On the model-based stochastic value gradient for continuous reinforcement learning

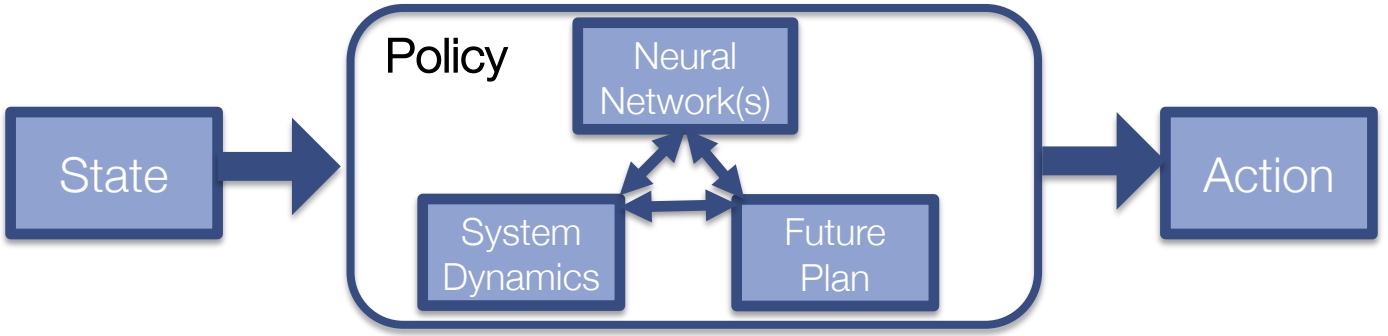
Brandon Amos[1] Samuel Stanton[2] Denis Yarats[1,2] Andrew Gordon Wilson[2]

[1]Facebook AI Research [2]NYU

L4DC 2021

Focus: Model-based methods for continuous control

Budding with methods and applications



PILCO (Deisenroth and Rasmussen, 2011)
MVE (Feinberg et al., 2018)
STEVE (Buckman et al., 2018)
IVG (Byravan et al., 2019)
Dreamer (Hafner et al., 2019)

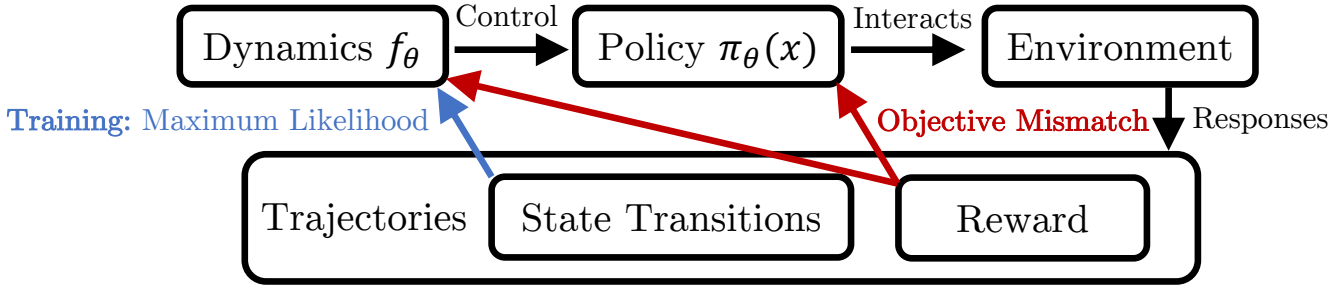
GPS (Levine and Koltun, 2013)
POPLIN (Wang and Ba, 2019)

METRPO (Kurutach et al., 2018)
MBPO (Janner et al., 2019)

(non-exhaustive)

But rife with problems Spoiler: We don't solve these

E.g., objective mismatch, short-horizon bias, inaccurate models, accumulating errors



