretical analysis of SAVED

In prior work, it has been shown that under mild assumptions, the *sampled safe set* $\mathcal{SS}^j$ and the associated value function can be used to design a controller which guarantees constraint satisfaction, convergence to $\mathcal{G}$, and iterative improvement [45]. This analysis specifically assumes known linear dynamics with bounded disturbances, that the limit of infinite data is used for policy evaluation at each iteration, and that the MPC objective can be solved exactly [45, 46]. We extend this analysis by showing that under the same assumptions, if task completion based costs (as defined in Section 3) are used and $\beta = 1$, then the same guarantees can be shown in expectation for SAVED for stochastic nonlinear systems. Define the closed-loop system with the policy defined by SAVED at iteration j as $x_{t+1}^j = f(x_t^j; \pi^j(x_t^j); w_t^j)$ for bounded disturbances $w_t \in \mathcal{W}$ and the expected cost-to-go as $J^{\pi^j}(x_0) = \mathbb{E}[\sum_{t=0}^{\infty} C(x_t^j, \pi^j(x_t^j))]$. In the supplementary material, we formally prove the following:

1. ***Recursive Feasibility:*** The trajectory generated by the closed-loop system at iteration j satisfies problem constraints. Equivalently stated, $\forall i \in \mathbb{N},\ x_i^j \in \mathcal{X}$ (see Lemma 1).
2. ***Convergence in Probability:*** If the closed-loop system converges in probability to $\mathcal{G}$ at the initial iteration, then it converges in probability at all subsequent iterations: $\lim_{t\to\infty} P(x_t^j \notin \mathcal{G}) = 0$ (see Lemma 2).
3. ***Iterative Improvement:*** The expected cost-to-go for SAVED is non-increasing: $\forall j \in \mathbb{N},\ J^{\pi^j}(x_0) \geq J^{\pi^{j+1}}(x_0)$ (see Theorem 1).

7 Experiments

We evaluate SAVED on simulated continuous control benchmarks and on real robotic tasks with the da Vinci Research Kit (dVRK) [24] against state-of-the-art deep RL algorithms and demonstrate that SAVED outperforms all baselines in terms of sample efficiency, success rate, and constraint satisfaction during learning. All tasks use $C(x,u) = \mathbb{1}_{\mathcal{G}^C}(x)$ (Section 3), which is equivalent to the time spent outside the goal set. All algorithms are given the same demonstrations, are evaluated on iteration cost, success rate, and constraint satisfaction rate (if applicable), and run 3 times to control for stochasticity in training. Tasks are only considered successfully completed if the agent reaches and stays in $\mathcal{G}$ until the end of the episode without ever violating constraints. For all simulated tasks, we give model-free methods 10,000 iterations since they take much longer to converge but sometimes have better asymptotic performance. See supplementary material for videos, and ablations with respect to choice of α, β, and demonstration quantity. We also include further details on baselines, network architectures, hyperparameters, and training procedures.

7.1 Baselines

We consider the following set of model-free and model-based baseline algorithms. To enforce constraints for model-based baselines, we augment the algorithms with the simulation based method

described in Section 5.2. Because model-free baselines have no such mechanism to readily enforce constraints, we instead apply a very negative reward when constraints are violated. See supplementary material for an ablation of the reward function used for model-free baselines.

1. **Behavior cloning (Clone)**: Supervised learning on demonstrator trajectories.
2. **PETS from demonstrations (PETSfD)**: Probabilistic ensemble trajectory sampling (PETS) from Chua et al [13] with the dynamics model initialized with demo trajectories and planning horizon long enough to plan to the goal (judged by best performance of SAVED).
3. **PETSfD Dense**: PETSfD with access to hand-engineered dense cost.
4. **Soft actor critic from demonstrations (SACfD)**: Model-free algorithm, Soft Actor Critic (SAC)[20], where only demo transitions are used for training on the first iteration.
5. **Overcoming exploration in reinforcement learning from demonstrations (OEFD)**: Model-free algorithm from Nair et al. [37] which uses DDPG [30] with Hindsight Experience Replay (HER)[3] and a behavior cloning loss to accelerate RL with demonstrations.
6. **SAVED (No SS)**: SAVED without the *sampled safe set* constraint described in Section 5.1.

7.2 Simulated navigation

To demonstrate if SAVED can efficiently and safely learn temporally extended tasks with complex constraints, we consider a set of tasks in which a point mass navigates to a unit ball centered at the origin. The agent can exert force in cardinal directions and experiences drag and Gaussian process noise in the dynamics. For each task, we supply 50-100 suboptimal demonstrations generated by running LQR along a hand-tuned safe trajectory. SAVED has a higher success rate than all other RL baselines using sparse costs, even including model-free baselines over the first 10,000 iterations, while never violating constraints across all navigation tasks. Only Clone and PETSfD Dense ever achieve a higher success rate, but Clone but does not improve upon demonstration performance (Figure 2) and PETSfD Dense has additional information about the task. Furthermore, SAVED learns significantly more efficiently than all RL baselines on all navigation tasks except for tasks 1 and 3, in which PETSfD Dense with a Euclidean norm cost function finds a better solution. While SAVED (No SS) can complete the tasks, it has a much lower success rate than SAVED, especially in environments with obstacles as expected, demonstrating the importance of the *sampled safe set* constraint. Note that SACfD, OEFD, and PETSfD make essentially no progress in the first 100 iterations and never complete any of the tasks in this time, although they mostly satisfy constraints.

7.3 Simulated robot experiments

To evaluate whether SAVED outperforms baselines even on standard unconstrained environments, we consider sparse versions of two common simulated robot tasks: the PR2 Reacher environment used in Chua et al. [13] with a fixed goal and on a pick and place task with a simulated, position-controlled Fetch robot [40, 56]. The reacher task involves controlling the end-effector of a simulated PR2 robot to a small ball in $\mathbb{R}^3$. The pick and place task involves picking up a block from a fixed location on a table and also guiding it to a small ball in $\mathbb{R}^3$. The task is simplified by automating the gripper motion, which is difficult for SAVED to learn due to bimodality of gripper controls, which is hard to capture with the unimodal truncated Gaussian distribution used during CEM sampling. SAVED still learns faster than all baselines on both tasks (Figure 3) and exhibits significantly more stable learning in the first 100 and 250 iterations for the reacher and pick and place tasks respectively.

8 Physical robot experiments

We evaluate the ability of SAVED to learn temporally-extended trajectory optimization tasks with nonconvex state space constraints on the da Vinci Research Kit (dVRK) [24]. The dVRK is cable-driven and has relatively imprecise controls, motivating model learning [49]. Furthermore, safety is paramount due to the cost and delicate structure of the arms. The goal here is to speed up demo trajectories by constraining learned trajectories to fall within a tight, 1 cm tube of the demos, making this difficult for many RL algorithms. Additionally, robot experiments are very time consuming, so training RL algorithms on limited physical hardware is difficult without sample efficient algorithms.

