Safe Policy Search for Lifelong Reinforcement Learning with Sublinear Regret

Penn Engineering


Haitham Bou Ammar
University of Pennsylvania


Rasul Tutunov
University of Pennsylvania


Eric Eaton
University of Pennsylvania

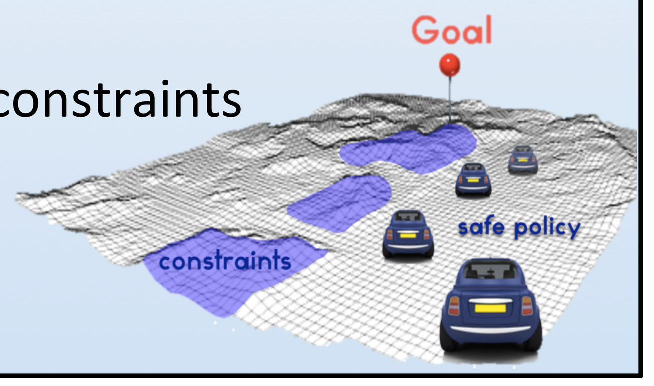
Summary

We developed a lifelong policy gradient learner that operates in an adversarial setting to learn multiple tasks online while enforcing safety constraints on the learned policies.

- Fully online learning of multiple, consecutive RL tasks
- Ensures "safe" policies by respecting safety constraints
- Exhibits *sublinear regret* for lifelong policy search
- Validated on benchmark dynamical systems and quadrotor control

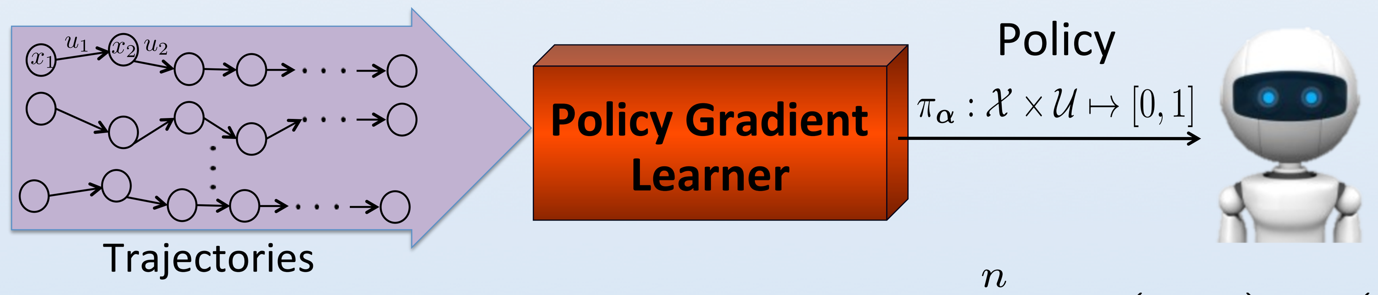
Motivations

1. Reuse knowledge from previously learned tasks to accelerate the learning of new control policies
   → Lifelong RL to learn multiple, consecutive tasks online
   → Exhibit vanishing regrets
2. Robotic control policies must obey safety constraints (e.g., prevent damage to the robot and environment, avoid catastrophic failure)
   → Incorporate constraints directly into the optimization



Background: Policy Gradient (PG) Methods

- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control



Goal: find policy π_{α} that minimizes $l(\alpha) = \sum_{k=1}^{n} p_{\alpha}\left(\tau^{(k)}\right) C\left(\tau^{(k)}\right)$

probability of trajectory: $p_{\alpha}\left(\tau^{(k)}\right) = \mathcal{P}_0\left(x_0^{(k)}\right) \prod_{m=0}^{M-1} \mathcal{P}\left(x_{m+1}^{(k)} | x_m^{(k)}, u_m^{(k)}\right) \pi_{\alpha}\left(u_m^{(k)} | x_m^{(k)}\right)$

cost of trajectory: $C\left(\tau^{(k)}\right) = \frac{1}{M} \sum_{m=0}^{M-1} c_{m+1}^{(k)}$

