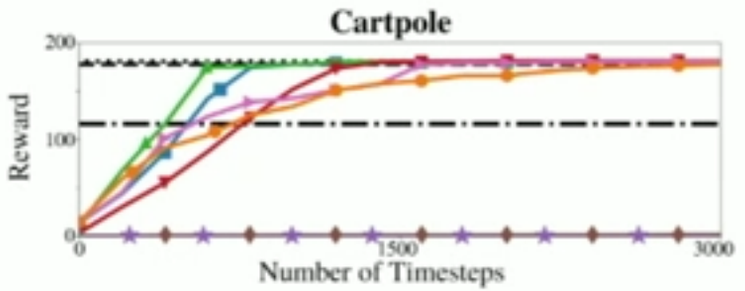
When to Trust Your Model: Model-Based Policy Optimization

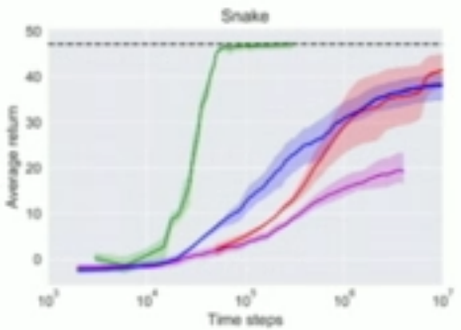
Michael Janner Justin Fu Marvin Zhang Sergey Levine
UC Berkeley

The problem of model bias

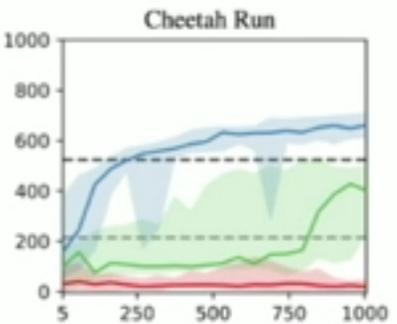
- Model-based reinforcement learning methods are fast!



PETS. Chua et al, 2018



ME-TRPO. Kurutach et al, 2018.



PlaNet. Hafner et al, 2019.

… but often have poor asymptotic performance on high-dimensional tasks due to modeling inaccuracies

- How can we best leverage a predictive model for learning a policy?

Outline

1. Model-based policy optimization (MBPO) algorithm
2. Monotonic improvement and model generalization in theory
3. Model generalization in practice
4. Experiments

Outline

1. Model-based policy optimization (MBPO) algorithm
2. Monotonic improvement and model generalization in theory
3. Model generalization in practice
4. Experiments

Outline

1. Model-based policy optimization (MBPO) algorithm
2. Monotonic improvement and model generalization in theory
3. Model generalization in practice
4. Experiments

Outline

1. Model-based policy optimization (MBPO) algorithm
2. Monotonic improvement and model generalization in theory
3. Model generalization in practice
4. Experiments

Algorithm 1 Model-Based Policy Optimization

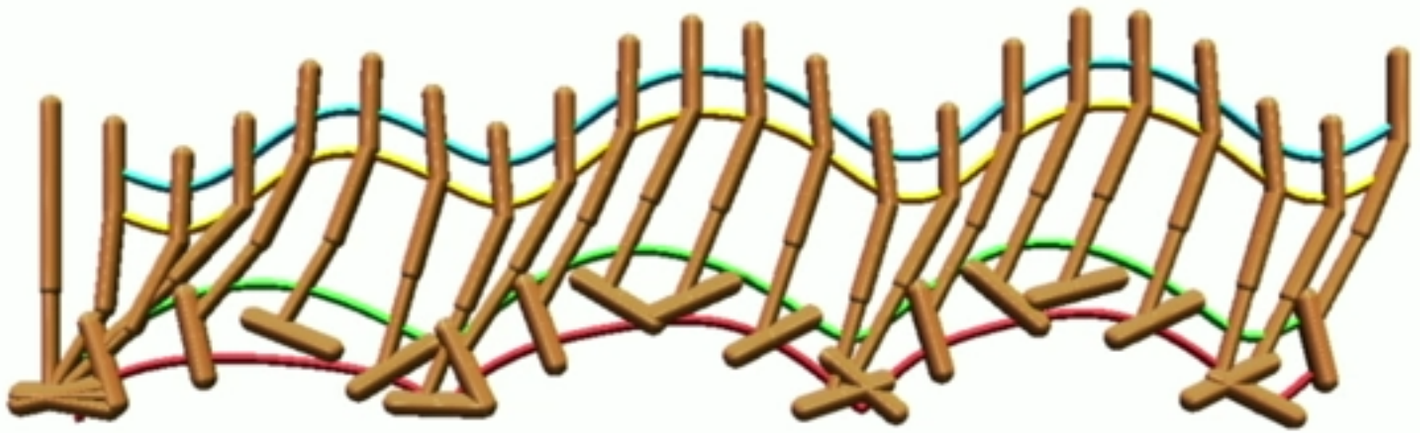
1: Initialize data-collecting policy π
2: **for** N epochs **do**
3: Collect data with π in real environment
4: Train model p_θ on real data
5: Optimize policy π under predictive model

Algorithm 1 Model-Based Policy Optimization

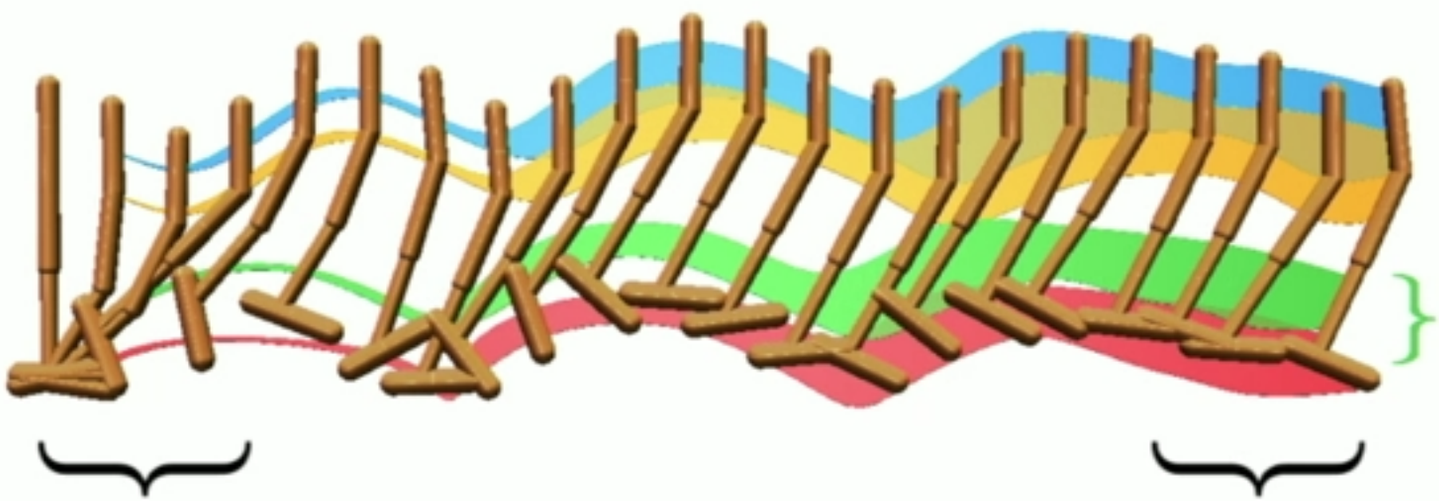
1: Initialize data-collecting policy π
2: **for** N epochs **do**
3: Collect data with π in real environment
4: Train model p_θ on real data
5: Optimize policy π under predictive model