Figure-8: Here, the agent tracks a figure-8 in state space. However, because there is an intersection in the middle of the desired trajectory, SAVED finds a shortcut to the goal state. Thus, the trajectory

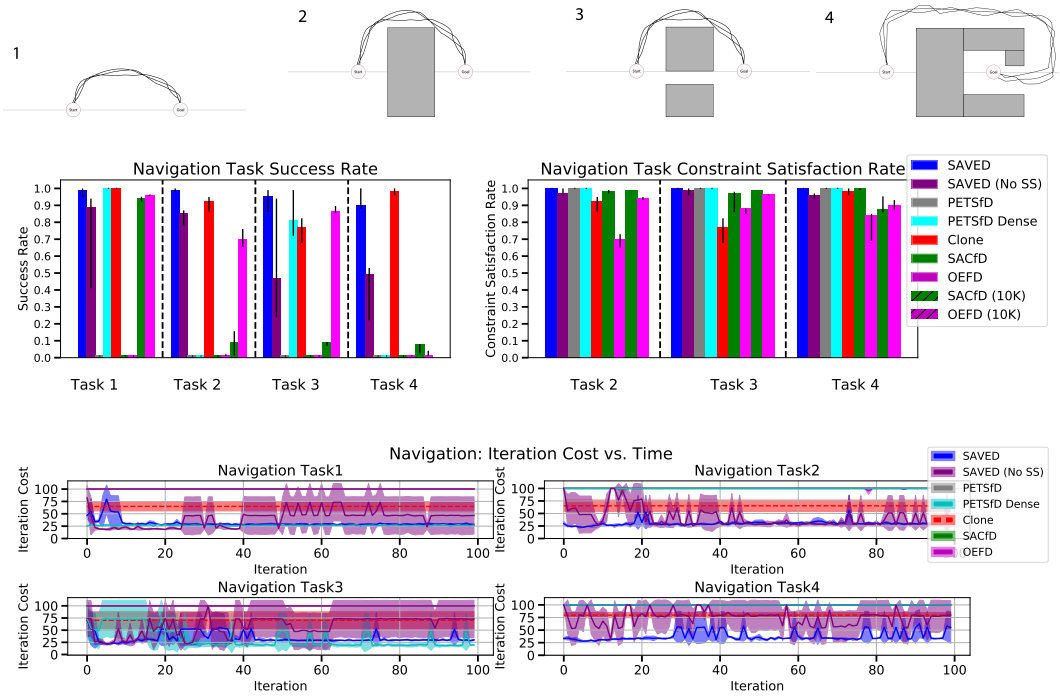


Figure 2: **Navigation Domains:** SAVED is evaluated on 4 navigation tasks. Tasks 2-4 contain obstacles, and task 3 contains a channel for passage to $\mathcal{G}$ near the x-axis. SAVED learns significantly faster than all RL baselines on tasks 2 and 4. In tasks 1 and 3, SAVED has lower iteration cost than baselines using sparse costs, but does worse than PETSfD Dense, which is given dense Euclidean norm costs to find the shortest path to the goal. For each task and algorithm, we report success and constraint satisfaction rates over the first 100 training iterations and also over the first 10,000 iterations for SACfD and OEFD. We observe that SAVED has higher success and constraint satisfaction rates than other RL algorithms using sparse costs across all tasks, and even achieves higher rates in the first 100 training iterations than model-free algorithms over the first 10,000 iterations.

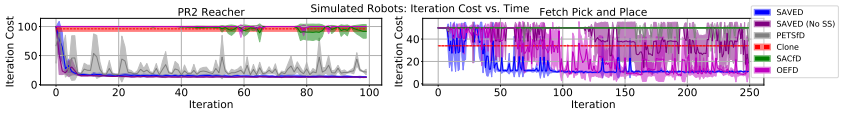


Figure 3: **Simulated Robot Experiments Performance:** SAVED achieves better performance than all baselines on both tasks. We use 20 demonstrations with average iteration cost of 94.6 for the reacher task and 100 demonstrations with average iteration cost of 34.4 for the pick and place task. For the reacher task, the safe set constraint does not improve performance, likely because the task is very simple, but for pick and place, we see that the safe set constraint adds significant training stability.

is divided into non-intersecting segments before SAVED separately optimizes each one. At execution-time, the segments are stitched together and we find that SAVED is robust enough to handle the uncertainty at the transition point. We hypothesize that this is because the dynamics and value function exhibit good generalization. SAVED quickly learns to smooth out demo trajectories while satisfying constraints with a success rate of over 80% while baselines violate constraints on nearly every iteration and never complete the task, as shown in Figure 4. Note that PETSfD almost always violates constraints, even though it enforces constraints in the same way as SAVED. We hypothesize that since we need to give PETSfD a long planning horizon to make it possible to complete the task (since it has no value function), this makes it unlikely that a constraint satisfying trajectory is sampled with CEM. See supplementary material for the other segment and the full combined trajectory.

Surgical knot tying: SAVED is used to optimize demonstrations of a surgical knot tying task on the dVRK, using the same multilateral motion as in [53]. Demonstrations are hand-engineered for the task, and then policies are optimized for one arm (arm 1), while a hand-engineered policy is used for the other arm (arm 2). We do this because while arm 1 wraps the thread around arm 2, arm 2 simply moves down, grasps the other end of the thread, and pulls it out of the phantom as shown in Figure 5. Thus, we only expect significant performance gain by optimizing the policy for the portion of the arm 1 trajectory which involves wrapping the thread around arm 2. We only model the motion of the end-effectors in 3D space. SAVED quickly learns to smooth out demo trajectories, with a success

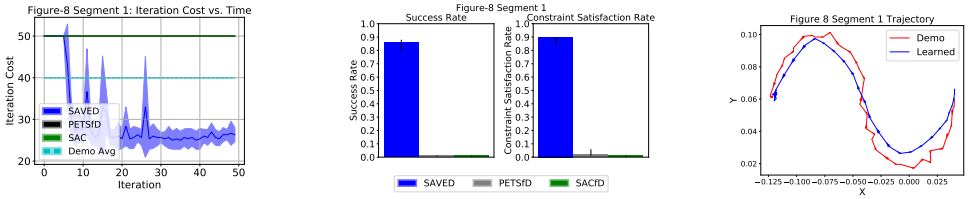


Figure 4: **Figure-8: Training Performance:** After just 10 iterations, SAVED consistently succeeds and converges to an iteration cost of 26, faster than demos which took an average of 40 steps. Neither baseline ever completes the task in the first 50 iterations; **Trajectories:** Demo trajectories satisfy constraints, but are noisy and inefficient. SAVED learns to speed up with only occasional constraint violations and stabilizes in the goal set.

rate of over 75% (Figure 5) during training, while baselines are unable to make sufficient progress in this time. PETSfD rarely violates constraints, but also almost never succeeds, while SACfD almost always violates constraints and never completes the task. Training SAVED directly on the real robot for 50 iterations takes only about an hour, making it practical to train on a real robot for tasks where data collection is expensive. At execution-time, we find that SAVED is very consistent, successfully tying a knot in 20/20 trials with average iteration cost of 21.9 and maximum iteration cost of 25 for the arm 1 learned policy, significantly more efficient than demos which have an average iteration cost of 34. See supplementary material for trajectory plots of the full knot tying trajectory.

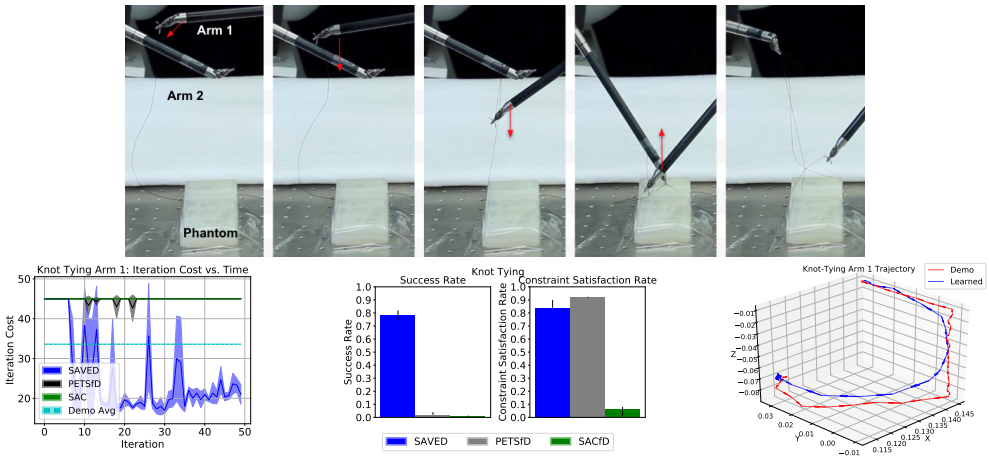


Figure 5: **Surgical Knot Tying: Knot tying motion:** Arm 1 wraps the thread around arm 2, which grasps the other end of the thread and tightens the knot; **Training Performance:** After just 15 iterations, the agent completes the task relatively consistently with only a few failures, and converges to a iteration cost of 22, faster than demos, which have an average iteration cost of 34. In the first 50 iterations, both baselines mostly fail, and are less efficient than demos when they do succeed; **Trajectories:** SAVED quickly learns to speed up with only occasional constraint violations and stabilizes in the goal set.

