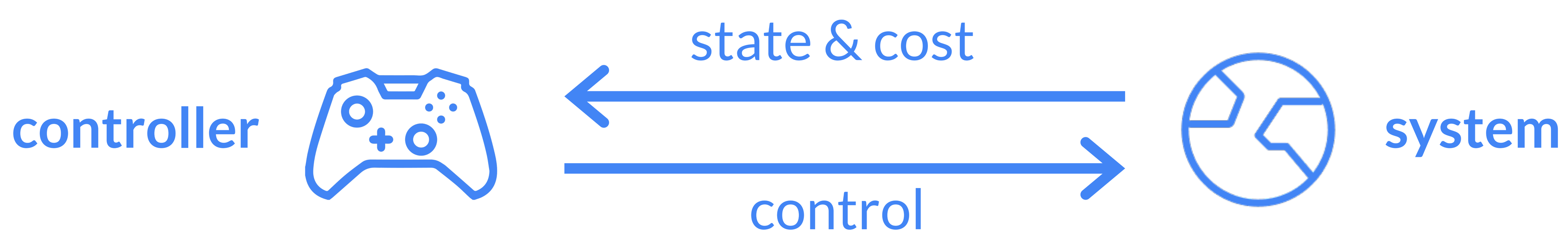
Implicit Bias of Policy Gradient in Linear Quadratic Control: Extrapolation to Unseen Initial States

Noam Razin*, Yotam Alexander*, Edo Cohen-Karlik, Raja Giryes, Amir Globerson, Nadav Cohen | *Equal Contribution

Have Only 2 Minutes? Read This

Setting: Policy Gradient (PG) for Optimal Control

Optimal Control (equivalent to Reinforcement Learning): Learn controller that minimizes cost over a dynamical system



Policy Gradient (PG): Parameterize controller (e.g. as neural network) and minimize cost via gradient descent

Issue of Prime Importance: Extrapolation to initial states unseen in training

Implicit Bias: Often multiple controllers minimize the training cost, so extrapolation is determined by an implicit bias of PG

Main Question

? How does the **implicit bias** of PG affect **extrapolation to initial states unseen in training?**

Theory for the Linear Quadratic Regulator (LQR)

Extrapolation is determined by exploration induced by the system from initial states that are seen in training

Experiments

Support theory for LQR and demonstrate its conclusions on **non-linear systems and neural network controllers**

Policy Gradient (PG) for the Linear Quadratic Regulator (LQR)

Linear System

$\mathbf{x}_{h+1} = \mathbf{A}\mathbf{x}_h + \mathbf{B}\mathbf{u}_h$

$\mathbf{x}_h \in \mathbb{R}^D$ - state , $\mathbf{u}_h \in \mathbb{R}^M$ - control

Quadratic Cost

$\sum_{h=0}^{H} \mathbf{x}_h^\top \mathbf{Q}\mathbf{x}_h + \mathbf{u}_h^\top \mathbf{R}\mathbf{u}_h$

H - horizon

Linear Controller

$\mathbf{u}_h = \mathbf{K}\mathbf{x}_h$

PG Training

Run gradient descent over cost for training set of initial states $\mathcal{S}$: $cost_{\mathcal{S}}(\mathbf{K}) = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}_0 \in \mathcal{S}} \sum_{h=0}^{H} \mathbf{x}_h^\top (\mathbf{Q} + \mathbf{K}^\top \mathbf{R}\mathbf{K})\mathbf{x}_h$

We study a practically motivated setting where multiple controllers minimize the training cost, and they differ in their extrapolation

Quantifying Extrapolation

Optimality Condition

Controller $\mathbf{K}$ minimizes the training cost if and only if $\underbrace{\|(\mathbf{A}+\mathbf{B}\mathbf{K})\mathbf{x}_0\|^2 = 0}_{\mathbf{K} \text{ sends } \mathbf{x}_0 \text{ to zero}}, \forall \mathbf{x}_0 \in \mathcal{S}$

Extrapolation Error

$\mathcal{E}(\mathbf{K}) := \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_0 \in \mathcal{U}} \|(\mathbf{A}+\mathbf{B}\mathbf{K})\mathbf{x}_0\|^2$

