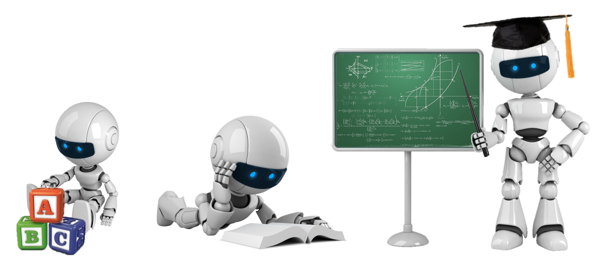
Safe Policy Search for Lifelong Reinforcement Learning with Sublinear Regret

Haitham Bou Ammar
Univ. of Pennsylvania

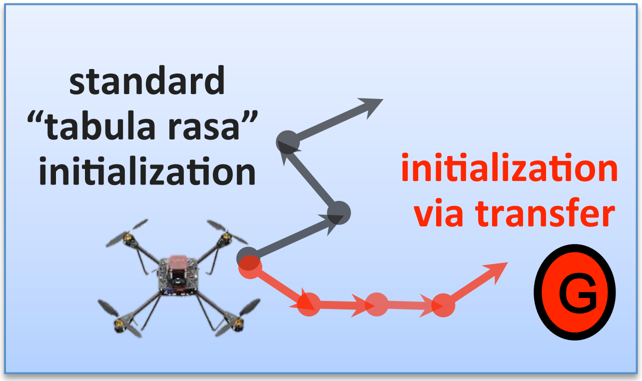
Rasul Tutunov
Univ. of Pennsylvania

Eric Eaton
Univ. of Pennsylvania

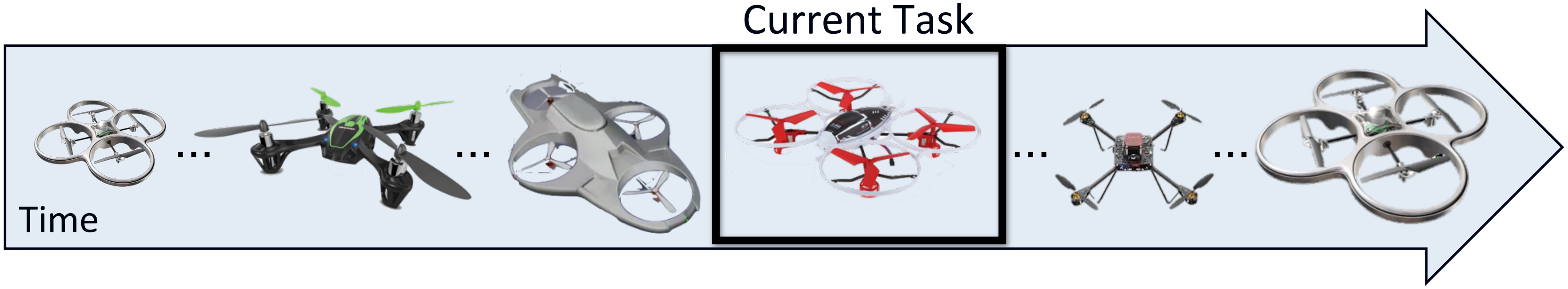


Motivation

Problem 1: Without prior knowledge, RL in a new task is slow



Idea: Reuse knowledge from previously learned tasks



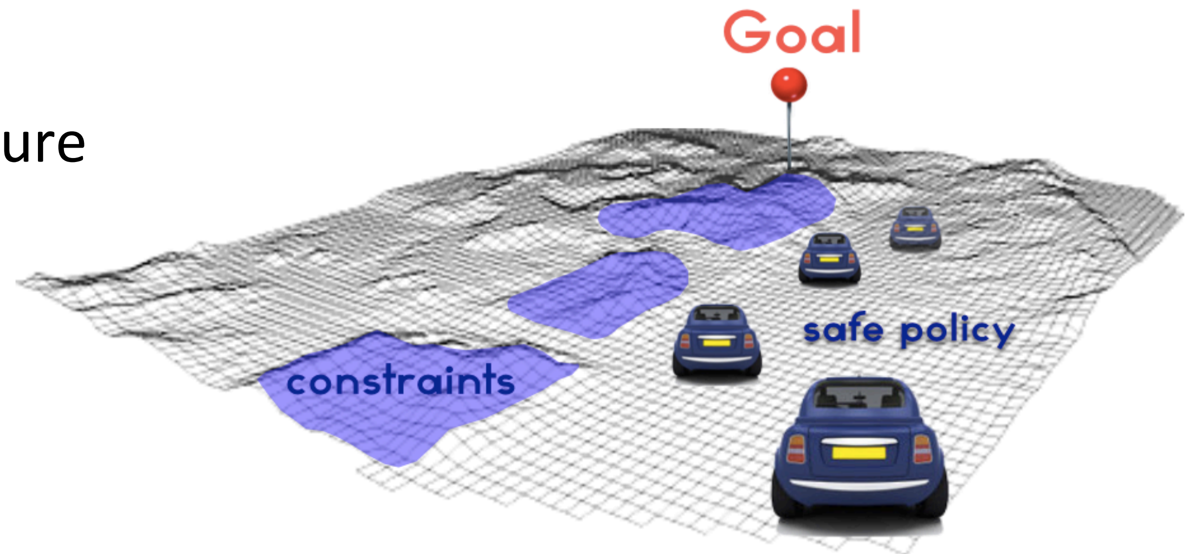
We focus on the lifelong learning case:

- Agent learns multiple tasks consecutively
- Want a fully online method with sublinear regret

Motivation

Problem 2: Robot control policies must obey safety constraints

- Prevent damage to the robot or environment
- Limit joint velocities
- Avoid catastrophic failure