9 Discussion and future work

This work presents SAVED, a model-based RL algorithm that can efficiently learn a variety of robotic control tasks in the presence of dynamical uncertainties, sparse cost feedback, and complex constraints. SAVED uses a small set of suboptimal demonstrations and a learned state-value function to guide learning with a novel method to constrain exploration to regions in which the agent is confident in task completion. We present iterative improvement guarantees in expectation for SAVED for stochastic nonlinear systems, extending prior work providing similar guarantees for stochastic linear systems. We then demonstrate that SAVED can handle complex state space constraints under uncertainty. We empirically evaluate SAVED on 6 simulated benchmarks and 2 complex maneuvers on a real surgical robot. Results suggest that SAVED is more sample efficient and has higher success and constraint satisfaction rates than all RL baselines and can be efficiently and safely trained on a real robot. We believe this work opens up opportunities to further study probabilistically safe RL, and we are particularly interested in exploring how these ideas can be extended to image space planning and multi-goal settings in future work.

References

[1] Joshua Achiam et al. "Constrained policy optimization". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 22–31.

[2] Dario Amodei et al. "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[3] Marcin Andrychowicz et al. "Hindsight Experience Replay". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5048–5058.

[4] Yusuf Aytar et al. "Playing hard exploration games by watching youtube". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2935–2945.

[5] Marc Bellemare et al. "Unifying count-based exploration and intrinsic motivation". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1471–1479.

[6] Felix Berkenkamp et al. "Safe Model-based Reinforcement Learning with Stability Guarantees". In: *NIPS*. 2017.

[7] Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive control for linear and hybrid systems*. Cambridge University Press, 2017.

[8] Zdravko I. Botev et al. *The Cross-Entropy Method for Optimization*.

[9] Douglas A Bristow, Marina Tharayil, and Andrew G Alleyne. "A survey of iterative learning control". In: *IEEE control systems magazine* 26.3 (2006), pp. 96–114.

[10] Daniel S. Brown et al. "Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations". In: abs/1904.06387 (2019). arXiv: 1904.06387.

[11] Insik Chin et al. "A two-stage iterative learning control technique combined with real-time feedback for independent disturbance rejection". In: *Automatica* 40.11 (2004), pp. 1913–1922.

[12] Kurtland Chua. *Experiment code for "Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models"*. https://github.com/kchua/handful-of-trials. 2018.

[13] Kurtland Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models". In: *Advances in Neural Information Processing Systems*. 2018, pp. 4754–4765.

[14] Sarah Dean et al. "On the sample complexity of the linear quadratic regulator". In: *Foundations of Computational Mathematics (FoCM) 2019*. 2019.

[15] MP. Deisenroth and CE. Rasmussen. "PILCO: A Model-Based and Data-Efficient Approach to Policy Search". In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*. Omnipress, 2011, pp. 465–472.

[16] Justin Fu, Sergey Levine, and Pieter Abbeel. "One-shot learning of manipulation skills with online dynamics adaptation and neural network priors". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 4019–4026.

[17] Justin Fu, John Co-Reyes, and Sergey Levine. "EX2: Exploration with exemplar models for deep reinforcement learning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2577–2587.

[18] Scott Fujimoto, Herke van Hoof, and Dave Meger. "Addressing Function Approximation Error in Actor-Critic Methods". In: *ICML*. 2018.

[19] Javier García and Fernando Fernández. "A Comprehensive Survey on Safe Reinforcement Learning". In: *J. Mach. Learn. Res.* 16.1 (Jan. 2015), pp. 1437–1480.

[20] Tuomas Haarnoja et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. 2018, pp. 1856–1865.

[21] Todd Hester et al. "Deep q-learning from demonstrations". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[22] Rishabh Jangir. *Overcoming-exploration-from-demos*. https://github.com/jangirrishabh/Overcoming-exploration-from-demos. 2018.

[23] Nitin R Kapania and J Christian Gerdes. "Design of a feedback-feedforward steering controller for accurate path tracking and stability at the limits of handling". In: *Vehicle System Dynamics* 53.12 (2015), pp. 1687–1704.

[24] Peter Kazanzides et al. "An Open-Source Research Kit for the da Vinci Surgical System". In: *IEEE Intl. Conf. on Robotics and Auto. (ICRA)*. Hong Kong, China, June 1, 2014, pp. 6434–6439.

[25] Jay H Lee and Kwang S Lee. "Iterative learning control applied to batch processes: An overview". In: *Control Engineering Practice* 15.10 (2007), pp. 1306–1318.

[26] Ian Lenz, Ross A. Knepper, and Ashutosh Saxena. "DeepMPC: Learning Deep Latent Features for Model Predictive Control". In: *Robotics: Science and Systems*. 2015.

[27] Sergey Levine, Nolan Wagener, and Pieter Abbeel. "Learning contact-rich manipulation skills with guided policy search". In: *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2015, pp. 156–163.

[28] Sergey Levine et al. “End-to-end training of deep visuomotor policies”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[29] Z. Li, U. Kalabić, and T. Chu. “Safe Reinforcement Learning: Learning with Supervision Using a Constraint-Admissible Set”. In: *2018 Annual American Control Conference (ACC)*. June 2018, pp. 6390–6395.

[30] Timothy P. Lillicrap et al. “Continuous control with deep reinforcement learning”. In: *CoRR* abs/1509.02971 (2015). arXiv: 1509.02971.

[31] Chung-Yen Lin, Liting Sun, and Masayoshi Tomizuka. “Robust principal component analysis for iterative learning control of precision motion systems with non-repetitive disturbances”. In: *2015 American Control Conference (ACC)*. IEEE. 2015, pp. 2819–2824.

[32] Kendall Lowrey et al. “Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control”. In: *International Conference on Learning Representations*. 2019.

[33] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), p. 529.

[34] Teodor M. Moldovan and Pieter Abbeel. “Risk Aversion in Markov Decision Processes via Near Optimal Chernoff Bounds”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 3131–3139.

[35] Teodor Mihai Moldovan and Pieter Abbeel. “Safe exploration in Markov decision processes”. In: *arXiv preprint arXiv:1205.4810* (2012).

[36] Anusha Nagabandi et al. “Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning”. In: May 2018, pp. 7559–7566.

[37] Ashvin Nair et al. “Overcoming Exploration in Reinforcement Learning with Demonstrations”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), pp. 6292–6299.

[38] Arkadi Nemirovski. “On safe tractable approximations of chance constraints”. In: *European Journal of Operational Research* 219.3 (2012), pp. 707–718.

[39] Takayuki Osogami. “Robustness and risk-sensitivity in Markov decision processes”. In: *Advances in Neural Information Processing Systems* 1 (Jan. 2012), pp. 233–241.

[40] Matthias Plappert et al. “Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research”. In: *CoRR* abs/1802.09464 (2018). arXiv: 1802.09464.

[41] Vitchyr Pong. *rlkit*. https://github.com/vitchyr/rlkit. 2018-2019.

[42] Oliver Purwin and Raffaello D’Andrea. “Performing aggressive maneuvers using iterative learning control”. In: *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, pp. 1731–1736.

[43] U. Rosolia, A. Carvalho, and F. Borrelli. “Autonomous Racing using Learning Model Predictive Control”. In: *Proceedings 2017 IFAC World Congress*. 2017.

[44] Ugo Rosolia and Francesco Borrelli. “Learning Model Predictive Control for Iterative Tasks. A Data-Driven Control Framework”. In: *IEEE Transactions on Automatic Control* 63.7 (July 2018), pp. 1883–1896.

[45] Ugo Rosolia and Francesco Borrelli. “Sample-Based Learning Model Predictive Control for Linear Uncertain Systems”. In: *CoRR* abs/1904.06432 (2019).

[46] Ugo Rosolia, Xiaojing Zhang, and Francesco Borrelli. “A Stochastic MPC Approach with Application to Iterative Learning”. In: *2018 IEEE Conference on Decision and Control (CDC)* (2018), pp. 5152–5157.