Background: Online Learning & Regret Analysis

Regret Minimization Game: Each round $j = 1 \ldots R$,

a.) agent chooses a prediction θ_j, and

b.) environment (i.e., the adversary) chooses a loss function l_j

Goal: minimize cumulative regret (modified for multi-task case)

$$\mathfrak{R}_R = \sum_{j=1}^{R} l_{t_j}(\boldsymbol{\theta}_j) - \inf_{u \in \mathcal{K}} \left[\sum_{j=1}^{R} l_{t_j}(\boldsymbol{u}) \right]$$

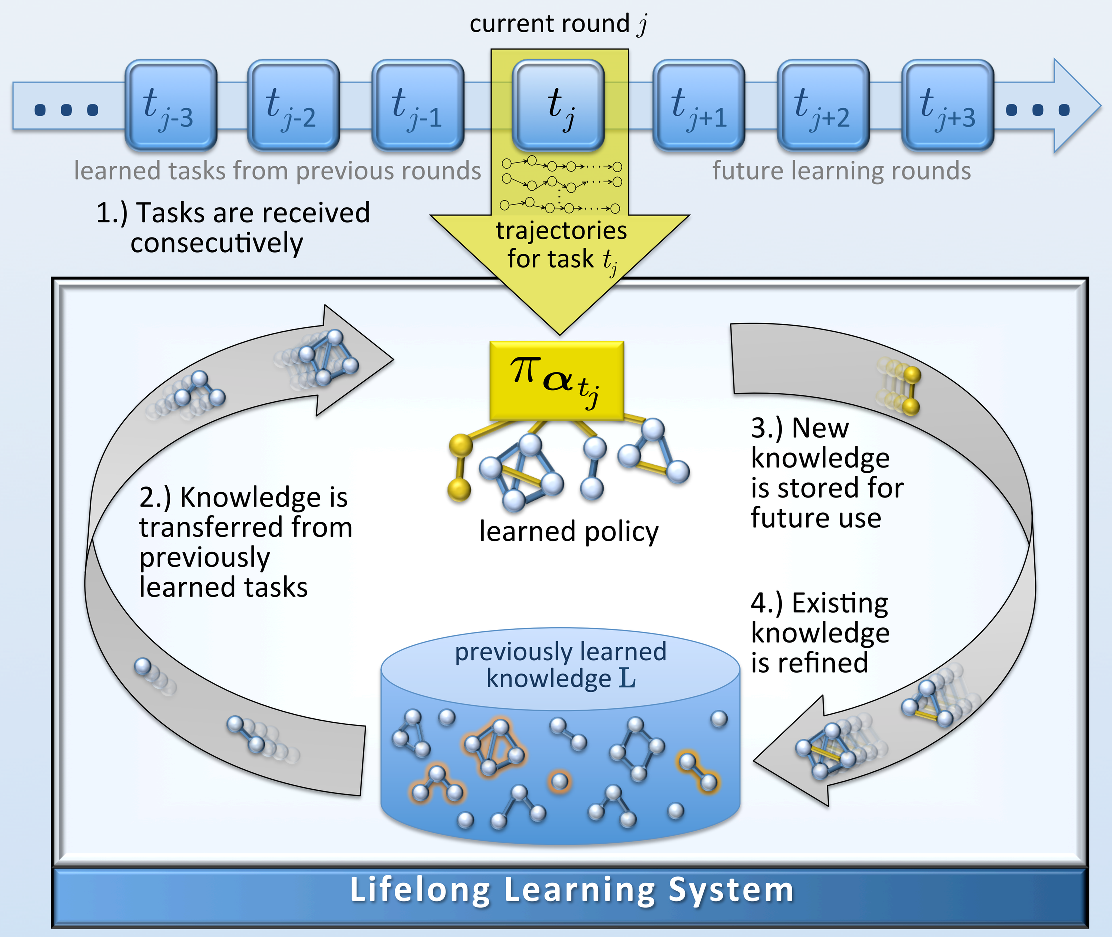
l_{t_j}: loss of task t at round j

Solve via "Follow the Regularized Leader":