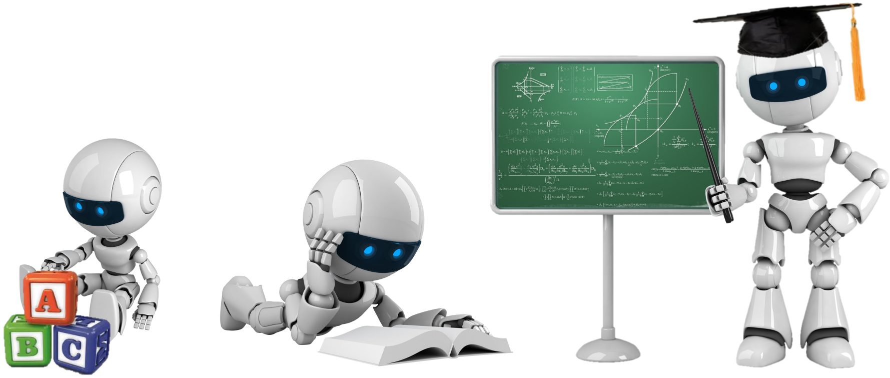


Idea: Incorporate constraints directly into policy optimization

Contribution

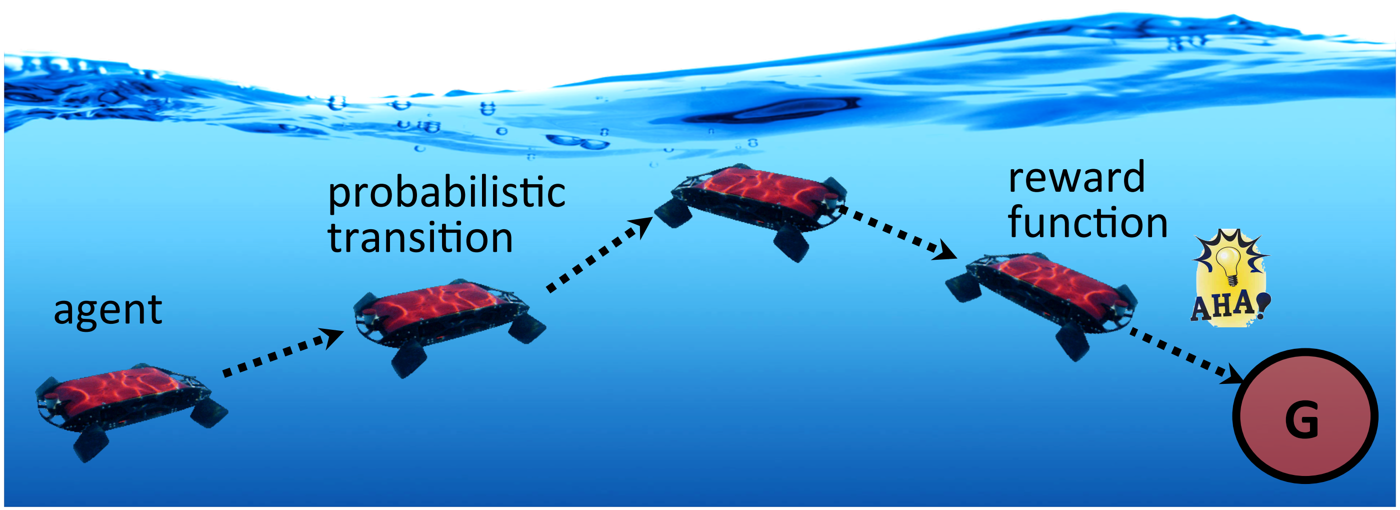
Safe lifelong policy gradient reinforcement learner

- Learns multiple, consecutive RL tasks online
- Operates in an adversarial setting
- Ensures that policies respect given safety constraints
- Exhibits sublinear regret for lifelong policy search



Background: Policy Gradient Methods for Control

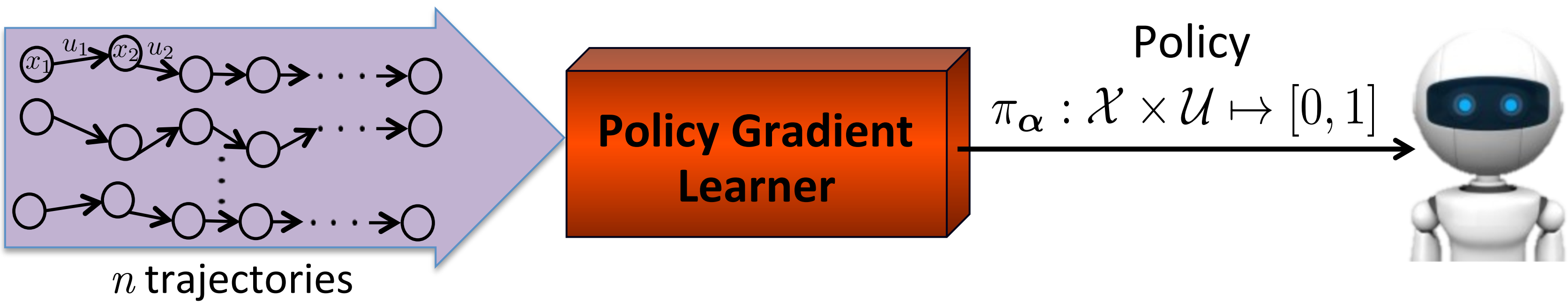
- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control
    [Kober & Peters 2011; Peters & Schaal 2008; Sutton et al. 2000]



Agent makes sequential decisions

Background: Policy Gradient Methods for Control

- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control
    [Kober & Peters 2011; Peters & Schaal 2008; Sutton et al. 2000]



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

probability of trajectory

$$p_{\boldsymbol{\alpha}}\left(\boldsymbol{\tau}^{(k)}\right) = \mathcal{P}_0\left(\boldsymbol{x}_0^{(k)}\right) \prod_{m=0}^{M-1} \mathcal{P}\left(\boldsymbol{x}_{m+1}^{(k)} | \boldsymbol{x}_m^{(k)}, \boldsymbol{u}_m^{(k)}\right) \pi_{\boldsymbol{\alpha}}\left(\boldsymbol{u}_m^{(k)} | \boldsymbol{x}_m^{(k)}\right)$$