[47] John Schulman et al. “Trust Region Policy Optimization”. In: Proceedings of Machine Learning Research 37 (July 2015). Ed. by Francis Bach and David Blei, pp. 1889–1897.

[48] John Schulman et al. “Proximal Policy Optimization Algorithms”. In: *CoRR* abs/1707.06347 (2017). arXiv: 1707.06347.

[49] D. Seita et al. “Fast and Reliable Autonomous Surgical Debridement with Cable-Driven Robots Using a Two-Phase Calibration Procedure”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. May 2018, pp. 6651–6658.

[50] Haoran Tang et al. “#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 2753–2762.

[51] Masayoshi Tomizuka. “Dealing with periodic disturbances in controls of mechanical systems”. In: *Annual Reviews in Control* 32.2 (2008), pp. 193–199.

[52] Stephen Tu and Benjamin Recht. “The Gap Between Model-Based and Model-Free Methods on the Linear Quadratic Regulator: An Asymptotic Viewpoint”. In: *CoRR* abs/1812.03565 (2018). arXiv: 1812.03565.

[53] Jur Van Den Berg et al. “Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations”. In: *2010 IEEE International Conference on Robotics and Automation*. IEEE. 2010, pp. 2074–2081.

[54] Matej Vecerik et al. “Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards”. In: *CoRR* abs/1707.08817 (2017).

[55] Grady Williams et al. “Information theoretic MPC for model-based reinforcement learning”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)* (2017), pp. 1714–1721.

[56] Melonee Wise et al. “Fetch and freight: Standard platforms for service robot applications”. In: *Workshop on Autonomous Mobile Service Robots*. 2016.

[57] Marvin Zhang et al. “SOLAR: Deep Structured Latent Representations for Model-Based Reinforcement Learning”. In: *ICML*. 2019.

10 Theoretical Results

10.1 Definitions

Consider the system

$$x_{t+1}^j = f(x_t^j, u_t^j, w_t^j) \tag{10.1.1}$$

where the state $x_t^j \in \mathbb{R}^n$, the input $u_t^j \in \mathbb{R}^d$ and the disturbance $w_t^j \in \mathcal{W}$.

We define the *sampled safe set* as:

$$\mathcal{SS}^j = \left\{ \bigcup_{k \in \mathcal{M}^j} \bigcup_{i=0}^{T} x_i^k \right\} \tag{10.1.2}$$

where $\mathcal{M}^j$ is the set of indices of all successful trajectories up to iteration j as in Rosolia et al. [44].

Recursively define the value function of π^j (SAVED at iteration j) in closed-loop with (10.1.1) as:

$$V^{\pi^j}(x) = \begin{cases} \mathbb{E}_w \left[C(x, \pi^j(x)) + V^{\pi^j}(f(x, \pi^j(x), w)) \right] & x \in \mathcal{SS}^j \\ +\infty & x \notin \mathcal{SS}^j \end{cases}$$

The solution to these modified Bellman equations computes the state-value function for the infinite task horizon controller. In practice, we train the value function using the standard TD-1 error corresponding to the standard Bellman equations. The penalty for states leaving the *sampled safe set* is approximated by the *sampled safe set* constraint described in the main paper. In future work, we will investigate using this information recursively to train the value function to capture this as well. We assume that we can solve the system of equations defining V^{π^j}.

Finally we define:

$$J_{t \to t+H}^j(x_t^j) = \min_{\pi_{t:t+H-1|t}} \mathbb{E}_{x_{t:t+H|t}^j} \left[\sum_{k=t}^{t+H-1} C(x_{k|t}^j, \pi_{k|t}(x_{k|t}^j)) + V^{\pi^{j-1}}(x_{t+H|t}^j) \right] \tag{10.1.3}$$

$$= \mathbb{E}_{x_{t:t+H|t}^j} \left[\sum_{k=t}^{t+H-1} C(x_{k|t}^j, \pi_{k|t}^*(x_{k|t}^j)) + V^{\pi^{j-1}}(x_{t+H|t}^j) \right] \tag{10.1.4}$$

$$J^{\pi_j}(x_0) = \mathbb{E}_{x^j} \left[\sum_{t=0}^{\infty} C(x_t^j, \pi_j(x_t^j)) \right] = V^{\pi^j}(x_0) \tag{10.1.5}$$

Equation 10.1.4 is the value of equation 10.1.3, which is the MPC cost-to-go at time t during iteration j. Equation 10.1.5 is the expected performance of the policy at iteration j.

10.2 Assumptions

1. **Known dynamics with disturbances:** The dynamics in (10.1.1) are known with disturbance realizations $w \in \mathcal{W}$ where the support of $\mathcal{W}$ is bounded.
2. **Optimization over the set of causal feedback policies:** We assume that at each timestep, we optimize over the set of causal feedback policies Π and not over individual controls. In SAVED, we optimize over the set of constant feedback policies $\mathcal{U} \subseteq \Pi$. We set $\beta = 1$ (robust constraints) and exactly constrain terminal states to robustly fall within the *sampled safe set*. Specifically, the optimization problem at time t of iteration j is

$$\begin{aligned} \pi_{t:t+H-1|t}^{*,j} = \underset{\pi_{t:t+H-1|t} \in \Pi^H}{\arg\min} \; & \mathbb{E}_{x_{t:t+H|t}^j} \left[\sum_{i=0}^{H-1} C(x_{t+i|t}^j, \pi_{t+i|t}(x_{t+i|t}^j)) + V^{\pi^{j-1}}(x_{t+H|t}^j) \right] \\ \text{s.t. } & x_{t+i+1|t}^j = f(x_{t+i|t}^j, \pi_{t+i|t}(x_{t+i}^j), w_{t+i}) \; \forall i \in \{0, \ldots, H-1\} \\ & x_{t+H|t}^j \in \mathcal{SS}^j, \; \forall w_t \in \mathcal{W} \\ & x_{t:t+H|t}^j \in \mathcal{X}^{H+1}, \; \forall w_t \in \mathcal{W} \end{aligned} \tag{10.2.6}$$

π^j is the policy (SAVED) at iteration j, where

$$u_t^j = \pi^j(x_t^j) = \pi_{t|t}^{*,j}(x_t^j) \tag{10.2.7}$$

is the control applied at state x_t. $J_{t\to t+H}^j(x_t^j)$ is defined as the value of 10.2.6. We assume we can solve this problem at each timestep. We also assume that we can exactly compute V^{π^j}.

3. **Robust Control Invariant Goal Set:** $\mathcal{G}$ is a robust control invariant set with respect to the dynamics and policy class as defined above. This means that from any state in $\mathcal{G}$, there exists a policy that keeps the system in $\mathcal{G}$ for all possible disturbance realizations.

4. **Robust Control Invariant Sampled Safe Set:** We assume that $\mathcal{SS}^j$ is a robust control invariant set with respect to the dynamics and policy class for all j. This is a strong assumption, but it can be shown in the limit of infinite samples from the control policy at each iteration, the sampled safe $\mathcal{SS}^j$ is robust control invariant [45]. Finally, if $\mathcal{SS}^j$ is a robust control invariant set and since $x_0 \in \mathcal{SS}^j$, $J_{0\to H}^j(x_0^j) < \infty$. This also implies that the demonstrations provide enough information to construct $\mathcal{SS}^1$ such that it is robust control invariant. The amount of data needed to approximately meet this assumption is related to the stochasticity of the environment.

5. **Constant Start State:** The start state x_0 is constant across iterations. This assumption is reasonable in the setting considered in this paper, since in all experiments, the start state distribution has low variance.

6. **Completion Cost Specification:** $\exists \varepsilon > 0$ s.t. $C(x,\cdot) \geq \varepsilon \mathbb{1}_{\mathcal{G}^C}(x)$ and $\forall x \in \mathcal{G}\ C(x,\cdot) = 0$. This is true in the experiments we consider in this paper, where all costs are specified as above with $\varepsilon = 1$.

10.3 Proofs of theoretical analysis

We will show that under the set of strong assumptions above, which are similar to those use in [45], the control strategy presented in this paper guarantees iterative improvement of expected performance. We proceed similarly to Rosolia et al. [45].