Little guidance on how to properly optimize a policy with a model

The problem with long rollouts



Environment rollout



Model rollout (x1000)

} mean prediction +/- std dev

accurate, low variance

low accuracy, high variance

Bounding returns

- Use policy performance under model rollouts to derive lower bound in real environment

$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\Big[\underbrace{\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k+2)\epsilon_\pi}{(1-\gamma)}}_{\text{discrepancy due to policy shift}} + \underbrace{\frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)}_{\text{discrepancy due to model error}}\Big]$$

See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.

Bounding returns

- Use policy performance under **branched length-*k*** model rollouts to derive lower bound in real environment

$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\Big[\underbrace{\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k+2)\epsilon_\pi}{(1-\gamma)}}_{\text{discrepancy due to policy shift}} + \underbrace{\frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)}_{\text{discrepancy due to model error}}\Big]$$

See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.

Bounding returns

- Use policy performance under **branched length-*k*** model rollouts to derive lower bound in real environment

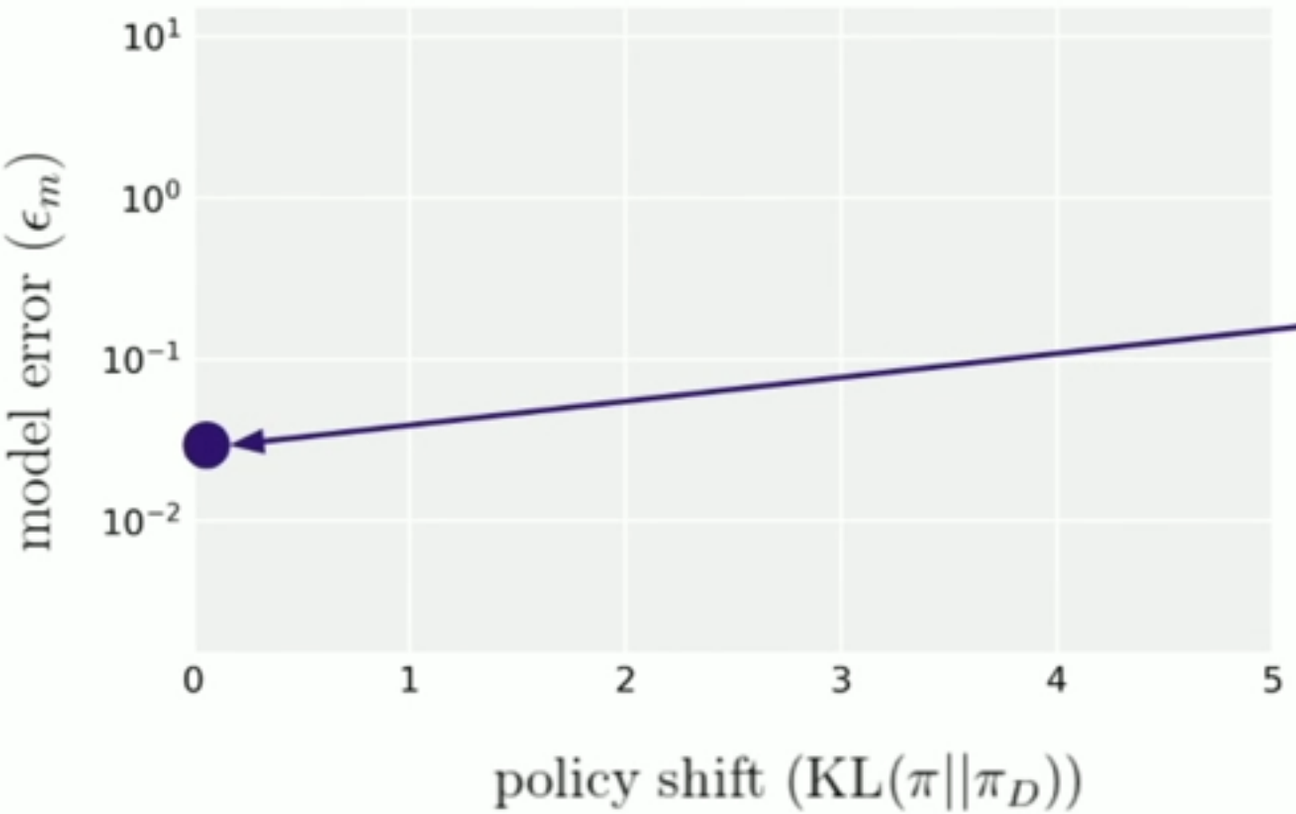
$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\Big[\underbrace{\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k+2)\epsilon_\pi}{(1-\gamma)}}_{\text{discrepancy due to policy shift}} + \underbrace{\frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)}_{\text{discrepancy due to model error}}\Big]$$

- Real off-policy data is always preferable to model-generated on-policy data!

  → *k* = 0?

