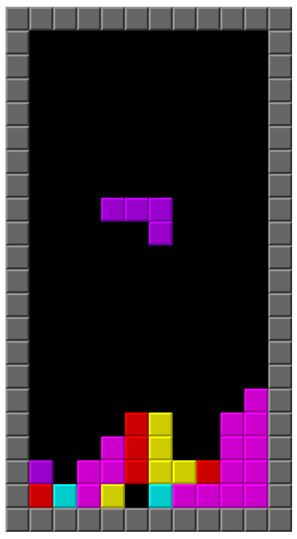
Contrasting Exploration in Parameter and Action Space

Anirudh Vemula, Wen Sun, Drew Bagnell
vemula@cmu.edu, {wensun,dbagnell}@cs.cmu.edu

Carnegie Mellon University
THE ROBOTICS INSTITUTE

Empirical Successes of Black-Box Policy Search



- Most of the success in tetris is by blackbox policy search
- Cross-entropy method achieves a score of **35 million lines!**
- Whereas, natural policy gradient gets **7000 lines**
- Gabillon et. al. 2013

- Very successful in obtaining good policies for diverse robotic tasks such as locomotion and helicopter control
- Tesch et. al. 2011, Bagnell and Schneider 2003

Exploration in Model-Free Policy Search

Parameter Space Exploration

- Find a direction of improvement directly in parameter space through random exploration
- Purely zeroth order approach
- Eg: Cross-entropy method, Evolutionary strategies, Augmented Random search etc.

Action Space Exploration

- Find a direction of improvement in action space through random exploration
- **Leverage Jacobian of policy** to update parameters
- A combination of zeroth and first order approach
- Eg: REINFORCE and its extensions

$$\nabla_\theta J(\theta) = \nabla_\pi J(\theta) \nabla_\theta \pi$$

Directly estimate using a zeroth order approach e.g. finite differencing

Jacobian of the policy

Estimate using a zeroth order approach

Analysis

	Linear Contextual Bandit	Model-Free RL
Parameter space	$\mathcal{O}(\frac{d^2}{\epsilon^2})$	$\mathcal{O}\left(\frac{d^2 Q \sigma^3}{\epsilon^3}\right)$
Action space	$\mathcal{O}(\frac{1}{\epsilon^4})$	$\mathcal{O}(\frac{p^2 H^4}{\epsilon^4}(Q^3 + \sigma^2 Q))$

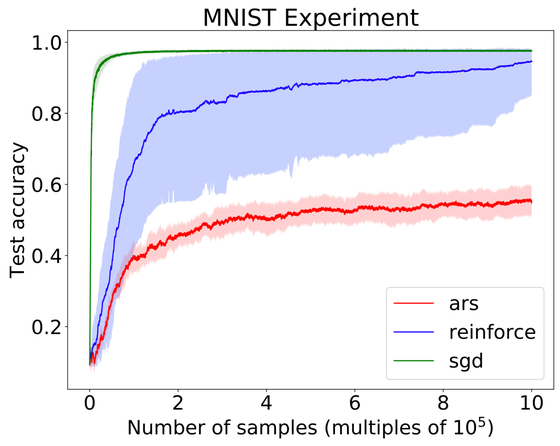
Linear Contextual Bandit : $\text{Avg. Regret} = \frac{1}{T}(\mathbb{E}[\sum_{t=1}^{T} c_i(\theta_i)] - \min_\theta \sum_{t=1}^{T} c_i(\theta))$

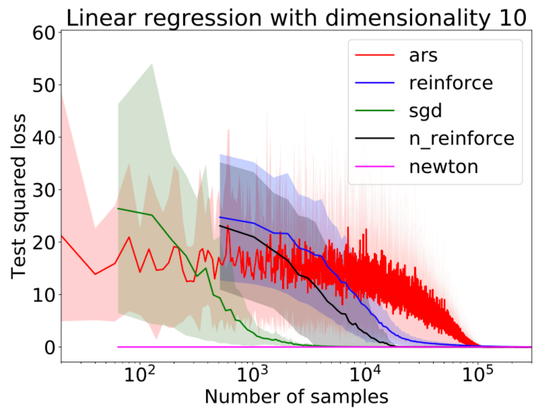
Model-Free RL : $\|\nabla_\theta J(\theta)\|_2^2 \leq \epsilon$ eps-stationary point

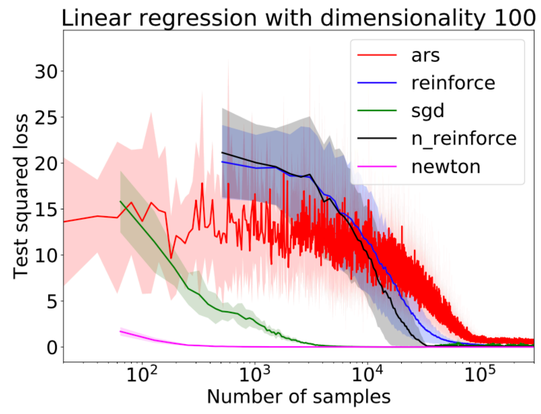
Dependence on parameter dimensionality
Independent of horizon length

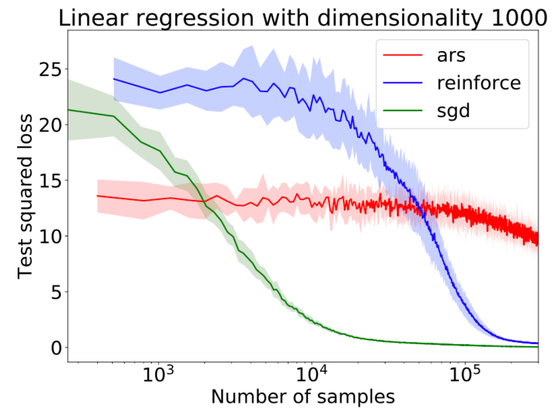
Dependence on horizon length
Dependence on action dimensionality
Independent of parameter dimensionality

Experiments









- Our analysis explains the success of black-box policy search methods like random search and evolutionary strategies in RL *(OpenAI gym tasks have very long horizons)*
- **For tasks with long horizons, exploration in parameter space should be preferred**
- **If the parametric complexity required is large, exploration in action space is better**