Lemma 1. ***Recursive Feasibility:*** *Consider system* (10.1.1) *in closed-loop with* (10.2.7). *Let the sampled safe set* $\mathcal{SS}^j$ *be defined as in* (10.1.2). *Let assumptions* (1)-(6) *hold, then the controller* (10.2.6) *and* (10.2.7) *is feasible for* $t \geq 0$ *and* $j \geq 0$. *Equivalently stated,* $\forall i \in \mathbb{N}$, $x_i^j \in \mathcal{X}$.

Proof of Lemma 1:
We proceed by induction. By assumption 4, $J_{0\to H}^j(x_0^j) < \infty$. Let $J_{t\to t+H}^j(x_t^j) < \infty$ for some $t \in \mathbb{N}$. Conditioning on the random variable x_t^j:

$$J^j_{t\to t+H}(x^j_t) = \mathbb{E}_{x^j_{t+1:t+H|t}}\left[\sum_{k=0}^{H-1} C(x^j_{t+k|t}, \pi^{*,j}_{t+k|t}(x^j_{t+k|t})) + V^{\pi^{j-1}}(x^j_{t+H|t})\right] \tag{10.3.8}$$

$$= C(x^j_t, \pi^{*}_{t|t}(x^j_t)) + \mathbb{E}_{x^j_{t+1:t+H|t}}\left[\sum_{k=1}^{H-1} C(x^j_{t+k|t}, \pi^{*,j}_{t+k|t}(x^j_{t+k|t})) + V^{\pi^{j-1}}(x^j_{t+H|t})\right] \tag{10.3.9}$$

$$= C(x^j_t, \pi^{*,j}_{t|t}(x^j_t)) +$$
$$\mathbb{E}_{x^j_{t+1:t+H+1|t}}\left[\sum_{k=1}^{H-1} C(x^j_{t+k|t}, \pi^{*,j}_{t+k|t}(x^j_{t+k|t})) + C(x^j_{t+H|t}, \pi^{j-1}(x^j_{t+H|t})) + V^{\pi^{j-1}}(x^j_{t+H+1|t})\right] \tag{10.3.10}$$

$$\geq C(x^j_t, \pi^{*,j}_{t|t}(x^j_t)) +$$
$$\mathbb{E}_{x^j_{t+1}}\left[\min_{\pi_{t+1:t+H|t+1}} \mathbb{E}_{x^j_{t+2:t+H+1|t+1}}\left[\sum_{k=1}^{H-1} C(x^j_{t+k|t}, \pi_{t+k|t+1}(x^j_{t+k|t+1})) + \right.\right. \tag{10.3.11}$$
$$\left.\left. C(x^j_{t+H|t+1}, \pi_{t+H|t+1}(x^j_{t+H|t+1})) + V^{\pi^{j-1}}(x^j_{t+H+1|t+1})\right]\right]$$

$$= C(x^j_t, \pi^j(x^j_t)) + \mathbb{E}_{x^j_{t+1}}\left[J^j_{1\to t+H+1}(x^j_{t+1})\right] \tag{10.3.12}$$

Equation 10.3.8 follows from the definition in 10.1.4, equation 10.3.10 follows from the definition of $V^{\pi_{j-1}}$. The inner expectation in equation 10.3.11 conditions on the random variable x^j_{t+1}, and the outer expectation integrates it out. The inequality in 10.3.11 follows from the fact that $[\pi^{*,j}_{t+1|t}, \ldots, \pi^{*,j}_{t+H-1|t}, \pi^{j-1}]$ is a possible solution to 10.3.11. Equation 10.3.12 follows from the definition in equation 10.1.4. By induction, $\mathbb{E}[J^j_{t\to t+H}(x^j_t)] < \infty \; \forall t \in \mathbb{N}$. Therefore, the controller is feasible at iteration j. □

Lemma 2. ***Convergence in Probability:*** *Consider the closed-loop system* (10.1.1) *and* (10.2.7)*. Let the sampled safe set $\mathcal{SS}^j$ be defined as in* (10.1.2)*. Let assumptions* (1)-(6) *hold. If the closed-loop system converges in probability to $\mathcal{G}$ at the initial iteration, then it converges in probability at all subsequent iterations. Precisely, at iteration j:* $\lim_{t\to\infty} P(x^j_t \notin \mathcal{G}) = 0$

Proof of Lemma 2:
By Lemma 1, assuming a cost satisfying assumption 6, $\forall L \in \mathbb{N}$,

$$\mathbb{E}_{x^j_{1:L}}\left[\sum_{k=0}^{L} C(x^j_k, \pi^j(x^j_k)) + J^j_{L\to L+H}(x^j_L)\right] \leq J^j_{0\to H}(x^j_0) \tag{10.3.13}$$

$$\implies \mathbb{E}_{x^j_L}\left[J^j_{L\to L+H}(x^j_L)\right] \leq J^j_{0\to H}(x^j_0) - \mathbb{E}_{x^j_{1:L}}\left[\sum_{k=0}^{L} C(x^j_k, \pi^j(x^j_k))\right] \leq J^j_{0\to H}(x^j_0) - \varepsilon \sum_{k=0}^{L} P(x^j_k \notin \mathcal{G}) \tag{10.3.14}$$

Line 10.3.14 follows from rearranging 10.3.13 and applying assumption 6. Because $\mathcal{G}$ is robust control invariant by assumption 3, $\{P(x^j_k \notin \mathcal{G})\}_{k=0}^{\infty}$ is a non-increasing sequence. Suppose $\lim_{k\to\infty} P(x^j_k \notin \mathcal{G}) = \delta > 0$ (the limit must exist by the Monotone Convergence Theorem). Then $\exists L \in \mathbb{N}$, s.t. $\forall l > L$, $P(x^j_l \notin \mathcal{G}) > \delta/2$. By the Archimedean principle, the RHS of 10.3.14 can be driven arbitrarily negative, which is impossible. By contradiction, $\lim_{k\to\infty} P(x^j_k \notin \mathcal{G}) = 0$. □

Theorem 1. ***Iterative Improvement:*** *Consider system* (10.1.1) *in closed-loop with* (10.2.7)*. Let the sampled safe set $\mathcal{SS}^j$ be defined as in* (10.1.2)*. Let assumptions* (1)-(6) *hold, then the expected cost-to-go (10.1.5) associated with the control policy* (10.2.7) *is non-increasing in iterations. More*

formally:

$$\forall j \in \mathbb{N},\ J^{\pi^j}(x_0) \geq J^{\pi^{j+1}}(x_0)$$

Furthermore, $\{J^{\pi^j}(x_0)\}_{j=0}^{\infty}$ *is a convergent sequence.*

Proof of Theorem 1:

Let $j \in \mathbb{N}$

$$J^j_{0\to H}(x_0) \geq C(x_0,u_0) + \mathbb{E}_{x^j_1}\left[J^j_{1\to H+1}(x^j_1)\right] \tag{10.3.15}$$

$$\geq \mathbb{E}_{x^j}\left[\sum_{t=0}^{\infty} C(x^j_t,\pi^j(x^j_t))\right] + \lim_{t\to\infty}\mathbb{E}_{x^j_t}\left[J^j_{t\to t+H}(x^j_t)\right] \tag{10.3.16}$$

$$= \mathbb{E}_{x^j}\left[\sum_{t=0}^{\infty} C(x^j_t,\pi^j(x^j_t))\right] + \lim_{t\to\infty}\mathbb{E}_{\mathbb{1}\{x^j_t\notin\mathcal{G}\}}\left[\mathbb{E}_{x^j_t}\left[J^j_{t\to t+H}(x_t)|\,\mathbb{1}\{x^j_t\notin\mathcal{G}\}\right]\right] \tag{10.3.17}$$

$$= \mathbb{E}_{x^j}\left[\sum_{t=0}^{\infty} C(x^j_t,\pi^j(x^j_t))\right] + \lim_{t\to\infty}\mathbb{E}_{x^j_t}\left[J^j_{t\to t+H}(x^j_t)|x^j_t\notin\mathcal{G}\right]P(x^j_t\notin\mathcal{G}) \tag{10.3.18}$$

$$\geq \mathbb{E}_{x^j}\left[\sum_{t=0}^{\infty} C(x^j_t,\pi^j(x^j_t))\right] + \lim_{t\to\infty}\varepsilon P(x^j_t\notin\mathcal{G}) \tag{10.3.19}$$

$$= \mathbb{E}_{x^j}\left[\sum_{t=0}^{\infty} C(x^j_t,\pi^j(x^j_t))\right] = J^{\pi^j}(x_0) \tag{10.3.20}$$