1.) Find $\tilde{\theta}$ via unconstrained optimization over accumulated losses

2.) Project $\tilde{\theta}$ onto the constraint set via Bregman projections

Lifelong Learning Framework

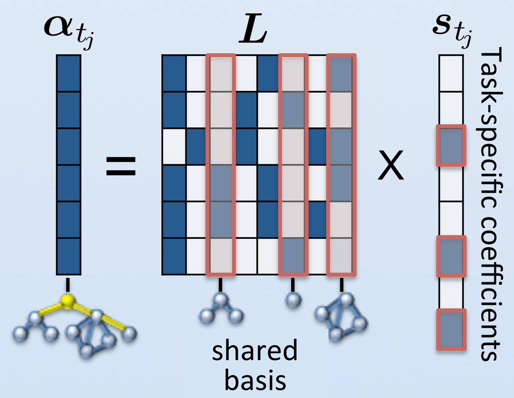


Each task has associated safety constraints

$$A_{t_j} \alpha_{t_j} \leq b_{t_j}$$

Policy parameters

$$\alpha_{t_j} = L s_{t_j}$$



Safe Lifelong Policy Search

Multi-task Optimization Problem after observing r rounds:

$$\min_{L,S} \sum_{j=1}^{r} \left[\eta_{t_j} l_{t_j}\left(L s_{t_j}\right)\right] + \mu_1 ||S||_F^2 + \mu_2 ||L||_F^2$$

$$\text{s.t.} \quad A_{t_j} \alpha_{t_j} \leq b_{t_j} \quad \forall t_j \in \mathcal{I}_r$$

$$\lambda_{\min}\left(LL^{\mathsf{T}}\right) \geq p \quad \text{and} \quad \lambda_{\max}\left(LL^{\mathsf{T}}\right) \leq q$$

Online Formulation

- Let $\boldsymbol{\theta} = [\text{vec}(L) \;\; \text{vec}(S)]^{\mathsf{T}}$ be the vector of all parameters
- The MTL objective can be written online as

$$\boldsymbol{\theta}_{r+1} = \arg\min_{\boldsymbol{\theta} \in \mathcal{K}} \boldsymbol{\Omega}_r(\boldsymbol{\theta}) \qquad \boldsymbol{\Omega}_0(\boldsymbol{\theta}) = \mu_2 \sum_{i=1}^{dk} \boldsymbol{\theta}_i^2 + \mu_1 \sum_{i=dk+1}^{dk+k|\mathcal{T}|} \boldsymbol{\theta}_i^2$$

$$\boldsymbol{\Omega}_j(\boldsymbol{\theta}) = \boldsymbol{\Omega}_{j-1}(\boldsymbol{\theta}) + \eta_{t_j} l_{t_j}(\boldsymbol{\theta})$$

where the loss for task t_j is the following bilinear product in $\boldsymbol{\theta}$:

$$l_{t_j}(\boldsymbol{\theta}) = -\frac{1}{n_{t_j}} \sum_{k=1}^{n_{t_j}} \sum_{m=0}^{M_{t_j}-1} \log\left[\pi^{(t_j)}_{\boldsymbol{\Theta}_L \boldsymbol{\Theta}_{s_{t_j}}}\left(u_m^{(k,\, t_j)} \mid x_m^{(k,\, t_j)}\right)\right]$$

$$\boldsymbol{\Theta}_L = \begin{bmatrix} \boldsymbol{\theta}_1 & \ldots & \boldsymbol{\theta}_{d(k-1)+1} \\ \vdots & \vdots & \vdots \\ \boldsymbol{\theta}_d & \ldots & \boldsymbol{\theta}_{dk} \end{bmatrix}, \quad \boldsymbol{\Theta}_{s_{t_j}} = \begin{bmatrix} \boldsymbol{\theta}_{dk+1} \\ \vdots \\ \boldsymbol{\theta}_{(d+1)k+1} \end{bmatrix}$$