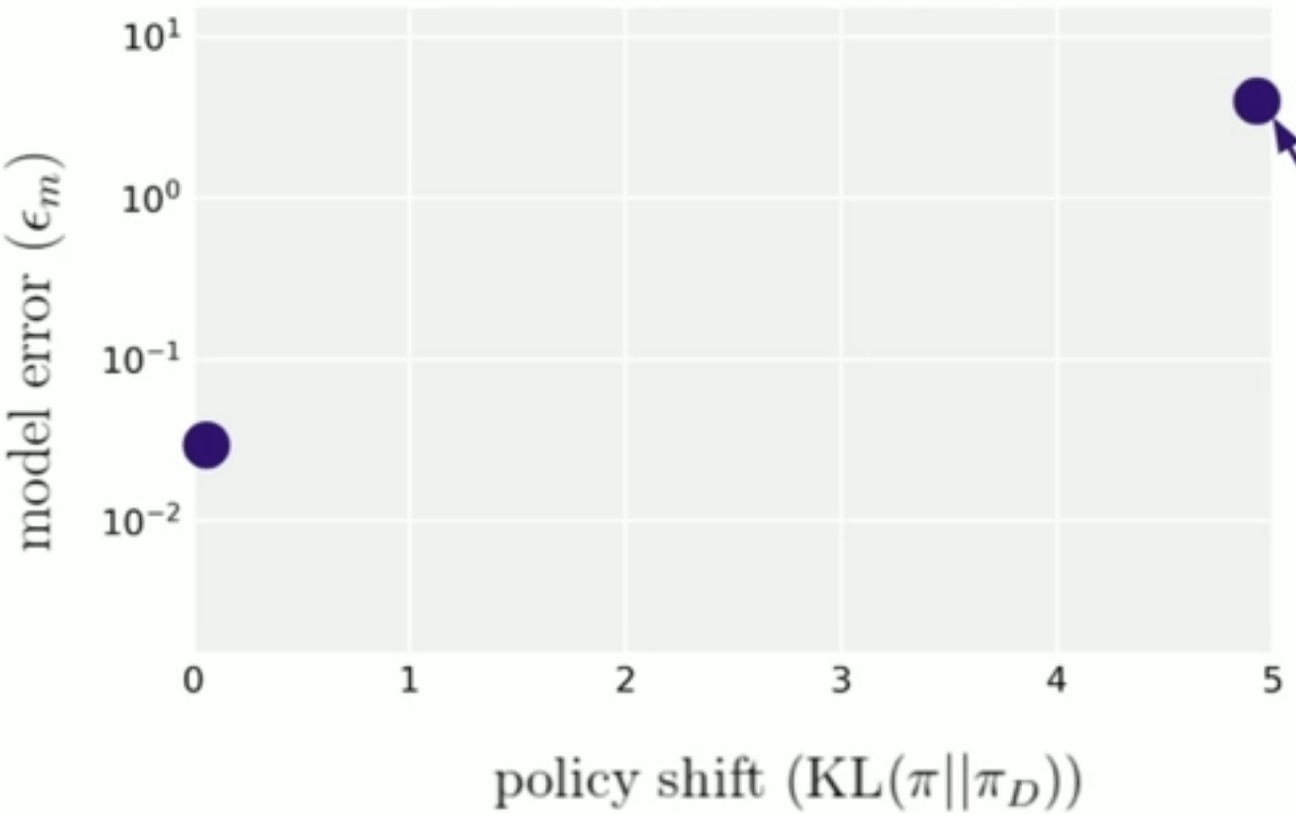
See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.

Model generalization in practice



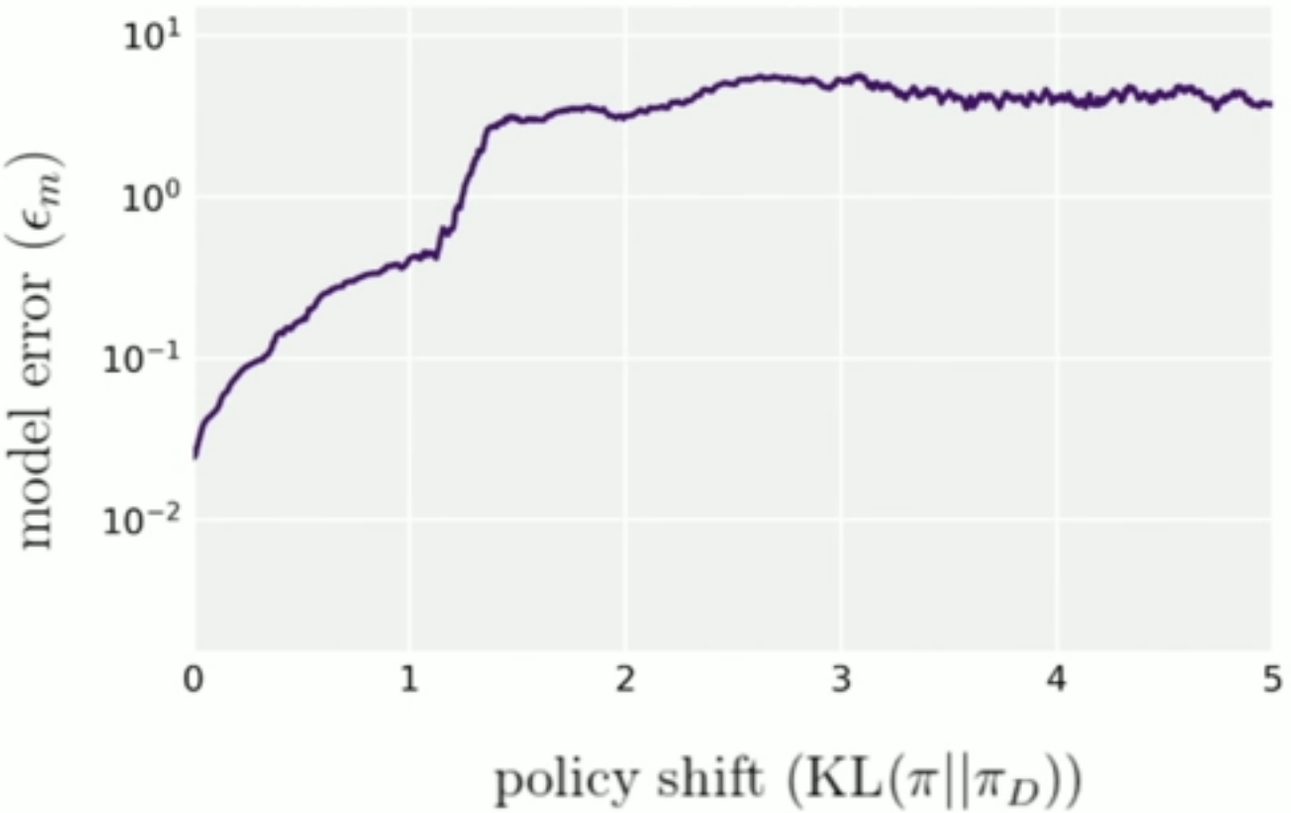
model error on data-collecting policy π_D