cost of trajectory

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

Background: Online Learning & Regret Analysis

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

a.) agent chooses a prediction $\boldsymbol{\alpha}_j$, and

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

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

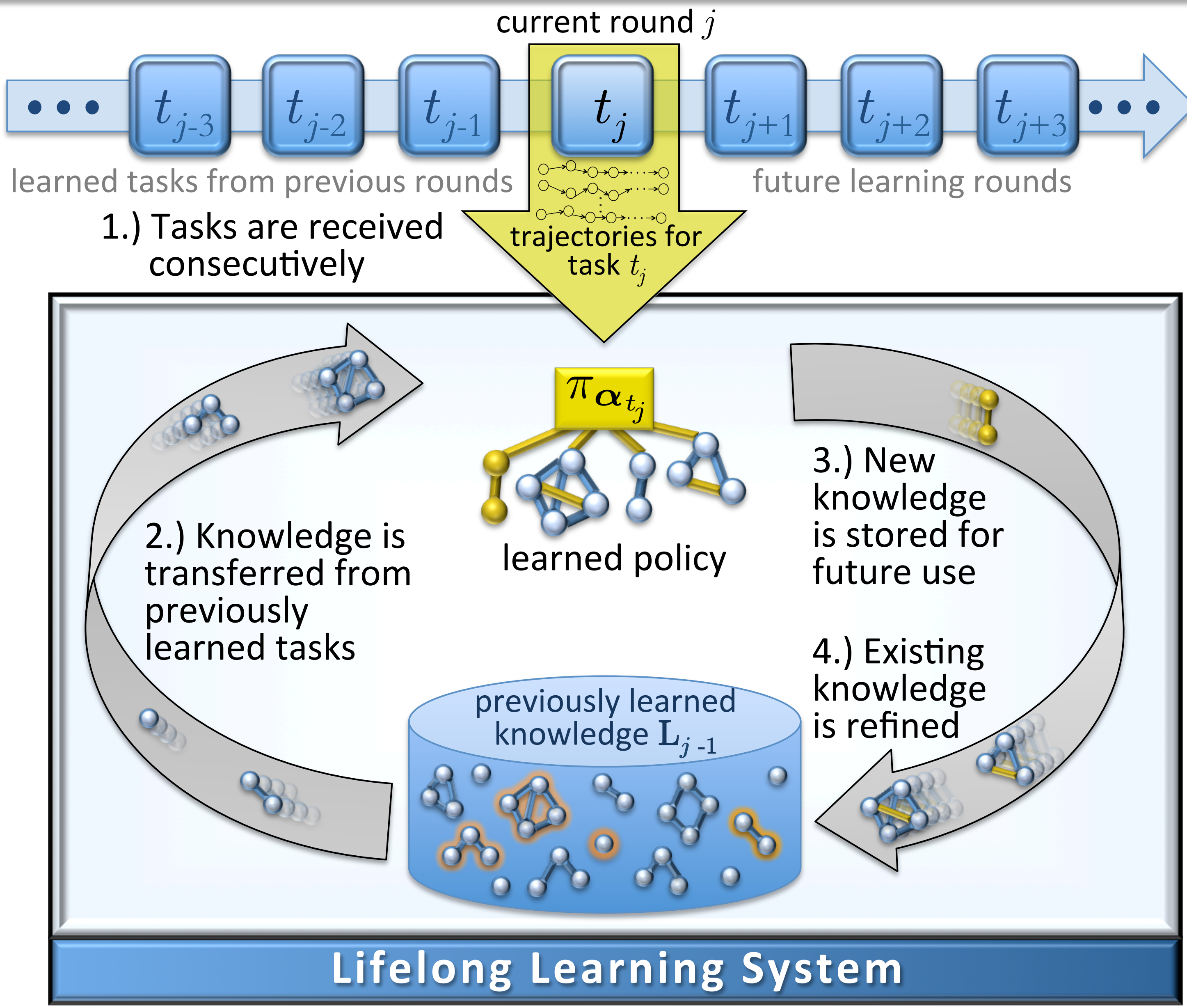
$$\mathfrak{R}_R = \sum_{j=1}^{R} l_{t_j}(\boldsymbol{\alpha}_j) - \inf_{\boldsymbol{\theta} \in \mathcal{K}} \left[\sum_{j=1}^{R} l_{t_j}(\boldsymbol{\theta}) \right]$$

loss of task t at round j

agent's total loss

best fixed loss in hindsight

Lifelong Machine Learning



Task Model

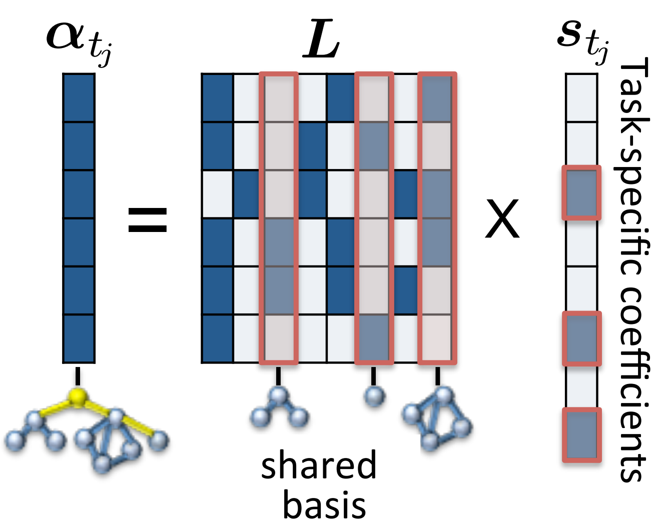
Policy gradient objective: $l(\boldsymbol{\alpha}) = \sum_{k=1}^{n} p_{\boldsymbol{\alpha}}\left(\boldsymbol{\tau}^{(k)}\right) \boldsymbol{C}\left(\boldsymbol{\tau}^{(k)}\right)$

- For a specific task t_j, find the optimal policy

$$\pi_{\boldsymbol{\alpha}^\star_{t_j}}(\boldsymbol{u} \mid \boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{\alpha}^\star_{t_j} = \min_{\boldsymbol{\alpha}} l_{t_j}(\boldsymbol{\alpha})$$