Equations 10.3.15 and 10.3.16 follow from repeated application of lemma 1 (10.3.12). Equation 10.3.17 follows from iterated expectation, equation 10.3.18 follows from the cost function assumption (6). Equation 10.3.19 follows again from the cost function assumption (incur a cost of at least ε for not being at the goal at time t). Then, Equation 10.3.20 follows from lemma 2. Using the above inequality with the definition of $J^{\pi^j}(x_0)$,

$$J^j_{0\to H}(x_0) \geq J^{\pi^j}(x_0) = \mathbb{E}_{x^j_{1:H}}\left[\sum_{t=0}^{H-1} C(x^j_t,\pi^j(x_t)) + V^{\pi^j}(x^j_H)\right] \tag{10.3.21}$$

$$\geq \mathbb{E}_{x^j_{1:H|0}}\left[\sum_{t=0}^{H-1} C(x^j_t,\pi^{*,j}_{t|0}(x_{t|0})) + V^{\pi^j}(x_{H|0})\right] = J^{j+1}_{0\to H}(x_0) \tag{10.3.22}$$

$$\geq J^{\pi^{j+1}}(x_0) \tag{10.3.23}$$

Equation 10.3.21 follows from equation 10.3.20, equation 10.3.22 follows from taking the minimum over all possible H-length sequences of policies in the policy class Π. Equation 10.3.23 follows from equation 10.3.20. By induction, this proves the theorem.

If the limit is not dropped in 10.3.16, then we can roughly quantify a rate of improvement:

$$J^{\pi^j}(x_0) \leq J^{\pi_0}(x_0) - \sum_{k=0}^{j}\lim_{t\to\infty}\mathbb{E}_{x^k_t}\left[J^k_{t\to t+H}(x^k_t)\right]$$

By the Monotone Convergence Theorem, this also implies convergence of $(J^{\pi^j}(x_0))_{j=0}^{\infty}$. □

11 Experimental details for SAVED and baselines

For all experiments, we run each algorithm 3 times to control for stochasticity in training and plot the mean iteration cost vs. time with error bars indicating the standard deviation over the 3 runs. Additionally, when reporting task success rate and constraint satisfaction rate, we show bar plots

indicating the median value over the 3 runs with error bars between the lowest and highest value over the 3 runs. Experiments are run on an Nvidia DGX-1 and on a desktop running Ubuntu 16.04 with a 3.60 GHz Intel Core i7-6850K, 12 core CPU and an NVIDIA GeForce GTX 1080. When reporting the iteration cost of SAVED and all baselines, any constraint violating trajectory is reported by assigning it the maximum possible iteration cost T, where T is the task horizon. Thus, any constraint violation is treated as a catastrophic failure. We plan to explore soft constraints as well in future work.

11.1 SAVED

11.1.1 Dynamics and value function

For all environments, dynamics models and value functions are each represented with a probabilistic ensemble of 5, 3 layer neural networks with 500 hidden units per layer with swish activations as used in Chua et al. [13]. To plan over the dynamics, the TS-∞ trajectory sampling method from [13] is used. We use 5 and 30 training epochs for dynamics and value function training when initializing from demonstrations. When updating the models after each training iteration, 5 and 15 epochs are used for the dynamics and value functions respectively. All models are trained using the Adam optimizer with learning rate 0.00075 and 0.001 for the dynamics and value functions respectively. Value function initialization is done by training the value function using the true cost-to-go estimates from demonstrations. However, when updated on-policy, the value function is trained using temporal difference error (TD-1) on a buffer containing all prior states. Since we use a probabilistic ensemble of neural networks to represent dynamics models and value functions, we built off of the provided implementation [12] of PETS in [13].

11.1.2 Constrained exploration

Define states from which the system was successfully stabilized to the goal in the past as safe states. We train density model ρ on a fixed history of safe states, where this history is tuned based on the experiment. We have found that simply training on all prior safe states works well in practice on all experiments in this work. We represent the density model using kernel density estimation with a tophat kernel. Instead of modifying δ for each environment, we set $\delta = 0$ (keeping points with positive density), and modify α (the kernel parameter/width). We find that this works well in practice, and allows us to speed up execution by using a nearest neighbors algorithm implementation from scikit-learn. We are experimenting with locality sensitive hashing, implicit density estimation as in Fu et al. [17], and have had some success with Gaussian kernels as well (at significant additional computational cost).

11.2 Behavior cloning

We represent the behavior cloning policy with a neural network with 3 layers of 200 hidden units each for navigation tasks and pick and place, and 2 layers of 20 hidden units each for the PR2 Reacher task. We train on the same demonstrations provided to SAVED and other baselines for 50 epochs.

11.3 PETSfD and PETSfD Dense

PETSfD and PETSfD Dense use the same network architectures and training procedure as SAVED and the same parameters for each task unless otherwise noted, but just omit the value function and density model ρ for enforcing constrained exploration. PETSfD uses a planning horizon that is long enough to complete the task, while PETSfD Dense uses the same planning horizon as SAVED.

11.4 SACfD

We use the rlkit implementation [41] of soft actor critic with the following parameters: batch size=128, discount=0.99, soft target $\tau = 0.001$, policy learning rate = $3e-4$, Q function learning rate = $3e-4$, and value function learning rate = $3e-4$, batch size = 128, replay buffer size = 1000000, discount factor = 0.99. All networks are two-layer multi-layer perceptrons (MLPs) with 300 hidden units. On the first training iteration, only transitions from demonstrations are used to train the critic. After this, SACfD is trained via rollouts from the actor network as usual. We use a similar reward function to that of SAVED, with a reward of -1 if the agent is not in the goal set and 0 if the agent is in the goal set. Additionally, for environments with constraints, we impose a reward of -100 when constraints are

violated to encourage constraint satisfaction. The choice of collision reward is ablated in section 16.2. This reward is set to prioritize constraint satisfaction over task success, which is consistent with the selection of β in the model-based algorithms considered.

11.5 OEFD

We use the implementation of OEFD provided by Jangir [22] with the following parameters: learning rate = 0.001, polyak averaging coefficient = 0.8, and L2 regularization coefficient = 1. During training, the random action selection rate is 0.2 and the noise added to policy actions is distributed as $\mathcal{N}(0, 1)$. All networks are three-layer MLPs with 256 hidden units. Hindsight experience replay uses the "future" goal replay and selection strategy with $k = 4$ [3]. Here k controls the ratio of HER data to data coming from normal experience replay in the replay buffer. We use a similar reward function to that of SAVED, with a reward of -1 if the agent is not in the goal set and 0 if the agent is in the goal set. Additionally, for environments with constraints, we impose a reward of -100 when constraints are violated to encourage constraint satisfaction. The choice of collision reward is ablated in section 16.2. This reward is set to prioritize constraint satisfaction over task success, which is consistent with the selection of β in the model-based algorithms considered.