Model generalization in practice

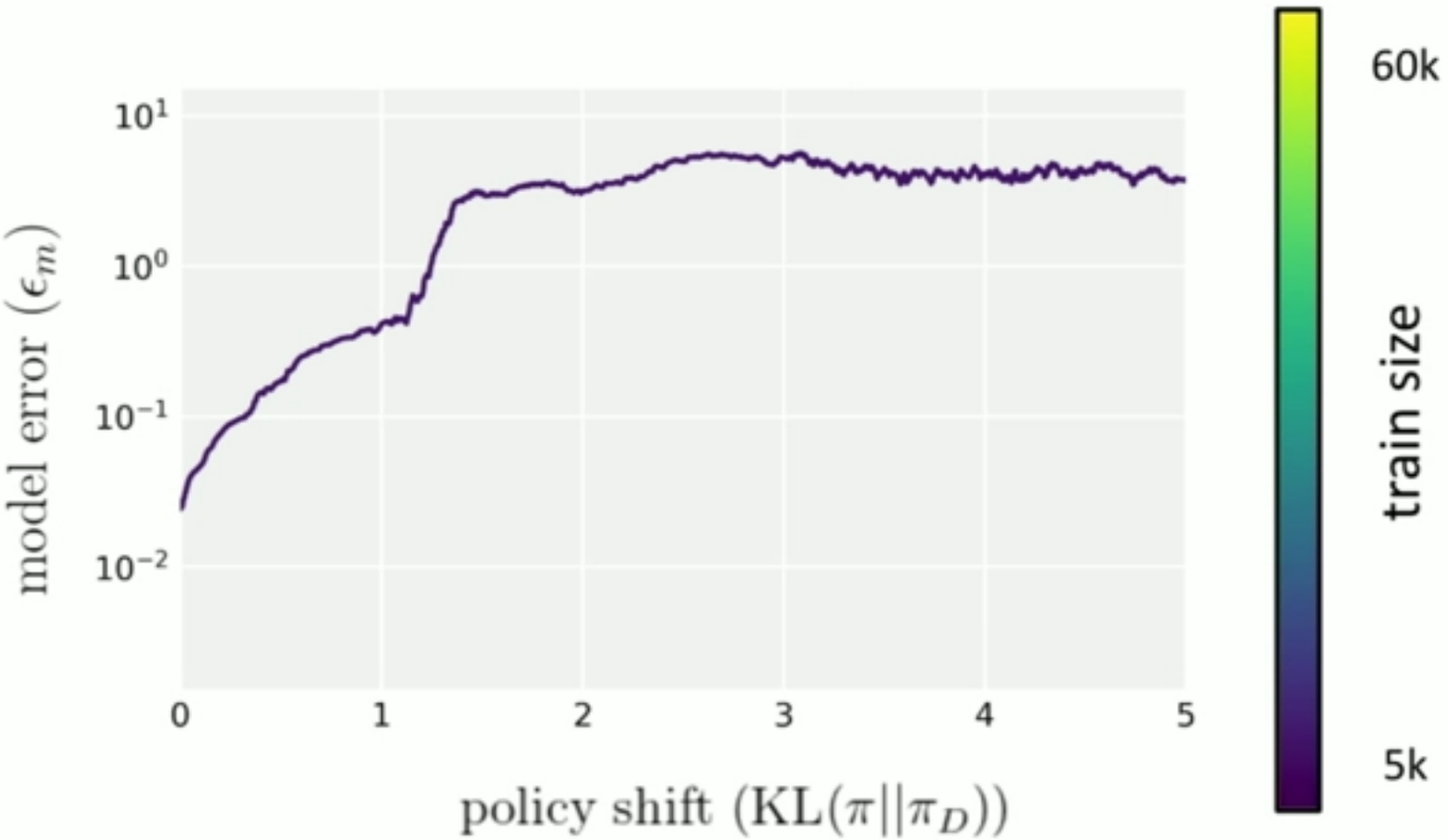


model error on *different* policy

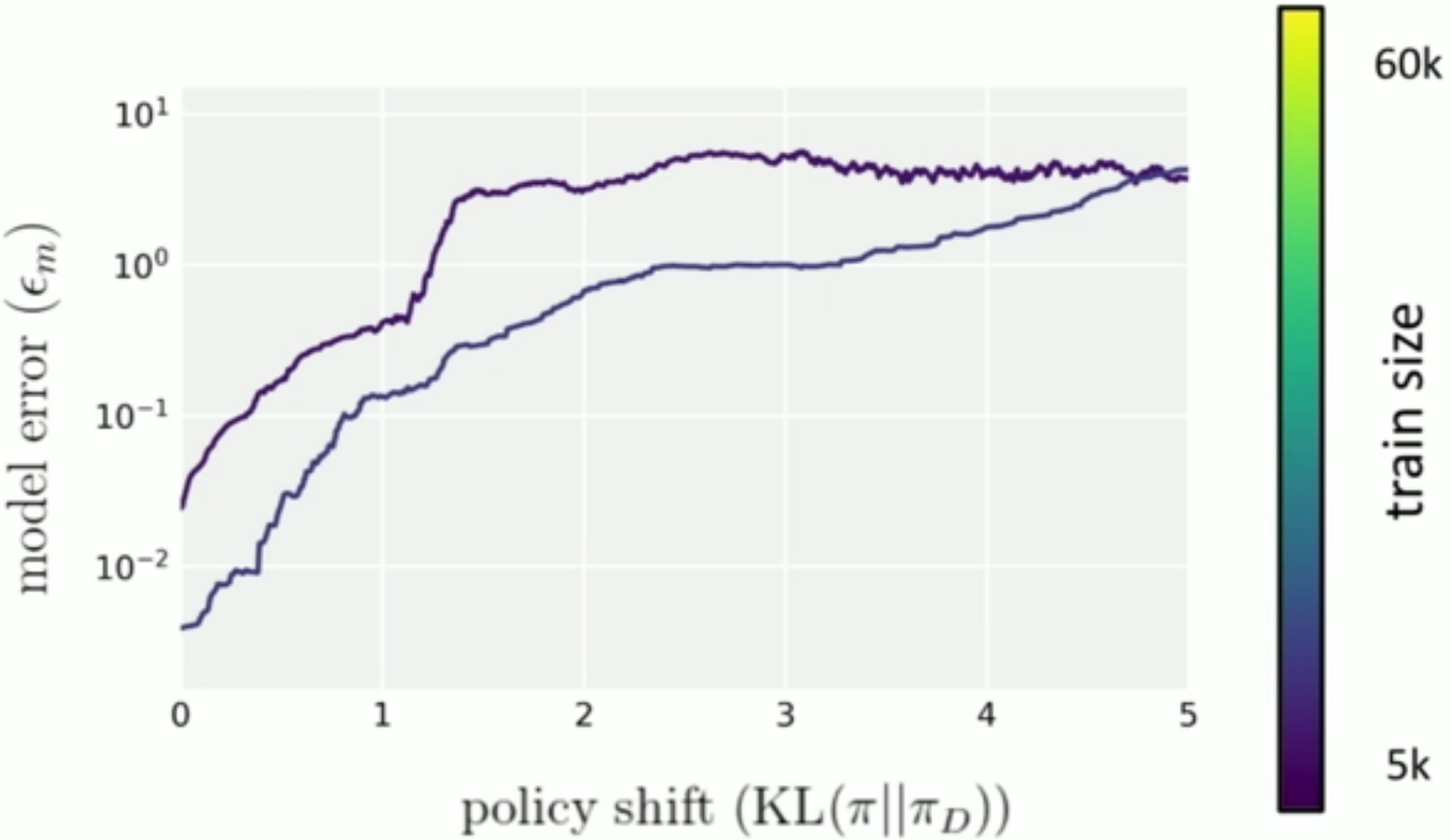
Model generalization in practice



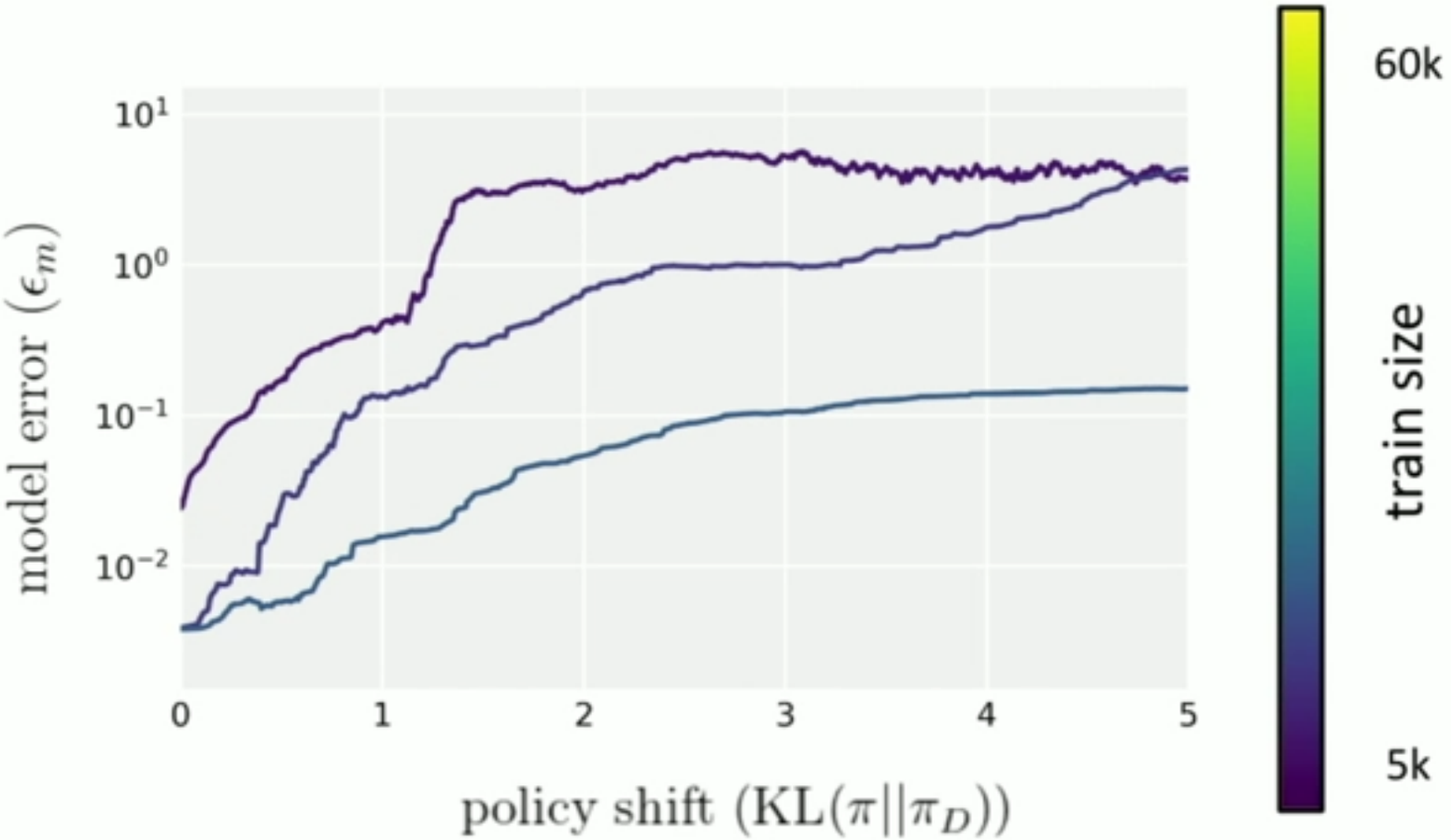
Model generalization in practice