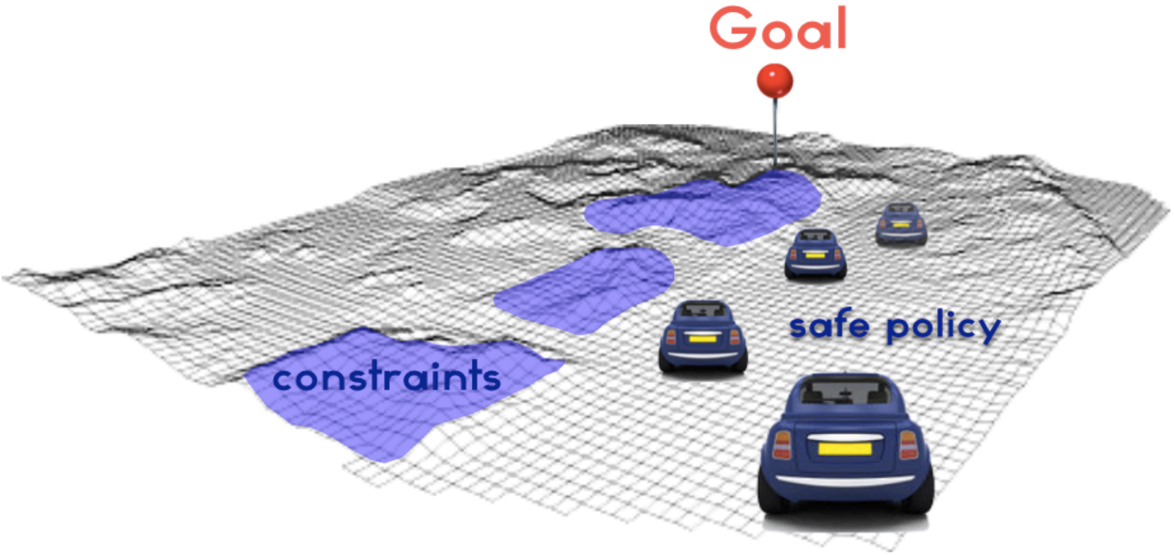
- The parameters $\boldsymbol{\alpha}_{t_j}$ are linear combinations of a shared basis $\boldsymbol{L}$

$$\boldsymbol{\alpha}_{t_j} = \boldsymbol{L}\boldsymbol{s}_{t_j} \qquad \boldsymbol{L} \in \mathbb{R}^{d \times k}, \boldsymbol{s}_{t_j} \in \mathbb{R}^k$$



Safety Constraints on Policy

Each task t_j has associated safety constraints $(\boldsymbol{A}_{t_j}, \boldsymbol{b}_{t_j})$ such that $\boldsymbol{A}_{t_j}\boldsymbol{\alpha}_{t_j} \leq \boldsymbol{b}_{t_j}$



Lifelong Learning Problem Definition

t_j Each round, we observe n_{t_j} trajectories of task t_j

Goal: minimize total cumulative loss-so-far

Online Multi-task Objective: After observing r rounds,

$$\min_{\boldsymbol{L},\boldsymbol{S}} \sum_{j=1}^{r} \left[\underbrace{\eta_{t_j} l_{t_j} \left(\boldsymbol{L}\boldsymbol{s}_{t_j} \right)}_{\text{loss for task } t_j} \right] + \underbrace{\mu_1 \|\boldsymbol{S}\|_{\mathsf{F}}^2 + \mu_2 \|\boldsymbol{L}\|_{\mathsf{F}}^2}_{\text{regularize projections and shared repository}}$$

$$\text{s.t.} \quad \mathbf{A}_{t_j} \boldsymbol{\alpha}_{t_j} \leq \mathbf{b}_{t_j} \quad \forall t_j \in \mathcal{I}_r$$

safety constraints

$$\underbrace{\boldsymbol{\lambda}_{\min}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \geq p \text{ and } \boldsymbol{\lambda}_{\max}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \leq q}_{\text{ensure "informative" policies by bounding } \|\mathbf{L}\|_{\mathsf{F}}}$$

Online Formulation

Online MTL Objective

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

$$\text{s.t.} \quad \mathbf{A}_{t_j}\boldsymbol{\alpha}_{t_j} \leq \mathbf{b}_{t_j} \quad \forall t_j \in \mathcal{I}_r$$

$$\boldsymbol{\lambda}_{\min}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \geq p \text{ and } \boldsymbol{\lambda}_{\max}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \leq q$$

Let $\boldsymbol{\theta} = [\text{vec}(\mathbf{L}), \text{vec}(\mathbf{S})]^{\mathsf{T}}$

We can re-write the objective as:

$$\boldsymbol{\theta}_{r+1} = \arg\min_{\boldsymbol{\theta}\in\mathcal{K}} \boldsymbol{\Omega}_r(\boldsymbol{\theta})$$

↑ set of safe policies

$$\boldsymbol{\Omega}_0(\boldsymbol{\theta}) = \mu_2 \sum_{i=1}^{dk} \boldsymbol{\theta}_i^2 + \mu_1 \sum_{i=1}^{dk+1} \boldsymbol{\theta}_i^2$$

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

Solution Strategy

Step 1: Unconstrained Solution

a.) Update $\mathbf{L}$, holding $\mathbf{S}$ fixed

$$\mathbf{L}_{\beta+1} = \mathbf{L}_{\beta} - \eta_{\mathbf{L}}^{\beta} \nabla_{\mathbf{L}} e_r(\mathbf{L}, \mathbf{S})$$

b.) Update $\mathbf{S}$, holding $\mathbf{L}$ fixed

$$\mathbf{s}_{\lambda+1}^{(t_j)} = \mathbf{s}_{\lambda}^{(t_j)} - \eta_{\mathbf{S}}^{\lambda} \nabla_{\mathbf{L}} e_r(\mathbf{L}, \mathbf{S})$$

$\tilde{\boldsymbol{\theta}}_{r+1}$ unconstrained solution

Step 2: Constrained Solution

Idea: Alternate to learn projection of $\tilde{\boldsymbol{\theta}}_{r+1}$ onto the constraint set

Problem: Computationally Expensive

Constrained Projection Learning

Learning the constrained solution is equivalent to:

$$\hat{\boldsymbol{\theta}}_{r+1} = \arg\min_{\boldsymbol{\theta} \in \mathcal{K}} \underbrace{\mathcal{B}_{\boldsymbol{\Omega}_r, \mathcal{K}}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_{r+1}\right)}_{\text{Bregman divergence}}$$

Reduce computational complexity by linearizing losses

$$l_{t_r}(\hat{\boldsymbol{u}}) = \hat{\mathbf{f}}_{t_r}\Big|_{\hat{\boldsymbol{\theta}}_r}^{\mathsf{T}} \hat{\mathbf{u}}$$

linearized loss around constrained solution to previous round

$$\hat{\mathbf{f}}_{t_r}\Big|_{\hat{\boldsymbol{\theta}}_r} = \underbrace{\left[\nabla_{\boldsymbol{\theta}} l_{t_r}(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_r}, \; l_{t_r}(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_r} - \nabla_{\boldsymbol{\theta}} l_{t_r}(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_r} \hat{\boldsymbol{\theta}}_r\right]^{\mathsf{T}}}_{\text{first-order term}}$$