Online Solution

Step 1: Unconstrained policy optimization via alternating optimization over $\mathbf{L}$ and $\mathbf{S}$ to yield unconstrained solution $\tilde{\boldsymbol{\theta}}_{r+1}$

Step 2: Project $\tilde{\boldsymbol{\theta}}_{r+1}$ to the constraint set via the Bregman divergence over $\boldsymbol{\Omega}_r$ — involves solving 2nd order cone and semi-definite programs

Main Theoretical Result

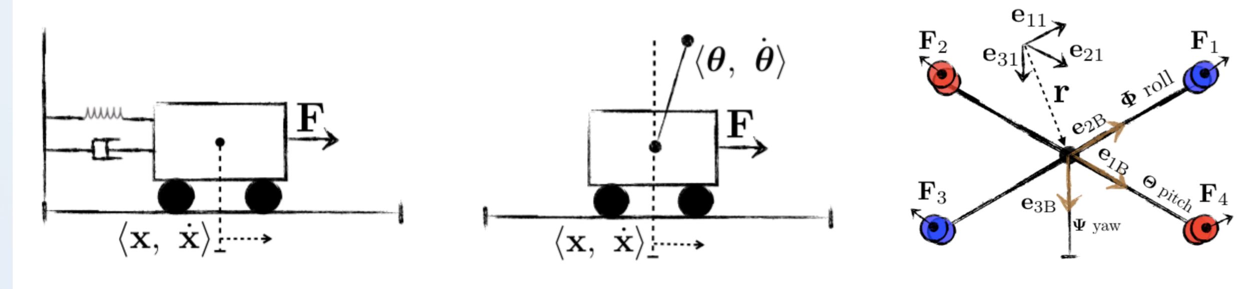
We prove that the safe lifelong policy search algorithm has sublinear regret of $\mathcal{O}(\sqrt{R})$ in the total number of rounds R.

Theorem 1 (Sublinear Regret). *After R rounds and choosing $\forall t_j \in \mathcal{I}_R \;\; \eta_{t_j} = \eta = \frac{1}{\sqrt{R}}$, $L\big|_{\hat{\theta}_1} = diag_{\boldsymbol{k}}(\zeta)$, with $diag_{\boldsymbol{k}}(\cdot)$ being a diagonal matrix among the $\boldsymbol{k}$ columns of L, $p \leq \zeta^2 \leq q$, and $S\big|_{\hat{\theta}_1} = 0_{k \times |\mathcal{T}|}$, the safe lifelong policy search reinforcement learner exhibits sublinear regret of the form:*

$$\sum_{j=1}^{R} l_{t_j}\left(\hat{\boldsymbol{\theta}}_j\right) - l_{t_j}(\boldsymbol{u}) = \mathcal{O}\left(\sqrt{R}\right) \quad \textit{for any } \boldsymbol{u} \in \mathcal{K}.$$