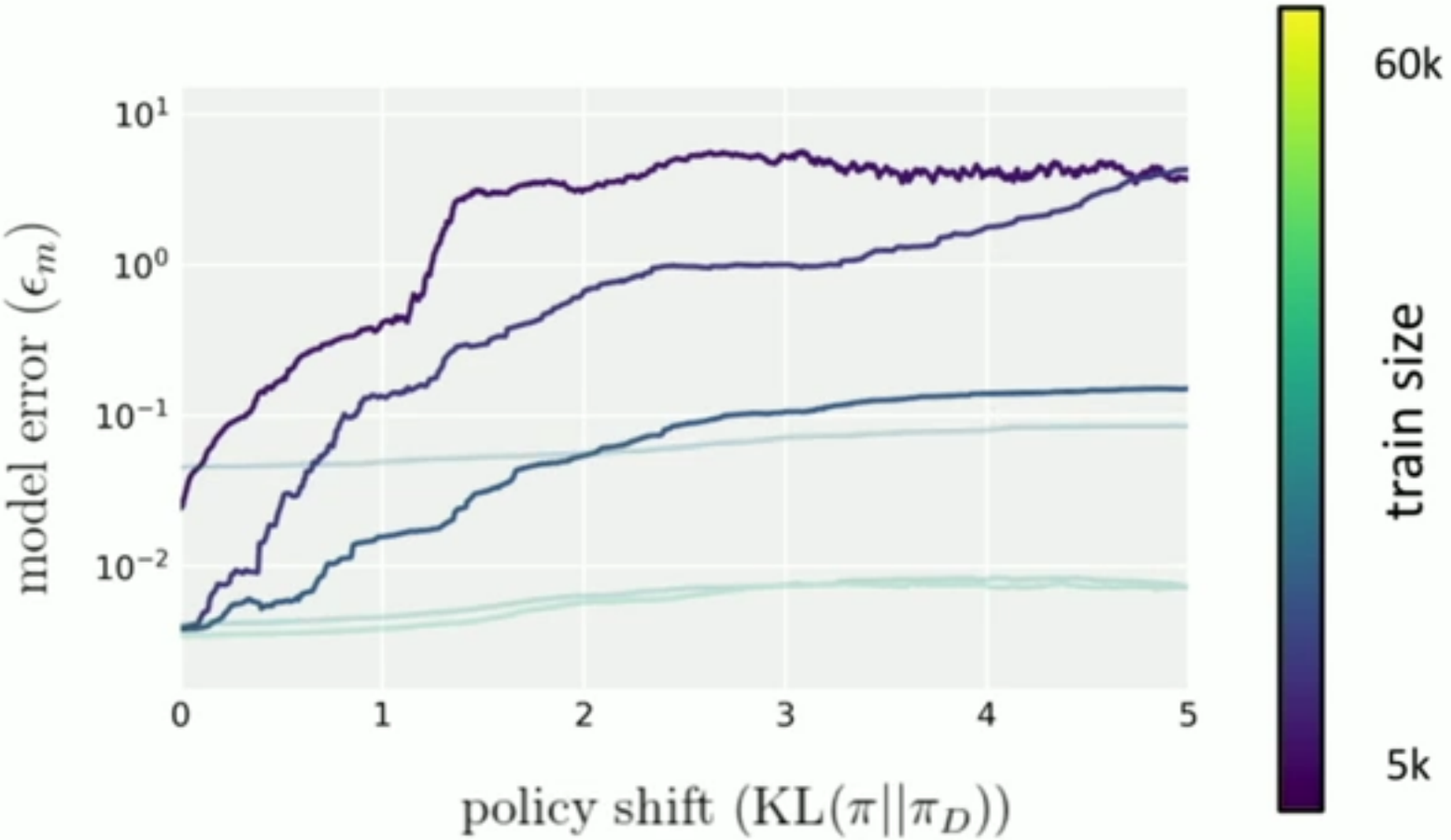
Model generalization in practice



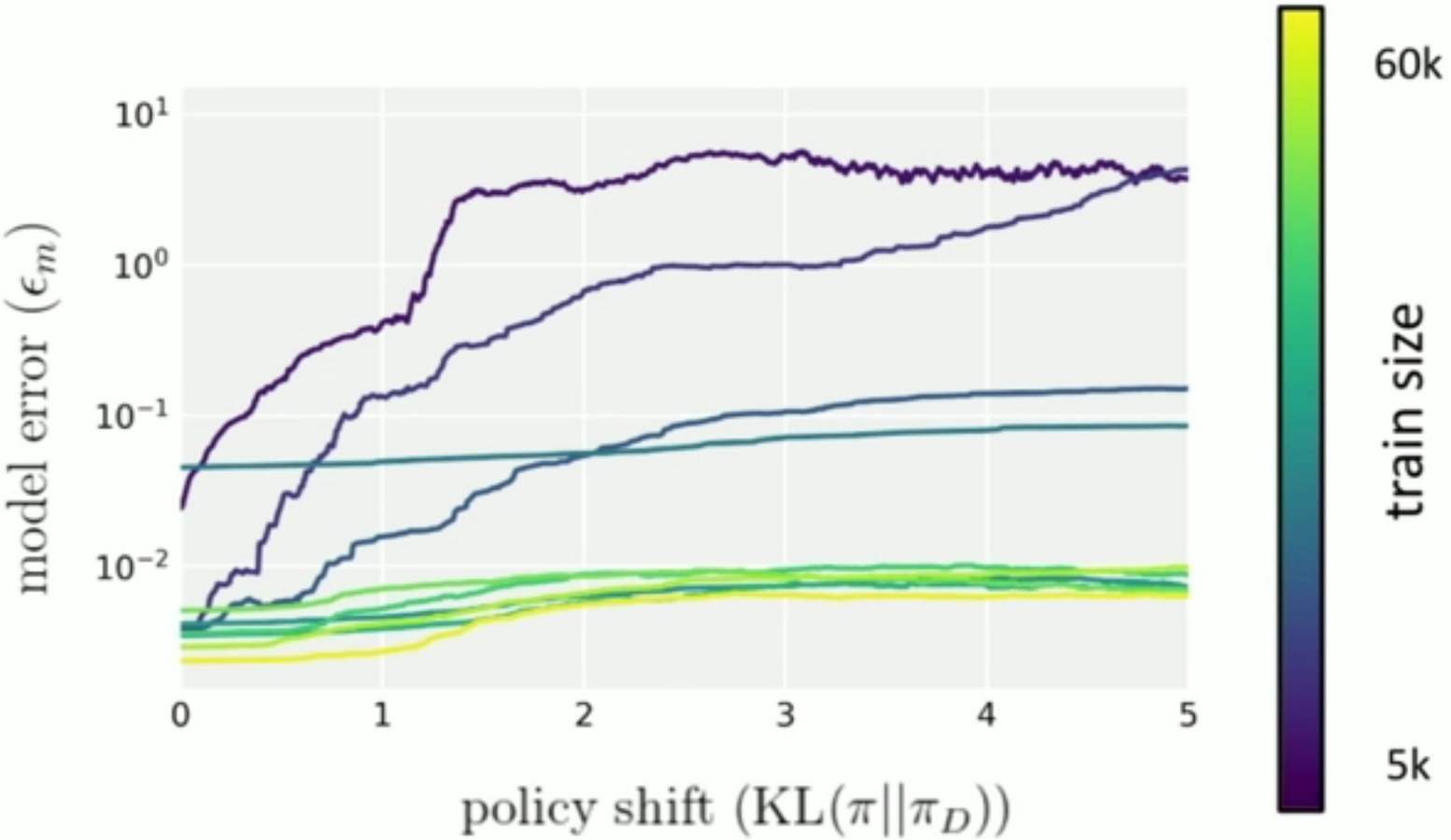
Model generalization in practice



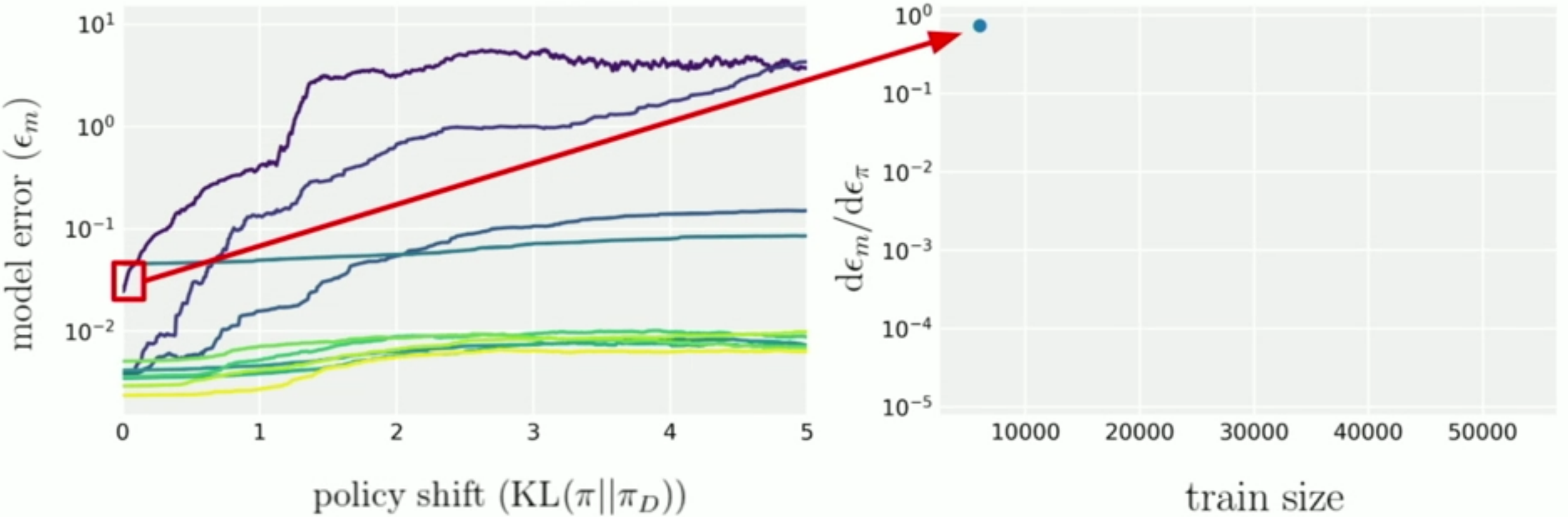
Model generalization in practice



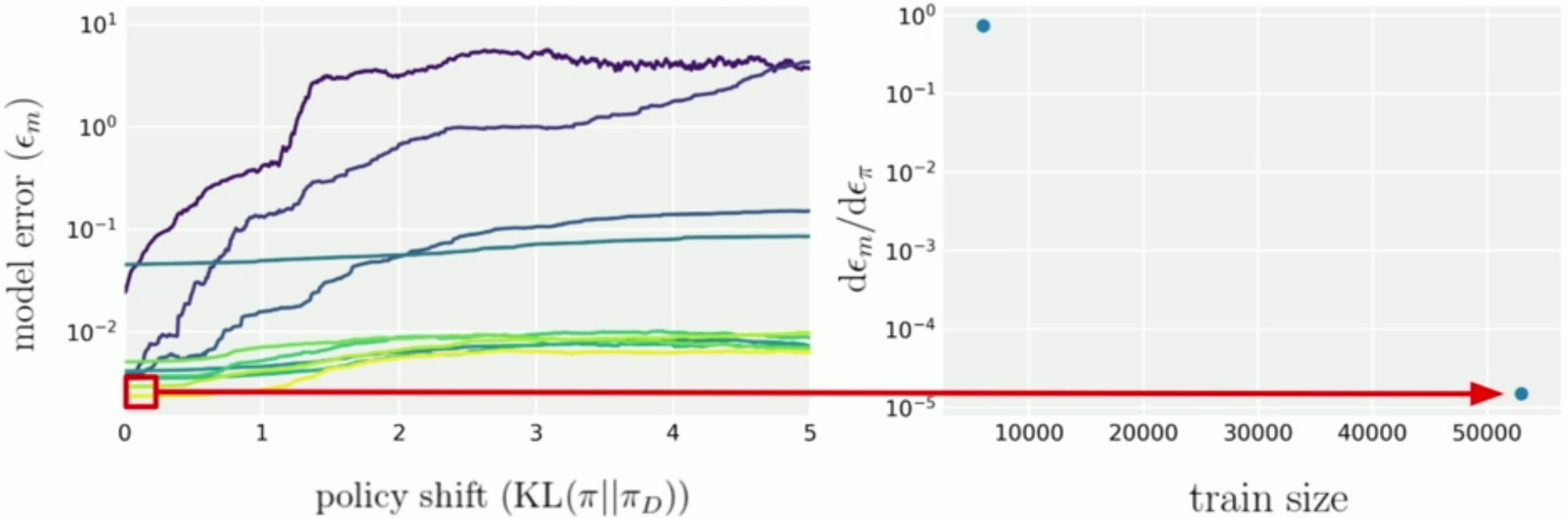
Model generalization in practice