Measures suboptimality on a basis $\mathcal{U}$ of $\mathcal{S}^\perp$ (unseen subspace)

Baseline Non-Extrapolating Controller

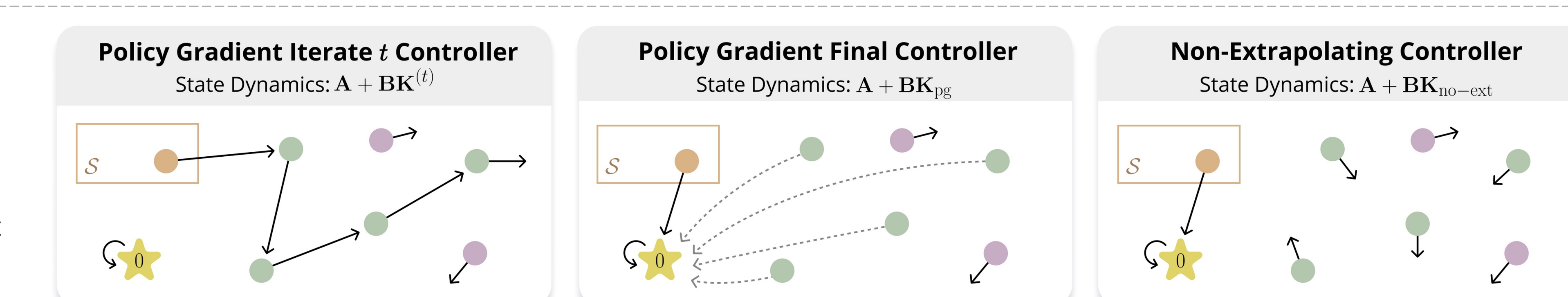
$\mathbf{K}_{\text{no-ext}} \begin{cases} \text{sends states in } \mathcal{S} \text{ to zero} \\ \text{assigns null controls to states in } \mathcal{U} \end{cases}$

minimizes training cost but has high extrapolation error

Theory: Extrapolation is Determined by Exploration

Intuition: Extrapolation is determined by exploration induced by the system from initial states seen in training

- initial state seen in training
- state explored during policy gradient
- state unexplored during policy gradient



Notation: $\mathbf{K}_{\text{pg}}$ - controller learned via PG , lr - learning rate of PG , D - state space dimension , H - horizon

Proposition

Extrapolation Requires Exploration

- For states orthogonal to those reached during PG, $\mathbf{K}_{\text{pg}}$ and $\mathbf{K}_{\text{no-ext}}$ produce identical controls
- There exist non-exploratory systems in which: $\mathcal{E}(\mathbf{K}_{\text{pg}}) = \mathcal{E}(\mathbf{K}_{\text{no-ext}})$

Proposition

Extrapolation in Exploration-Inducing Setting

There exist exploration-inducing settings in which PG leads to substantial extrapolation:

$$\mathcal{E}(\mathbf{K}_{\text{pg}}) << \mathcal{E}(\mathbf{K}_{\text{no-ext}})$$

*If the horizon H is infinite then $\mathcal{E}(\mathbf{K}_{\text{pg}}) = 0$

Theorem

Extrapolation in Typical Setting

When $\mathbf{A}$ is random Gaussian, a single step of PG already leads to non-trivial extrapolation:

$$\mathbb{E}[\mathcal{E}(\mathbf{K}_{\text{pg}})] \leq \mathbb{E}[\mathcal{E}(\mathbf{K}_{\text{no-ext}})] - \Omega(\text{lr} \cdot \tfrac{H^2}{D})$$

*Extrapolation occurs w.h.p. if D is large

Experiments with Non-Linear Systems and Neural Network Controllers

Our Theory: If a linear system induces exploration from initial states seen in training, then a linear controller typically extrapolates

Experiments: Phenomenon extends to non-linear systems with neural network controllers!

Pendulum Control

(analogous experiments for a quadcopter control problem)

- target state
- initial state seen in training
- initial state unseen in training