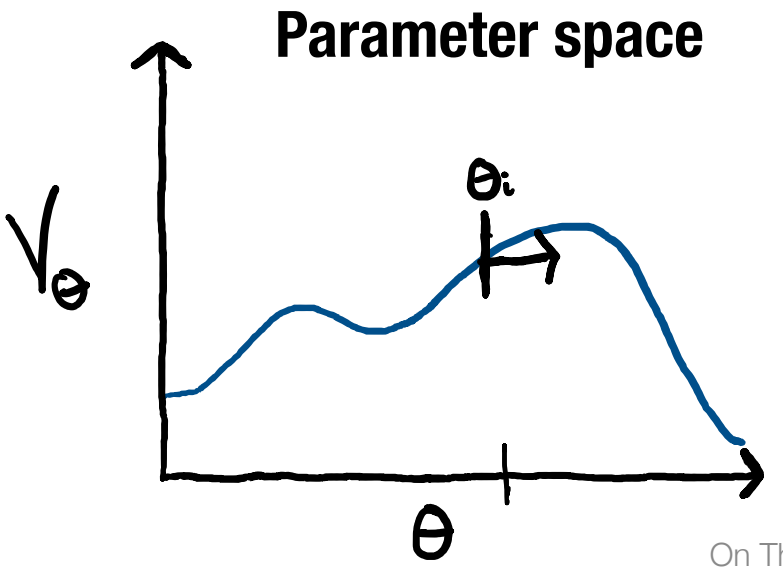
Background: The Stochastic Value Gradient

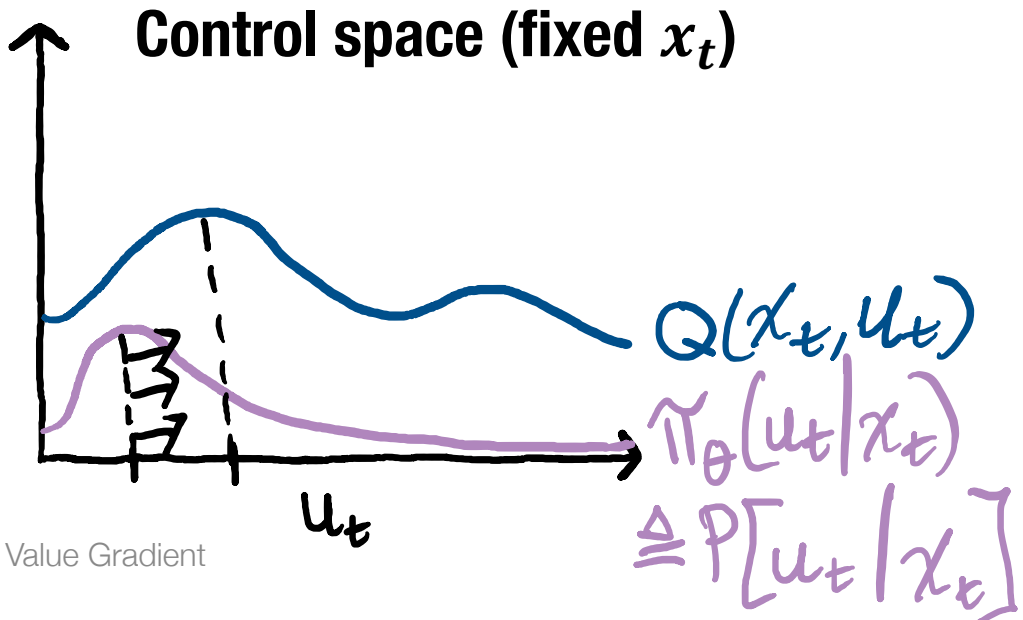
Heess et al. (NeurIPS 2015)

Learn a **stochastic** policy π_θ with **gradient** of a **value estimate**:

$$\nabla_\theta \mathbb{E}_{x_t}[V_\theta(x_t)]$$

Value estimate can be **model-based**, **model-free**, or **both**
Can also add an **entropy penalty** to the actions





Background: The Soft Actor-Critic

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine

Model-free deep reinforcement learning (RL) algorithms have been demonstrated on a range of challenging decision making and control tasks. However, these methods typically suffer from two major challenges: very high sample complexity and brittle convergence properties, which necessitate meticulous hyperparameter tuning. Both of these challenges severely limit the applicability of such methods to complex, real-world domains. In this paper, we propose soft actor-critic, an off-policy actor-critic deep RL algorithm based on the maximum entropy reinforcement learning framework. In this framework, the actor aims to maximize expected reward while also maximizing entropy. That is, to succeed at the task while acting as randomly as possible. Prior deep RL methods based on this framework have been formulated as Q-learning methods. By combining off-policy updates with a stable stochastic actor-critic formulation, our method achieves state-of-the-art performance on a range of continuous control benchmark tasks, outperforming prior on-policy and off-policy methods. Furthermore, we demonstrate that, in contrast to other off-policy algorithms, our approach is very stable, achieving very similar performance across different random seeds.