Constrained Projection Learning

Using linearized losses, the constrained solution simplifies to:

$$\hat{\boldsymbol{\theta}}_{r+1} = \arg\min_{\boldsymbol{\theta}\in\mathcal{K}} \mathcal{B}_{\boldsymbol{\Omega}_0,\mathcal{K}}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_{r+1}\right)$$

Constrained Problem for Determining Safe Policies

$$\min_{\mathbf{L},\mathbf{S}} \mu_1||\mathbf{S}||_{\mathsf{F}}^2 + \mu_2||\mathbf{L}||_{\mathsf{F}}^2 + 2\mu_1 \mathrm{tr}\left(\mathbf{S}\Big|_{\tilde{\boldsymbol{\theta}}_{r+1}}^{\mathsf{T}} \mathbf{S}\right) + 2\mu_2 \mathrm{tr}\left(\mathbf{L}\Big|_{\tilde{\boldsymbol{\theta}}_{r+1}}^{\mathsf{T}} \mathbf{L}\right)$$

$$\text{s.t.} \quad \mathbf{A}_{t_j}\boldsymbol{L}\boldsymbol{\alpha}_{t_j} \leq \mathbf{b}_{t_j} \quad \forall t_j \in \mathcal{I}_r$$

$$\boldsymbol{\lambda}_{\min}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \geq p \text{ and } \boldsymbol{\lambda}_{\max}(\mathbf{L}\mathbf{L}^{\mathsf{T}}) \leq q$$

Solved via (1) a 2nd order cone program for $\mathbf{S}$ and (2) a semi-definite program for $\mathbf{L}$

Regret Guarantees

Theorem (Sublinear Regret):

After R rounds, our algorithm attains sublinear regret:

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

Proof Sketch:

Bound $\left\| \hat{\mathbf{f}}_{t_r} \Big|_{\hat{\boldsymbol{\theta}}_r} \right\|_2^{\star}$

$$\left\| \hat{\mathbf{f}}_{t_r} \Big|_{\hat{\boldsymbol{\theta}}_r} \right\|_2 \leq \underbrace{\left| l_{t_r}(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_r} \right|}_{\text{constant}} + \underbrace{\left\| \nabla_{\boldsymbol{\theta}} l_{t_r}(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_r} \right\|_2}_{\text{bounded in terms of local losses}} + \left\| \nabla_{\boldsymbol{\theta}} l_{t_r}(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_r} \right\|_2 \underbrace{\left\| \hat{\boldsymbol{\theta}}_r \right\|_2}_{\text{constraints}}$$

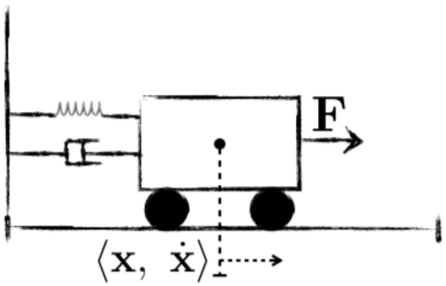
constant

bounded in terms
of local losses

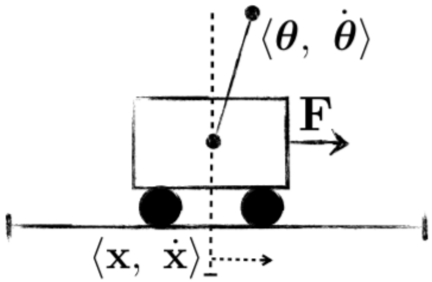
constraints

Experiments

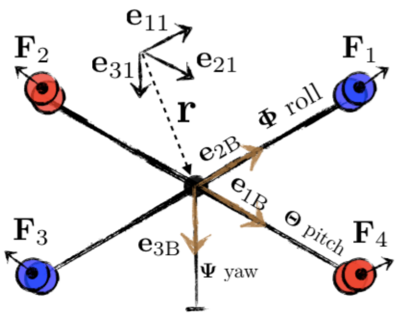
Goal: Learn policies for consecutive control tasks on three types of dynamical systems



Simple Mass



Cart Pole



Quadrotor

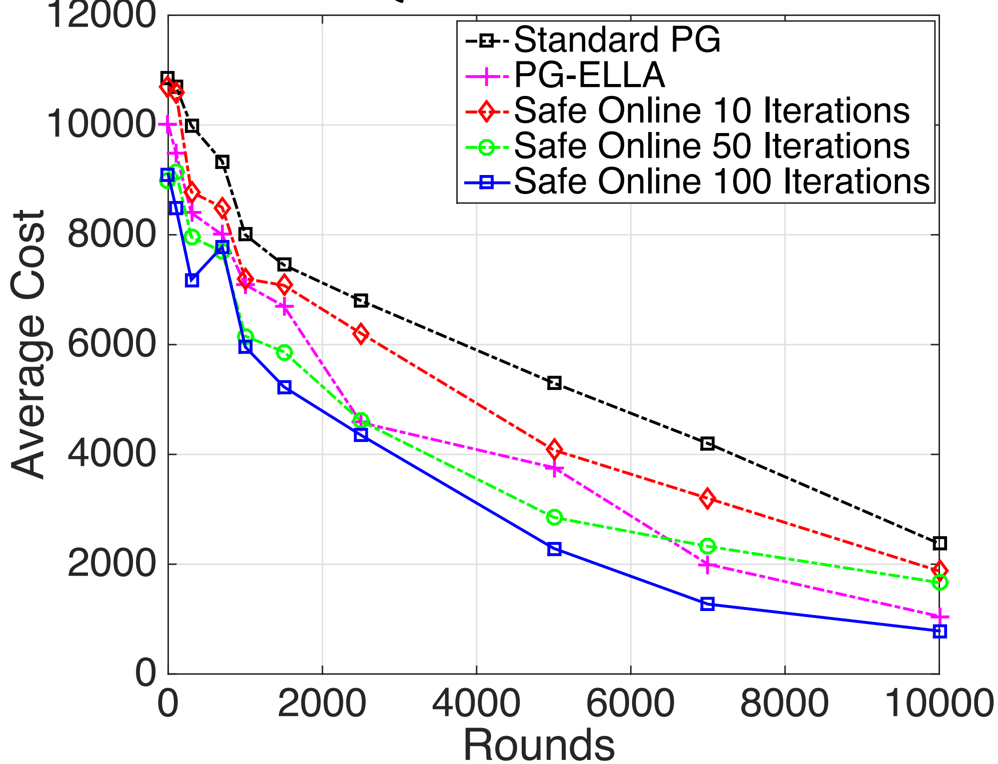
Generated 10 tasks per system by varying specifications