12 Simulated experimental details

12.1 Navigation

We consider a 4-dimensional (x, y, v_x, v_y) navigation task in which a point mass is navigating to a goal set, which is a unit ball centered at the origin. The agent can exert force in cardinal directions and experiences drag coefficient ψ and Gaussian process noise $z_t \sim \mathcal{N}(0, \sigma^2 I)$ in the dynamics. We use $\psi = 0.2$ and $\sigma = 0.05$ in all experiments in this domain. Demonstrations trajectories are generated by guiding the robot along a suboptimal hand-tuned trajectory for the first half of the trajectory before running LQR on a quadratic approximation of the true cost. Gaussian noise is added to the demonstrator policy. We train state density estimator ρ on all prior successful trajectories for the navigation tasks. Additionally, we use a planning horizon of 15 for SAVED and 25, 30, 30, 35 for PETSfD for tasks 1-4 respectively. The 4 experiments run on this environment are:

1. $x_0 = (-100, 0)$ Long navigation task to the origin. For experiments, 50 demonstrations with average return of 73.9 were used for training. We use kernel width $\alpha = 3$. SACfD and OEFD on average achieve a best iteration cost of 23.7 over 10,000 iterations of training averaged over the 3 runs.
2. $x_0 = (-50, 0)$ and a large obstacle blocking the x axis. This environment is difficult for approaches that use a Euclidean norm cost function due to local minima. For experiments, 50 demonstrations with average return of 67.9 were used for training. We use kernel width $\alpha = 3$ and chance constraint parameter $\beta = 1$. SACfD and OEFD achieve a best iteration cost of 21 and 21.7 respectively over 10,000 iterations of training averaged over the 3 runs.
3. $x_0 = (-50, 0)$ and a large obstacle near the path directly to the origin with a small channel near the x axis for passage. This environment is difficult for the algorithm to optimally solve since the iterative improvement of paths taken by the agent is constrained. For experiments, 50 demonstrations with average return of 67.9 were used for training. We use kernel width $\alpha = 3$ and chance constraint parameter $\beta = 1$. SACfD and OEFD achieve a best iteration cost of 17.3 and 19 respectively over 10,000 iterations of training averaged over the 3 runs.
4. $x_0 = (-50, 0)$ and a large obstacle surrounds the goal set with a small channel for entry. This environment is extremely difficult to solve without demonstrations. We use 100 demonstrations with average return of 78.3 and kernel width $\alpha = 3$ and chance constraint parameter $\beta = 1$. SACfD and OEFD achieve a best iteration cost of 23.7 and 40 respectively over 10,000 iterations of training averaged over the 3 runs.

12.2 PR2 reacher

We use 20 demonstrations for training, with no demonstration achieving total iteration cost less than 70, and with average iteration cost of 94.6. We use $\alpha = 15$ for all experiments. No other constraints are imposed, so the chance constraint parameter β is not used. The state representation consists of 7 joint positions, 7 joint velocities, and the goal position. The goal set is specified by a 0.05 radius

Euclidean ball in state space. SACfD and OEFD achieve a best iteration cost of 9 and 60 respectively over 10,000 iterations of training averaged over the 3 runs. We train state density estimator ρ on all prior successful trajectories for the PR2 reacher task. Additionally we use a planning horizon of 25 for both SAVED and PETSfD.

12.3 Fetch pick and place

We use 100 demonstrations generated by a hand-tuned PID controller with average iteration cost of 34.4. For SAVED, we set $\alpha = 0.05$. No other constraints are imposed, so the chance constraint parameter β is not used. The state representation for the task consists of (end effector relative position to object, object relative position to goal, gripper jaw positions). We find the gripper closing motion to be difficult to learn with SAVED, so we automate this motion by automatically closing it when the end effector is close enough to the object. We hypothesize that this is due to a combination of instability in the value function in this region and the difficulty of sampling bimodal behavior using CEM (open and close). SACfD and OEFD achieve a best iteration cost of 6 over 10,000 iterations of training averaged over the 3 runs. We train state density estimator ρ on the last 5000 safe states (100 trajectories) for the Fetch pick and place task. Additionally we use a planning horizon of 10 for SAVED and 20 for PETSfD.

13 Physical experimental details

For all experiments, $\alpha = 0.05$ and a set of 100 hand-coded trajectories with a small amount of Gaussian noise added to the controls is generated. For all physical experiments, we use $\beta = 1$ for PETSfD since we found this gave the best performance when no signal from the value function was provided. In all tasks, the goal set is represented with a 1 cm ball in $\mathbb{R}^3$. The dVRK is controlled via delta-position control, with a maximum action magnitude set to 1 cm during learning for safety. We train state density estimator ρ on all prior successful trajectories for the physical robot experiments.

13.1 Figure-8

The agent is constrained to remain within a 1 cm pipe around a reference trajectory with chance constraint parameter $\beta = 0.8$ for SAVED and $\beta = 1$ for PETSfD. We use 100 inefficient but successful and constraint-satisfying demonstrations with average iteration cost of 40 steps for both segments. Additionally we use a planning horizon of 10 for SAVED and 30 for PETSfD.

13.2 Knot tying

The agent is again constrained to remain within a 1 cm tube around a reference trajectory as in prior experiments with chance constraint parameter $\beta = 0.65$ for SAVED and $\beta = 1$ for PETSfD. Provided demonstrations are feasible but noisy and inefficient due to hand-engineering and have average iteration cost of 34 steps. Additionally we use a planning horizon of 10 for SAVED and 20 for PETSfD.

14 Simulated experiments additional results

In Figure 6, we show the task success rate for the PR2 reacher and Fetch pick and place tasks for SAVED and baselines. We note that SAVED outperforms RL baselines (except SAVED (No SS) for the reacher task, most likely because the task is relatively simple so the *sampled safe set* constraint has little effect) in the first 100 and 250 iterations for the reacher and pick and place tasks respectively. Note that although behavior cloning has a higher success rate, it does not improve upon demonstration performance. However, although SAVED's success rate is not as different from the baselines in these environments as those with constraints, this result shows that SAVED can be used effectively in a general purpose way, and still learns more efficiently than baselines in unconstrained environments as seen in the main paper.

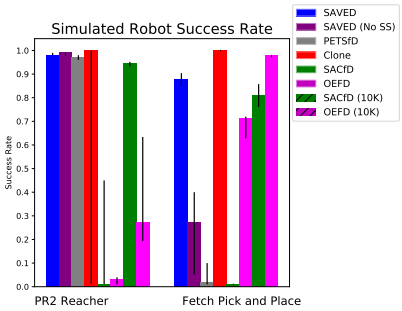


Figure 6: SAVED has comparable success rate to Clone, PETSfD, and SAVED (No SS) on the reacher task in the first 100 iterations. For the pick and place task, SAVED outperforms all baselines in the first 250 iterations except for Clone, which does not improve upon demonstration performance.

15 Physical experiments additional results

15.1 Figure-8

For the other segment of the Figure-8, SAVED still quickly learns to smooth out demo trajectories while satisfying constraints, with a success rate of over 80% while baselines violate constraints on nearly every iteration and never complete the task, as shown in Figure 7.

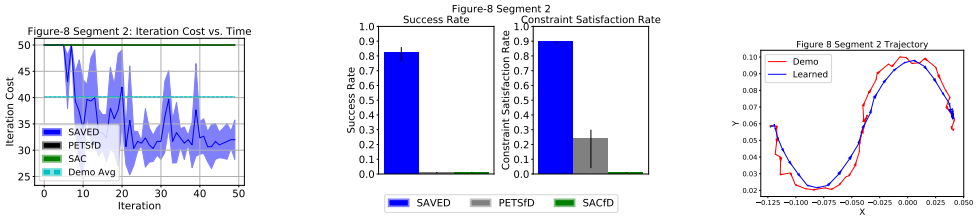


Figure 7: **Figure-8: Training Performance:** After 10 iterations, the agent consistently completes the task and converges to an iteration cost of around 32, faster than demos which took an average of 40 steps. Neither baseline ever completed the task in the first 50 iterations; **Trajectories:** Demo trajectories are constraint-satisfying, but noisy and inefficient. SAVED quickly learns to speed up demos with only occasional constraint violations and successfully stabilizes in the goal set. Note that due to the difficulty of the tube constraint, it is hard to avoid occasional constraint violations during learning, which are reflected by spikes in the iteration cost.