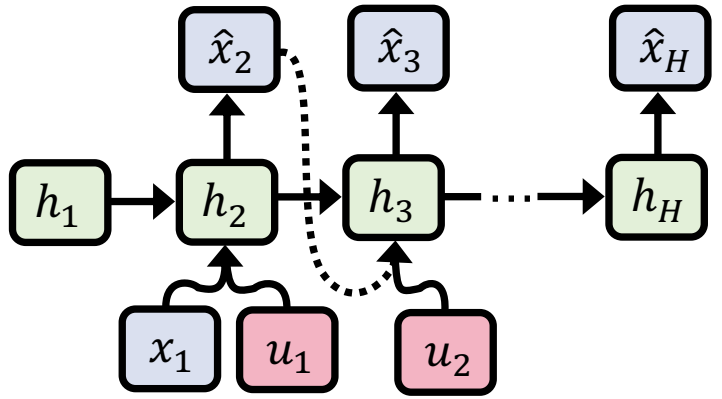
SAC-SVG(H): A model-based SAC extension

Observation (S4.1 of our paper). The soft actor-critic (SAC) policy update is just a value gradient with a **model-free value estimate** and **entropy regularization**

Idea: Replace SAC's **value estimate** with a **more accurate model-based expansion** for H steps.

	SAC-SVG(0): equivalent to SAC Model-free value estimate	**SAC-SVG(H)** when H>0 Model-based value expansion with model-free estimate at the end
Value Estimate $V_\theta(x_t)$	V_t	$r_t \rightarrow r_{t+1} \rightarrow \cdots \rightarrow V_{t+H}$
Policy update	The stochastic value gradient $\nabla_\theta \mathbb{E}_{x_t}[V_\theta(x_t)]$	

Also use a **simple** recurrent **dynamics model** (no ensembling) trained for **multi-step predictions**