Compared to (1) standard PG and

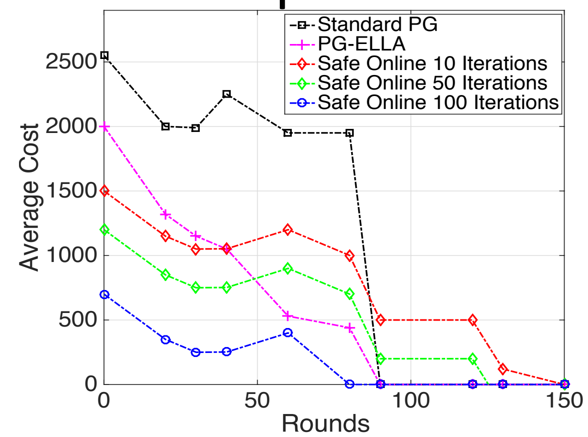
(2) PG-ELLA lifelong learner [Bou Ammar et al, ICML'14]

Results: Performance

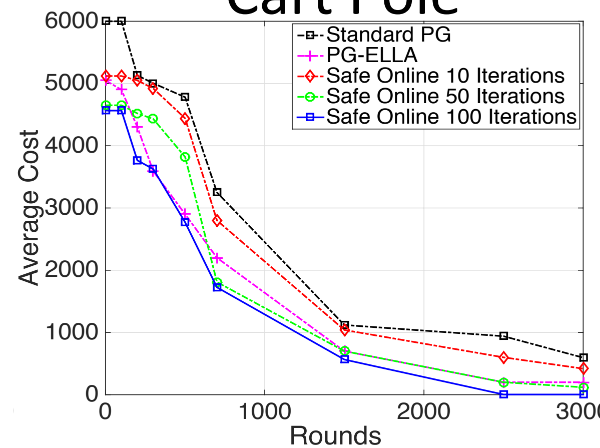
Quadrotor



Simple Mass



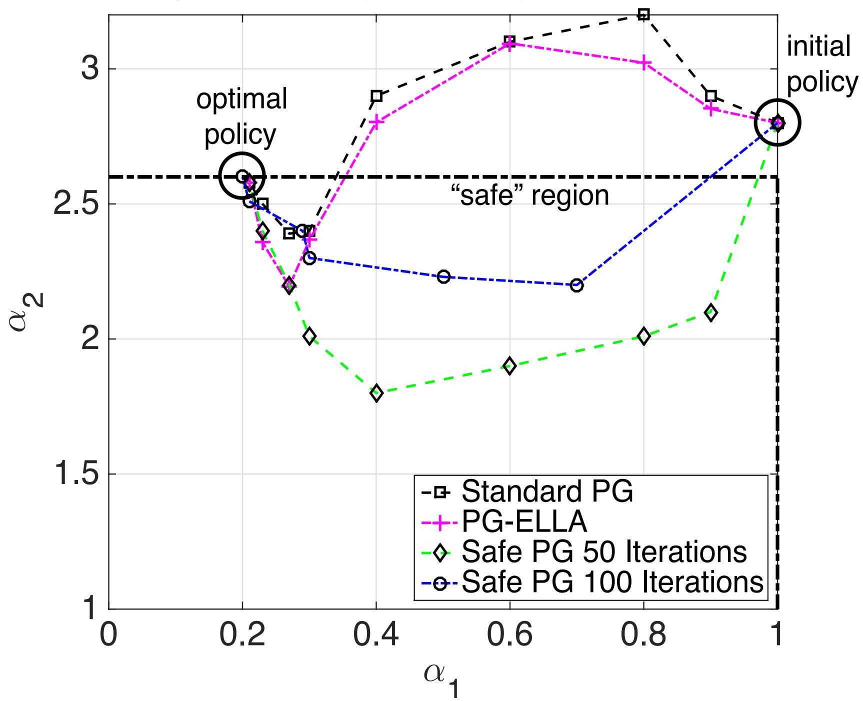
Cart Pole



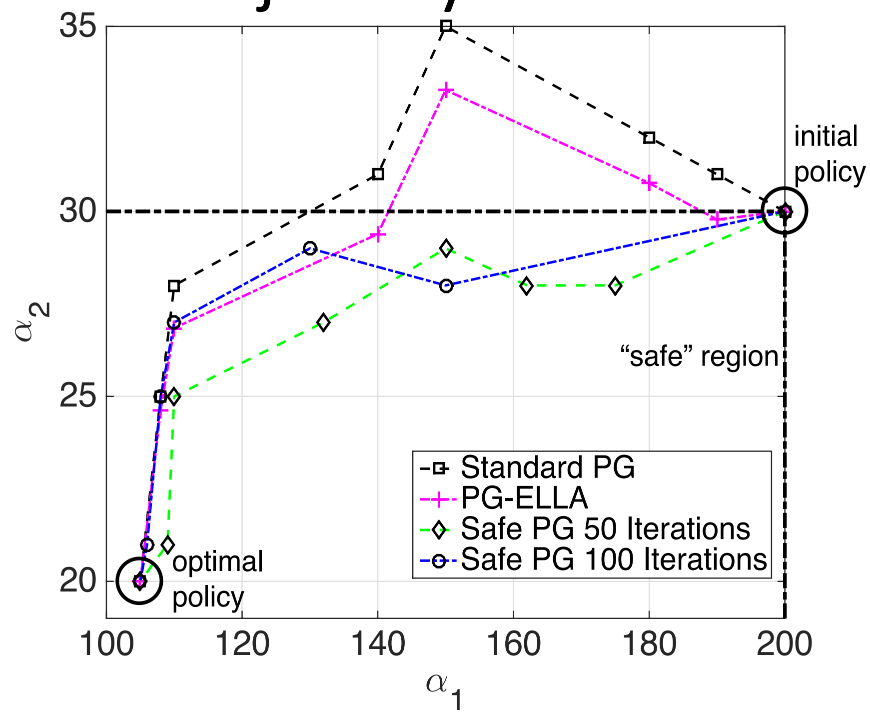
Safe lifelong learner shows superior performance