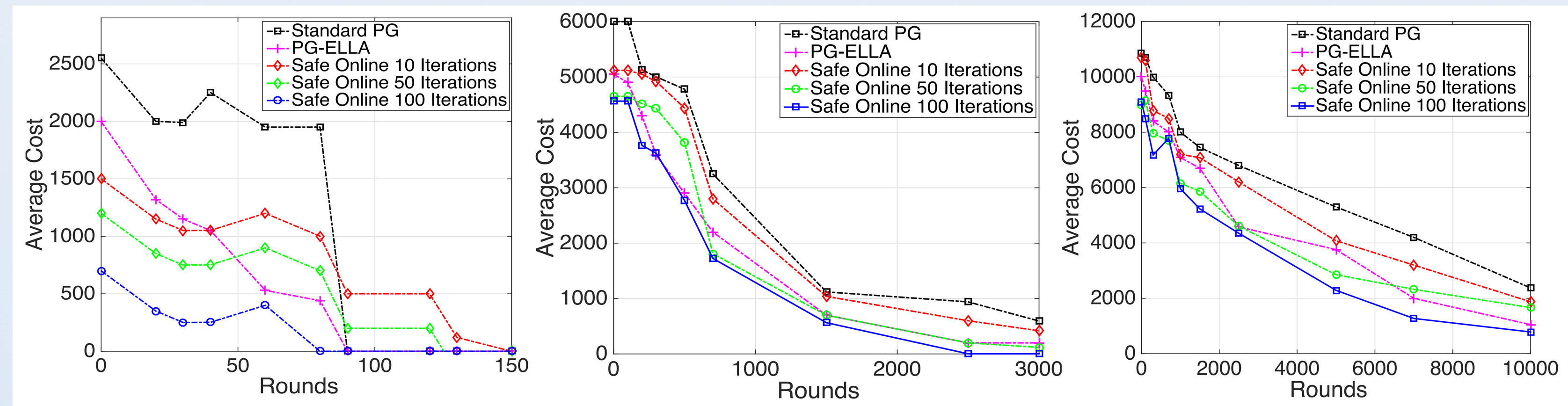
Experimental Results

Learn policies for consecutive control tasks on three types of systems



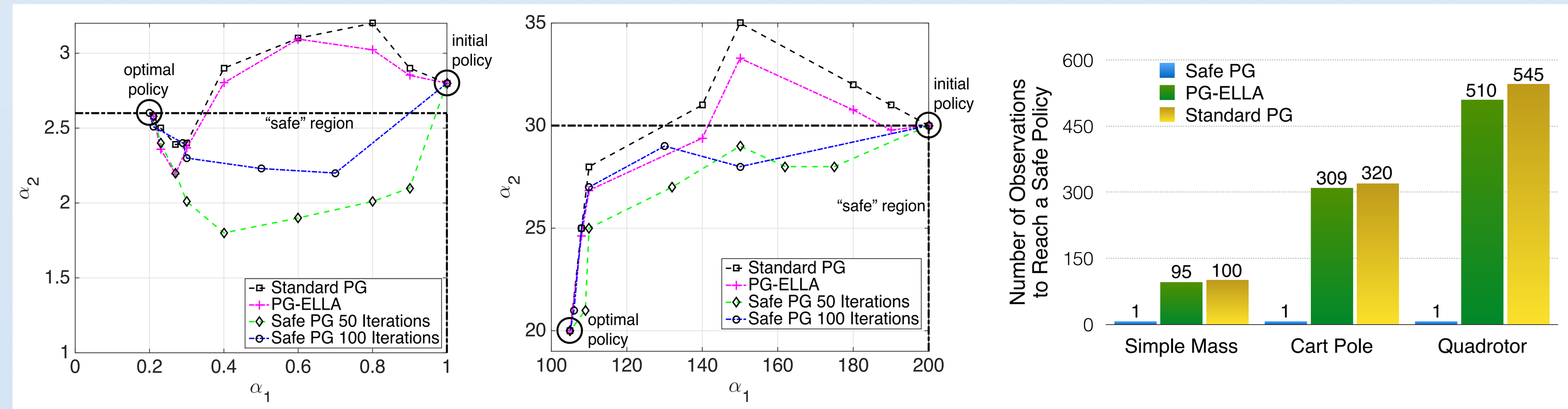
Simple Mass | Cart Pole | Quadrotor

Superior Performance over standard PG and the lifelong learner PG-ELLA



Simple Mass | Cart Pole | Quadrotor

Enforces the Given Safety Constraints, unlike alternative methods



Trajectory Simple Mass | Trajectory Cart Pole | Number of Observations to Reach a Safe Policy

- Note that our approach immediately projects policies to safe regions even during the policy search process, unlike other methods

Acknowledgements

This research was supported by ONR grant #N00014-11-1-0139 and AFRL grant #FA8750-14-1-0069.

We thank Ali Jadbabaie for assistance with the optimization solution, and the anonymous reviewers for their helpful feedback.