SAC-SVG excels in locomotion tasks

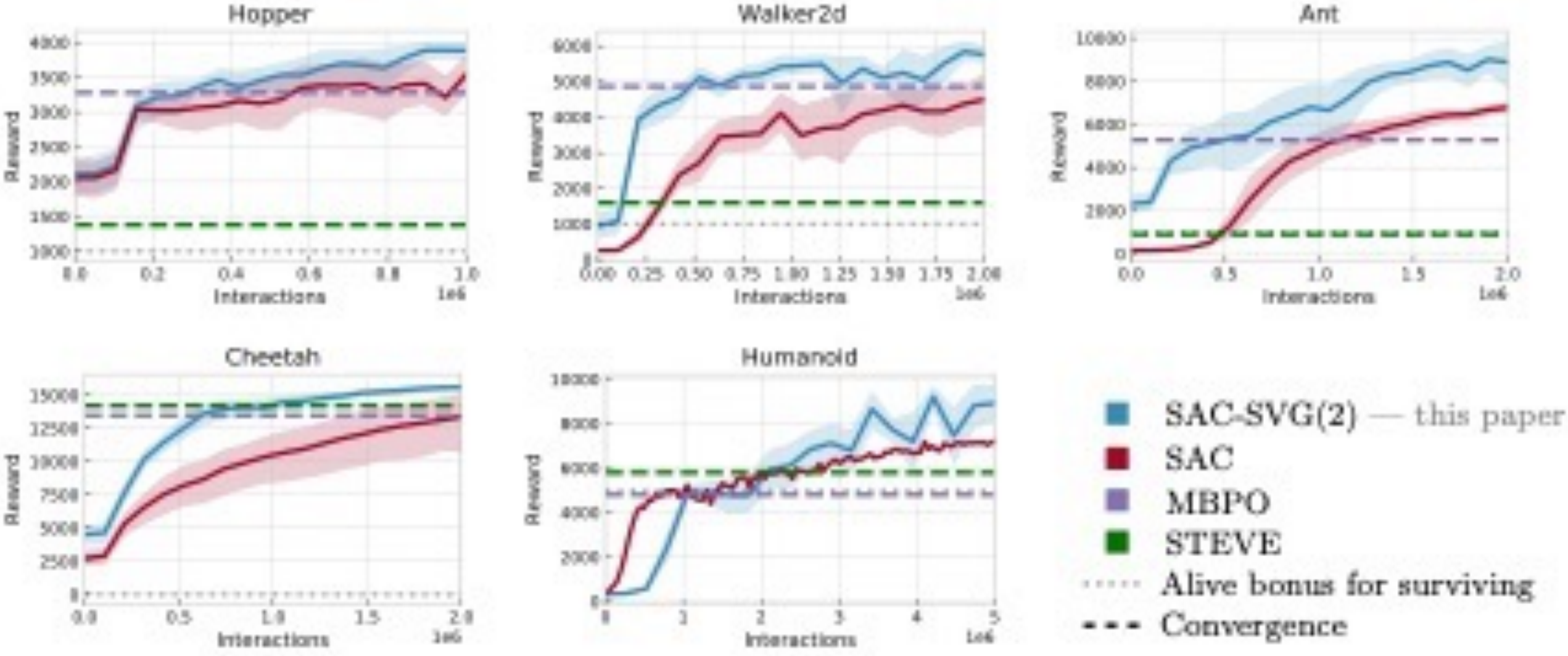
Evaluation Rewards

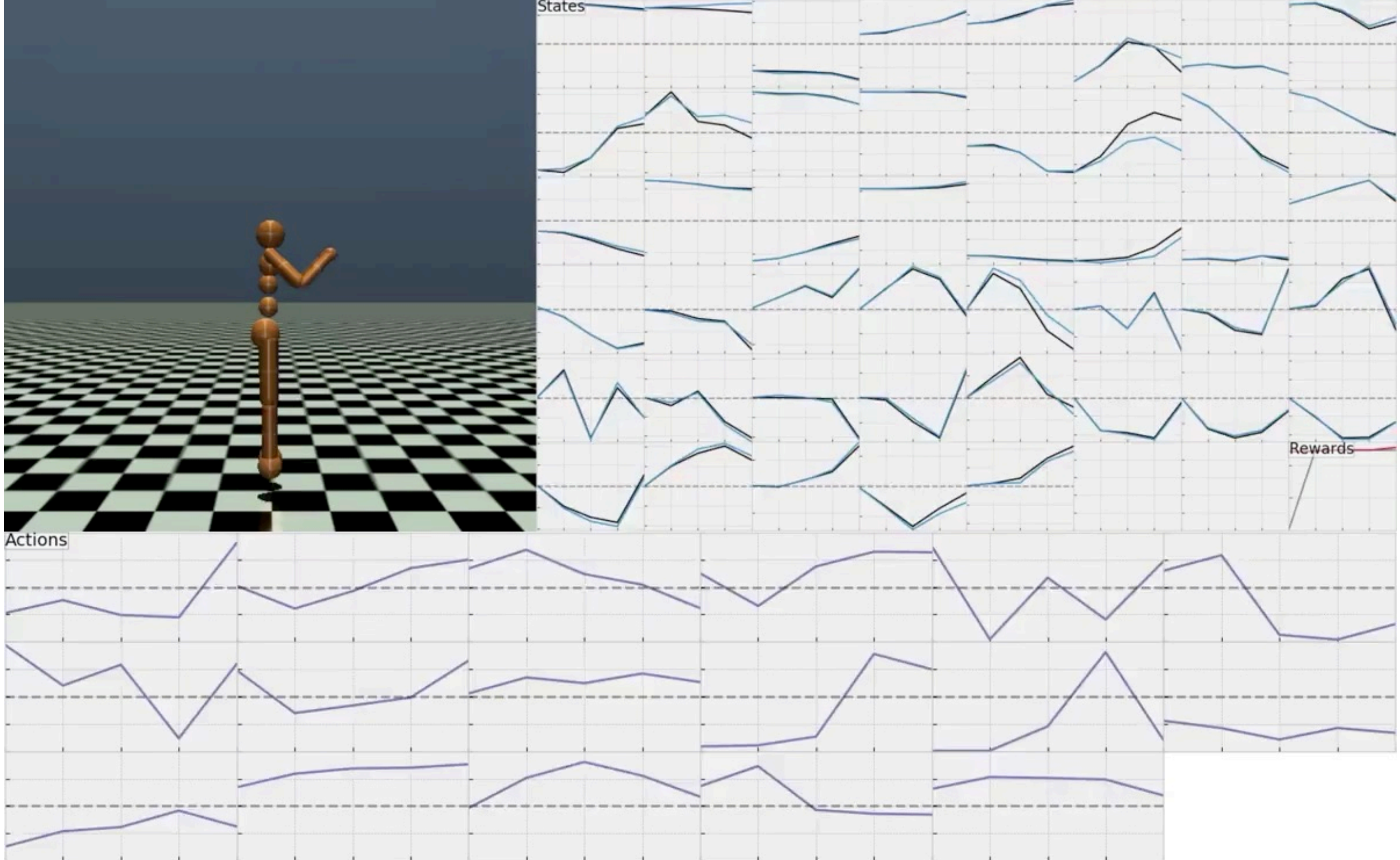
this paper

	Ant	Hopper	Swimmer	Cheetah	Walker2d	PETS Cheetah
SAC-SVG(1)	3691.00 ± 1096.77	1594.43 ± 1689.01	348.40 ± 8.32	6890.20 ± 1860.49	-291.54 ± 659.52	5321.23 ± 1507.00
SAC-SVG(2)	4473.36 ± 893.44	2851.90 ± 361.07	350.22 ± 3.63	8751.76 ± 1785.66	447.68 ± 1139.51	5799.59 ± 1266.93
SAC-SVG(3)	3833.12 ± 1418.15	2024.43 ± 1981.51	340.75 ± 13.46	9220.39 ± 1431.77	877.77 ± 1533.08	5636.93 ± 2117.52
SAC-SVG(4)	2896.77 ± 1444.40	2062.16 ± 1245.33	348.03 ± 6.35	8175.29 ± 3226.04	1852.18 ± 967.61	5807.69 ± 1008.60
SAC-SVG(5)	3221.66 ± 1576.25	608.58 ± 2105.60	340.99 ± 4.58	6129.02 ± 3519.98	1309.20 ± 1281.76	4896.22 ± 1033.33
SAC-SVG(10)	1389.30 ± 981.59	-2511.05 ± 881.26	303.16 ± 10.57	2477.25 ± 2596.43	-2328.08 ± 735.48	4248.25 ± 802.54
POPLIN-P (Wang and Ba, 2019)	2330.1 ± 320.9	2055.2 ± 613.8	334.4 ± 34.2	4235.0 ± 1133.0	597.0 ± 478.8	12227.9 ± 5652.8
SAC* (Haarnoja et al., 2018)	548.1 ± 146.6	788.3 ± 738.2	204.6 ± 69.3	3459.8 ± 1326.6	164.5 ± 1318.6	1745.9 ± 839.2
SAC (our run)	510.56 ± 76.38	2180.33 ± 977.30	351.24 ± 5.27	6514.83 ± 1100.61	1265.13 ± 1317.00	3259.99 ± 1219.94
PETS* (Chua et al., 2018)	1165.5 ± 226.9	114.9 ± 621.0	326.2 ± 12.6	2288.4 ± 1019.0	282.5 ± 501.6	4204.5 ± 789.0
METRPO* (Kurutach et al., 2018)	282.2 ± 18.0	1272.5 ± 500.9	225.5 ± 104.6	2283.7 ± 900.4	-1609.3 ± 657.5	-744.8 ± 707.1
TD3* (Fujimoto et al., 2018c)	870.1 ± 283.8	1816.6 ± 994.8	72.1 ± 130.9	3015.7 ± 969.8	-516.4 ± 812.2	218.9 ± 593.3