Results: Safety Constraint Enforcement

Trajectory Simple Mass

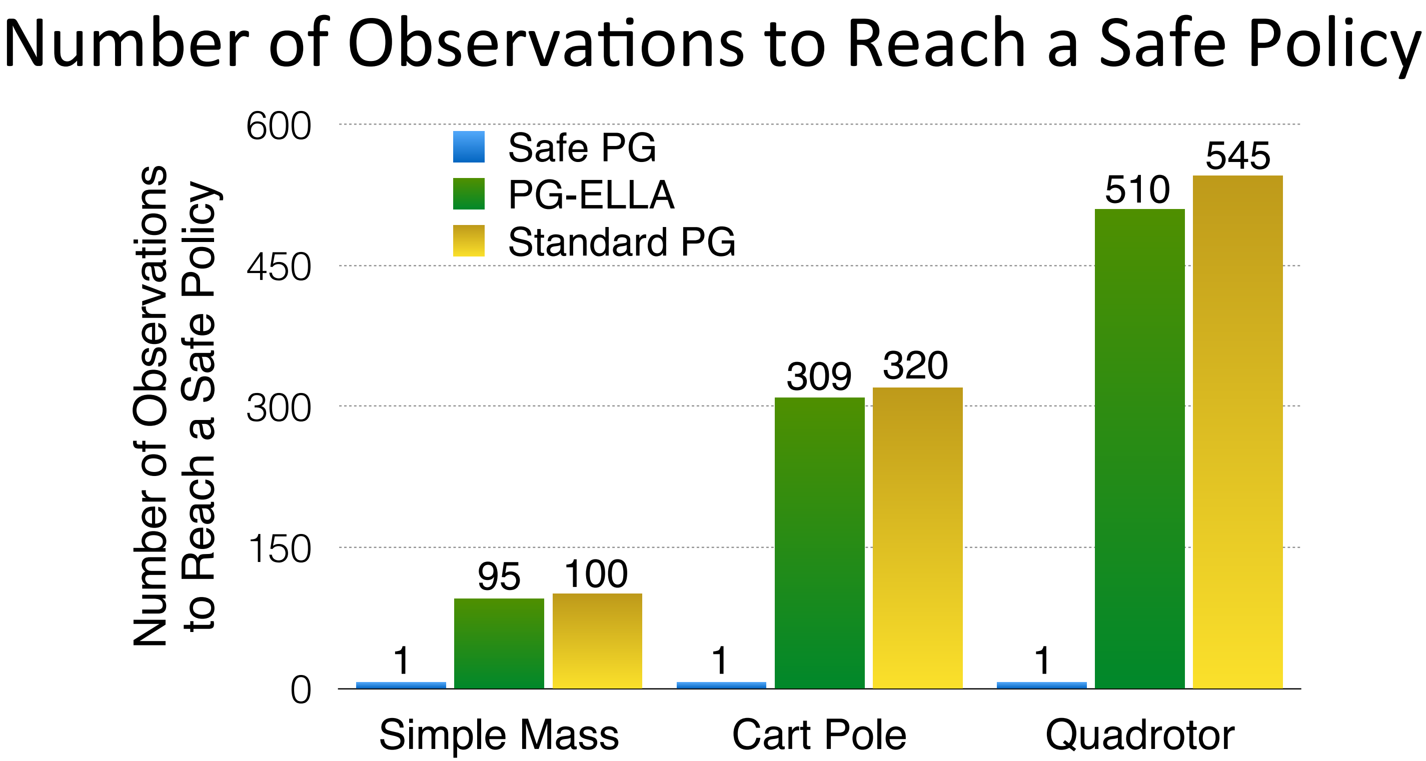


Trajectory Cart Pole



Enforces safety constraints, unlike alternative methods

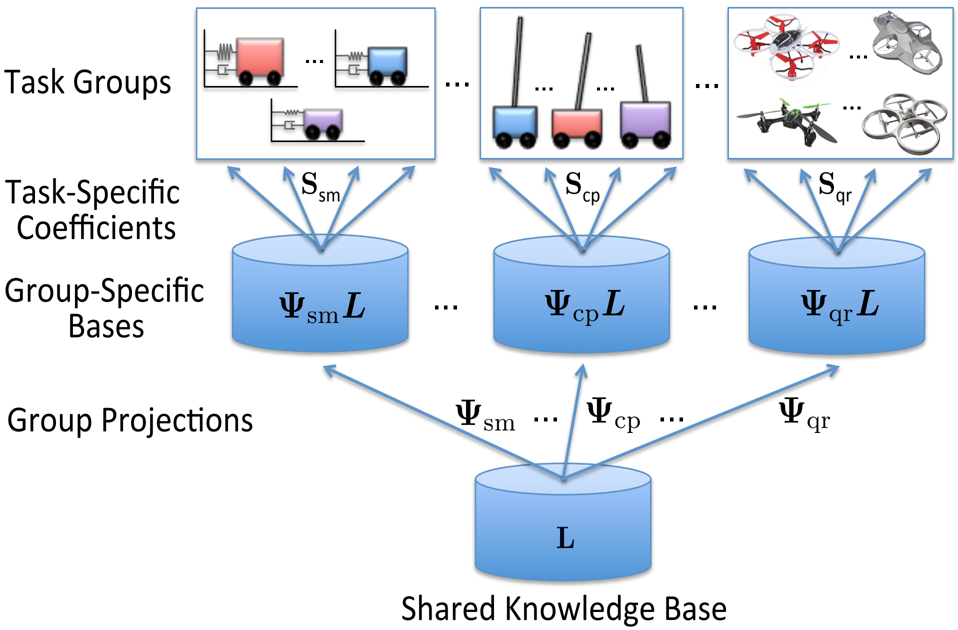
Results: Safety Constraint Enforcement



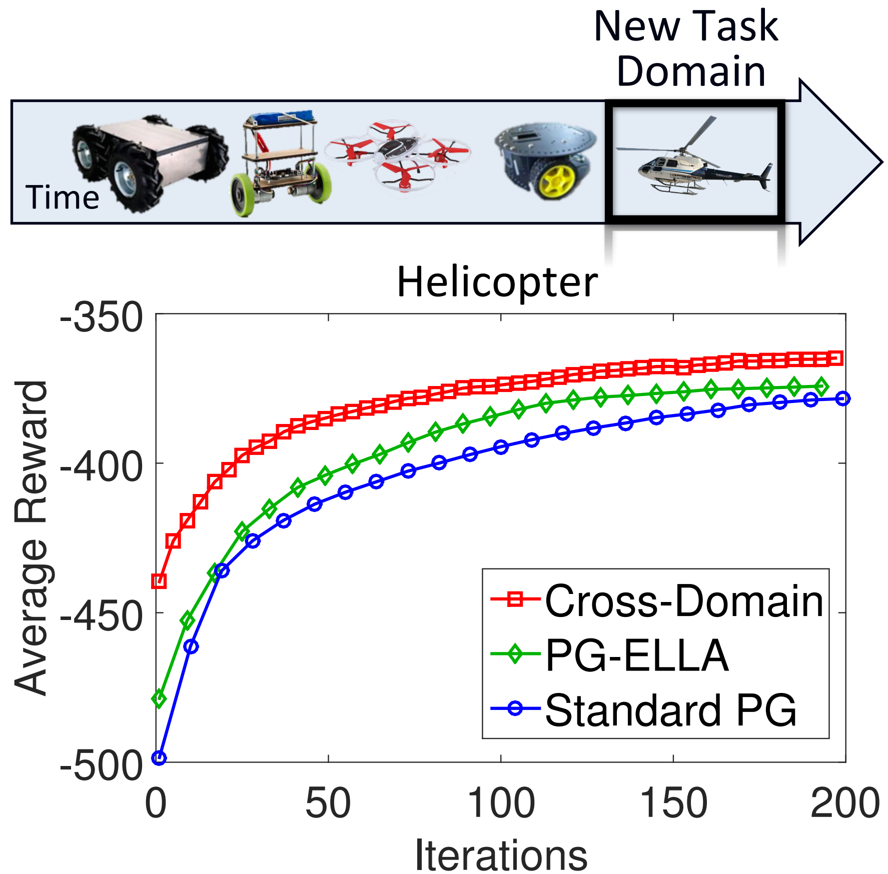
Our approach immediately projects policies to safe regions, even during the policy search process

Teaser: Autonomous Cross-Domain Transfer

Key Idea: Use projections to specialize a shared KB to individual task domains for lifelong RL



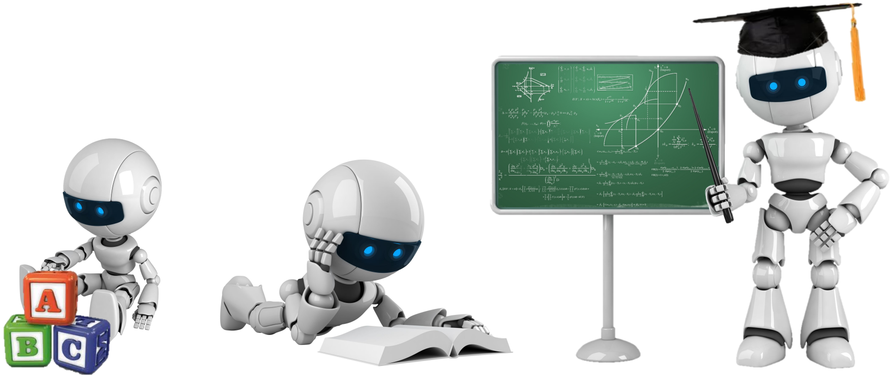
[Bou Ammar, Eaton, et al., IJCAI'15]



Conclusion

The safe lifelong policy gradient learner:

- 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



Thank you!

Questions?

Haitham Bou Ammar
haithamb@seas.upenn.edu

Rasul Tutunov
tutunov@seas.upenn.edu

Eric Eaton
eeaton@cis.upenn.edu

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.

Backup Slides

Constrained Solution

Alternate to determine safety-constrained $\mathbf{L}$ and $\mathbf{S}$:

Semi-Definite Program for $\mathbf{L}$:

$$\min_{\boldsymbol{X} \subset \mathcal{S}_{++}} \mu_2 \text{trace}(\boldsymbol{X}) + 2\mu_2 \left|\left| \boldsymbol{L}\Big|_{\tilde{\boldsymbol{\theta}}_{r+1}} \right|\right|_{\mathsf{F}} \sqrt{\text{trace}\left(\boldsymbol{X}\right)}$$

$$\text{s.t.} \quad \boldsymbol{s}_{t_j}^{\mathsf{T}} \boldsymbol{X} \boldsymbol{s}_{t_j} = \boldsymbol{a}_{t_j}^{\mathsf{T}} \boldsymbol{a}_{t_j} \quad \forall t_j \in \mathcal{I}_r$$

$$\boldsymbol{X} \leq p\boldsymbol{I} \quad \text{and} \quad \boldsymbol{X} \geq q\boldsymbol{I}\,, \quad \text{with} \quad \boldsymbol{X} = \boldsymbol{L}^{\mathsf{T}}\boldsymbol{L}$$

Second-Order Cone Program for $\mathbf{S}$:

$$\min_{\boldsymbol{s}_{t_1},\ldots,\boldsymbol{s}_{t_j},\boldsymbol{c}_{t_1},\ldots,\boldsymbol{c}_{t_j}} \mu_1 \sum_{j=1}^{r} \|\boldsymbol{s}_{t_j}\|_2^2 + 2\mu_1 \sum_{j=1}^{r} \boldsymbol{s}_{t_j}^{\mathsf{T}}\Big|_{\hat{\boldsymbol{\theta}}_r} \boldsymbol{s}_{t_j}$$

$$\text{s.t.} \quad \boldsymbol{A}_{t_j} \boldsymbol{L} \boldsymbol{s}_{t_j} = \boldsymbol{b}_{t_j} - \boldsymbol{c}_{t_j}$$

$$\boldsymbol{c}_{t_j} > 0 \quad \|\boldsymbol{c}_{t_j}\|_2^2 \leq \boldsymbol{c}_{\max}^2 \; \forall t_j \in \mathcal{I}_r \;.$$