In Figure 8, we show the full trajectory for the figure-8 task when both segments are combined at execution-time. This is done by rolling out the policy for the first segment, and then starting the policy for the second segment as soon as the policy for the first segment reaches the goal set. We see that even given uncertainty in the dynamics and end state for the first policy (it could end anywhere in a 1 cm ball around the goal position), SAVED is able to smoothly navigate these issues and interpolate between the two segments at execution-time to successfully stabilize at the goal at the end of the second segment. Each segment of the trajectory is shown in a different color for clarity. We suspect that SAVED's ability to handle this transition is reflective of good generalization of the learned dynamics and value functions.

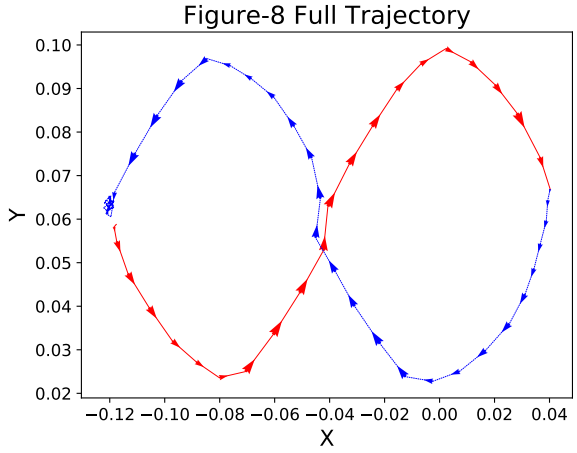


Figure 8: **Full figure-8 trajectory:** We show the full figure-8 trajectory, obtained by evaluating learned policies for the first and second figure-8 segments in succession. Even when segmenting the task, the agent can smoothly interpolate between the segments, successfully navigating the uncertainty in the transition at execution-time and stabilizing in the goal set.

15.2 Knot tying

In Figure 9, we show the full trajectory for both arms for the surgical knot tying task. We see that the learned policy for arm 1 smoothly navigates around arm 2, after which arm 1 is manually moved down along with arm 2, which grasps the thread and pulls it up through the resulting loop in the thread, completing the knot.

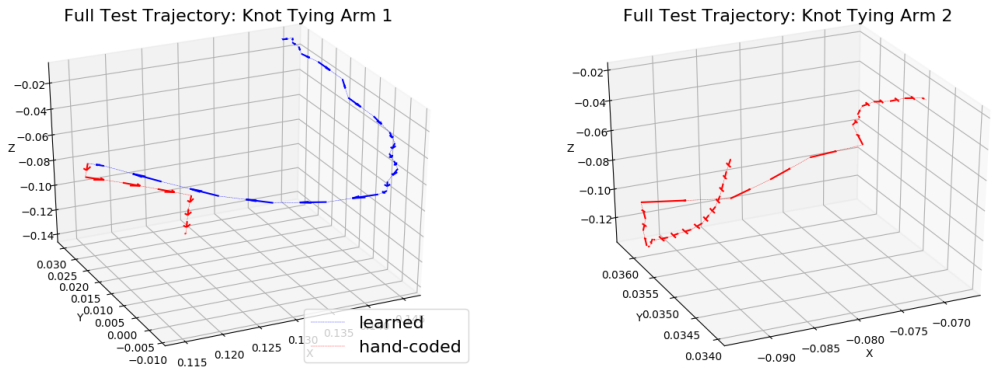


Figure 9: **Knot-Tying full trajectories:** **(a) Arm 1 trajectory:** We see that the learned part of the arm 1 trajectory is significantly smoothed compared to the demonstrations at execution-time as well, consistent with the training results. Then, in the hand-coded portion of the trajectory, arm 1 is simply moved down towards the phantom along with arm 2, which grasps the thread and pulls it up; **(b) Arm 2 trajectory:** This trajectory is hand-coded to move down towards the phantom after arm 1 has fully wrapped the thread around it, grasp the thread, and pull it up.

16 Ablations

16.1 SAVED

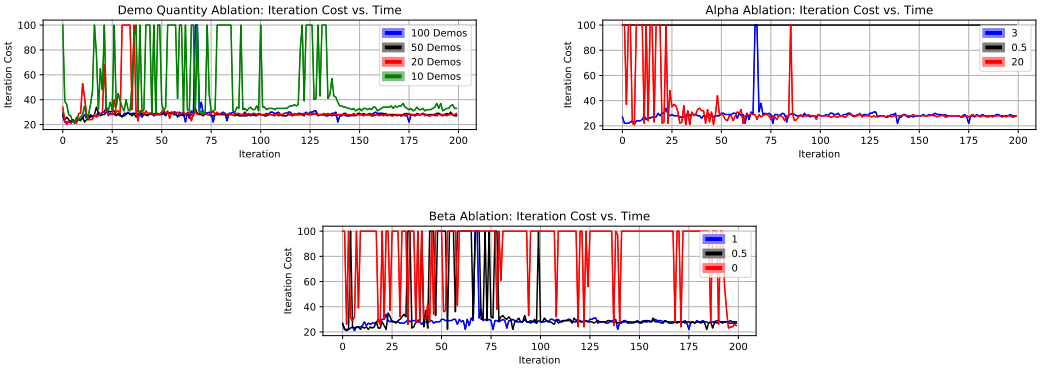


Figure 10: **SAVED Ablations on Navigation Task 2: Number of Demonstrations:** We see that SAVED is able to complete the task with just 20 demonstrations, but more demonstrations result in increased stability during learning; **Kernel width α:** We see that α must be chosen to be high enough such that SAVED is able to explore enough to find the goal set, but not so high that SAVED starts to explore unsafe regions of the state space; **Chance constraint parameter β:** Decreasing β results in many more collisions with the obstacle. Ignoring the obstacle entirely results in the majority of trials ending in collision or failure.

We investigate the impact of kernel width α, chance constraint parameter β, and the number of demonstrator trajectories used on navigation task 2. Results are shown in Figure 10. We see that SAVED is able to complete the task well even with just 20 demonstrations, but is more consistent with more demonstrations. We also notice that SAVED is relatively sensitive to the setting of kernel width α. When α is set too low, we see that SAVED is overly conservative, and thus can barely explore at all. This makes it difficult to discover regions near the goal set early on and leads to significant model mismatch, resulting in poor task performance. Setting α too low can also make it difficult for SAVED to plan to regions with high density further along the task, resulting in SAVED failing to make progress. On the other extreme, making α too large causes a lot of initial instability as the agent explores unsafe regions of the state space. Thus, α must be chosen such that SAVED is able to sufficiently explore, but does not explore so aggressively that it starts visiting states from which it has low confidence in being able reach the goal set. Reducing β allows the agent to take more risks, but this results in many more collisions. With $\beta = 0$, most rollouts end in collision or failure as expected. In the physical experiments, we find that allowing the agent to take some risk during exploration is useful due to the difficult tube constraints on the state space.

16.2 Model-Free

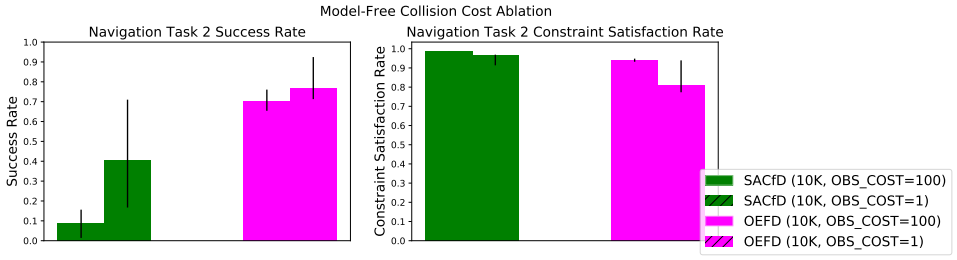


Figure 11: A high cost for constraint violations results in conservative behavior that learns to avoid the obstacle, but also makes it take longer to learn to perform the task. Setting the cost low results in riskier behavior that more often achieves task success.

To convey information about constraints to model-free methods, we provide an additional cost for constraint violations. We ablate this parameter for navigation task 2 in Figure 11. We find that a high

cost for constraint violations results in conservative behavior that learns to avoid the obstacle, but also takes much longer to achieve task success. Setting the cost low results in riskier behavior that succeeds more often. This trade-off is also present for model-based methods, as seen in the prior ablations.