Training Timesteps	200000	200000	50000	200000	200000	50000

*Denotes the baseline results reported in Wang and Ba (2019).

SAC-SVG excels in locomotion tasks




States
Rewards
Actions

Discussion

1. Is **epistemic uncertainty** important for **exploration** in Mujoco environments?
   A **target action entropy schedule** worked well for us
2. Should the **value expansion** be used for actor and/or critic updates?
   We found actor updates to be the best
3. **Model becomes inaccurate** quickly
   Short horizons work best for us (and MBPO)
4. **Benchmarks and baselines** are challenging to do properly
   E.g., from papers tweaking environments or not re-tuning parameters

Concluding remarks

The **stochastic value gradient is competitive** with **short-horizon model rollouts** and **action entropy**
We attain these results with a simple deterministic **LSTM** for the **world model**

Key **future directions** include:

1. **Refinement** or **semi-amortization**
2. **Constrained MDPs**
3. **Model-based extensions** of other **model-free algorithms**
4. More sophisticated **world models**, e.g., self-supervised
5. Going **beyond the single-agent, single-task setting**

On the model-based stochastic value gradient for continuous reinforcement learning

Brandon Amos[1] Samuel Stanton[2] Denis Yarats[1,2] Andrew Gordon Wilson[2]

[1]Facebook AI Research [2]NYU

L4DC 2021

github.com/facebookresearch/svg

brandondamos samscub denisyarats andrewgwils

bamos.github.io samuelstanton.github.io cs.nyu.edu/~dy1042 cims.nyu.edu/~andrewgw