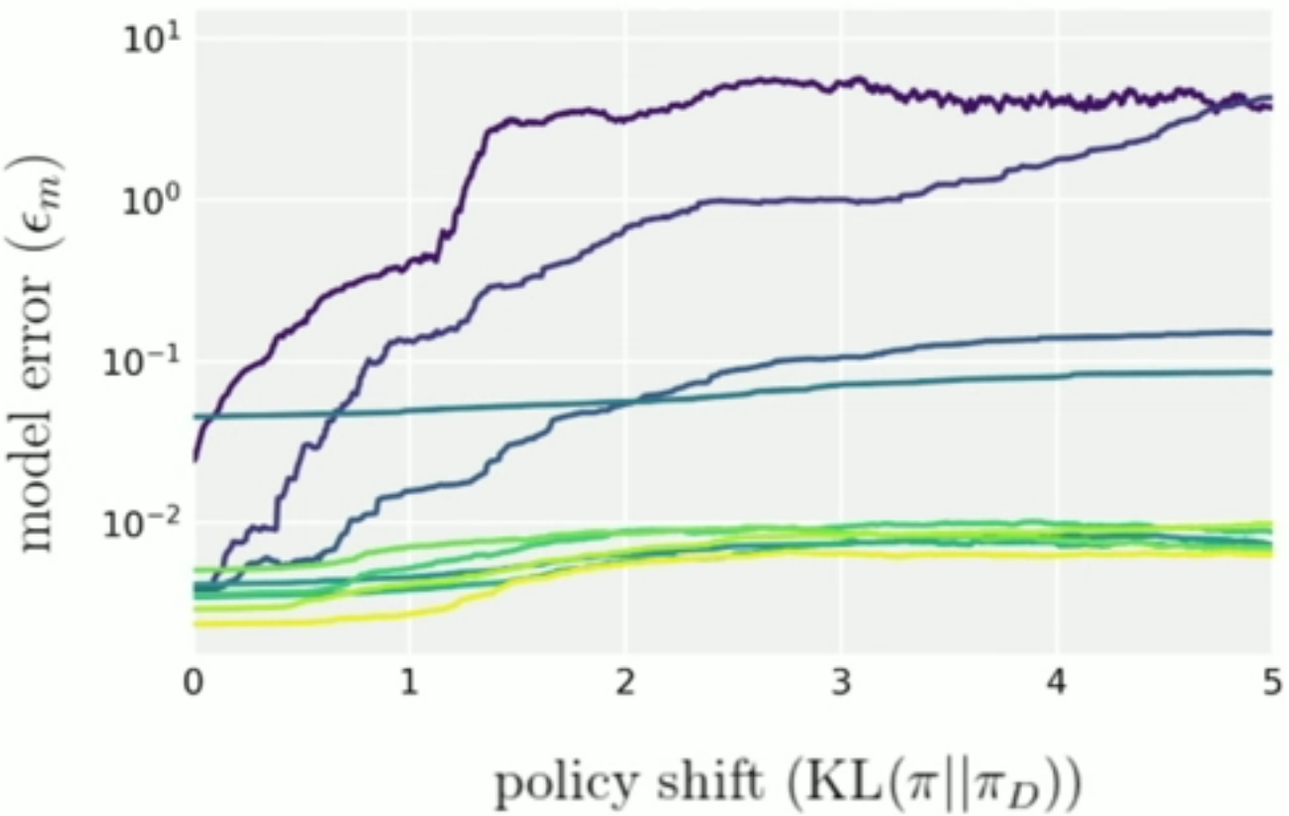
Model generalization in practice

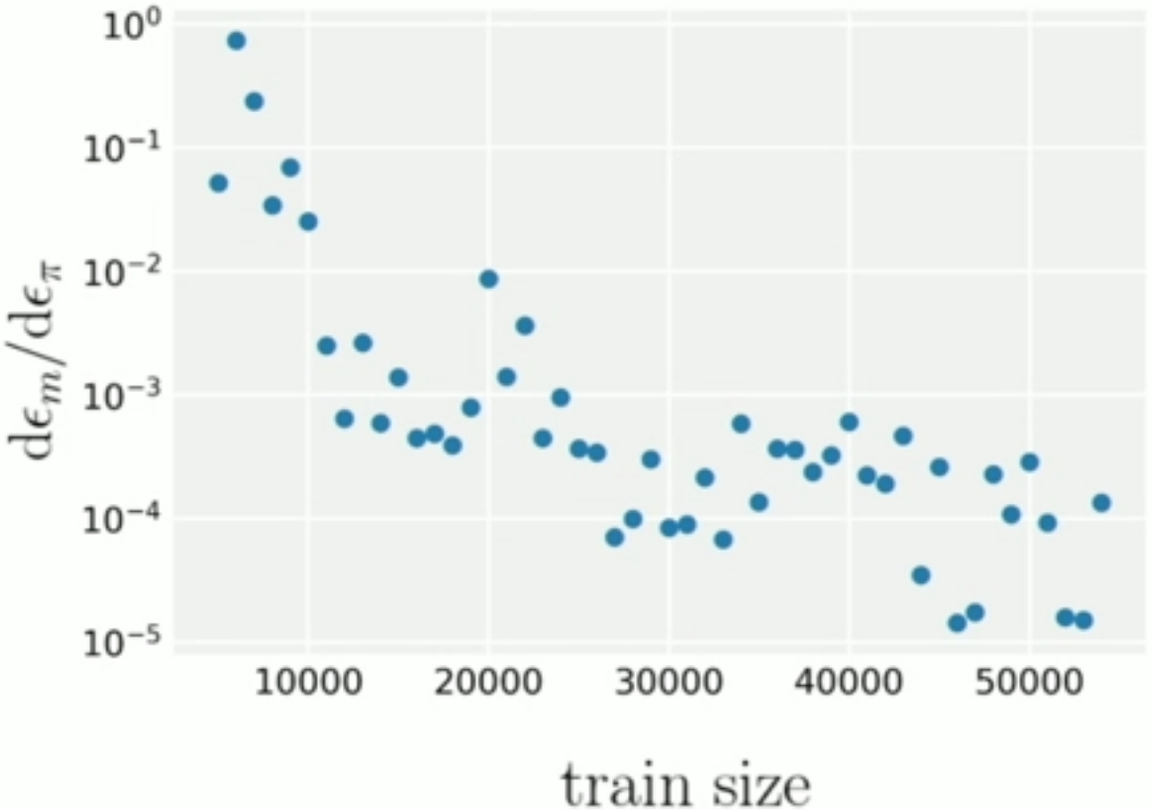


Model generalization in practice



Model generalization in practice





(Better) bounds on returns

- Use policy performance under model rollouts to derive lower bound in real environment

$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\left[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k+2)\epsilon_\pi}{(1-\gamma)} + \frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)\right]$$

(Better) bounds on returns

- Use policy performance under model rollouts to derive lower bound in real environment

$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\Big[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k+2)\epsilon_\pi}{(1-\gamma)} + \frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)\Big]$$

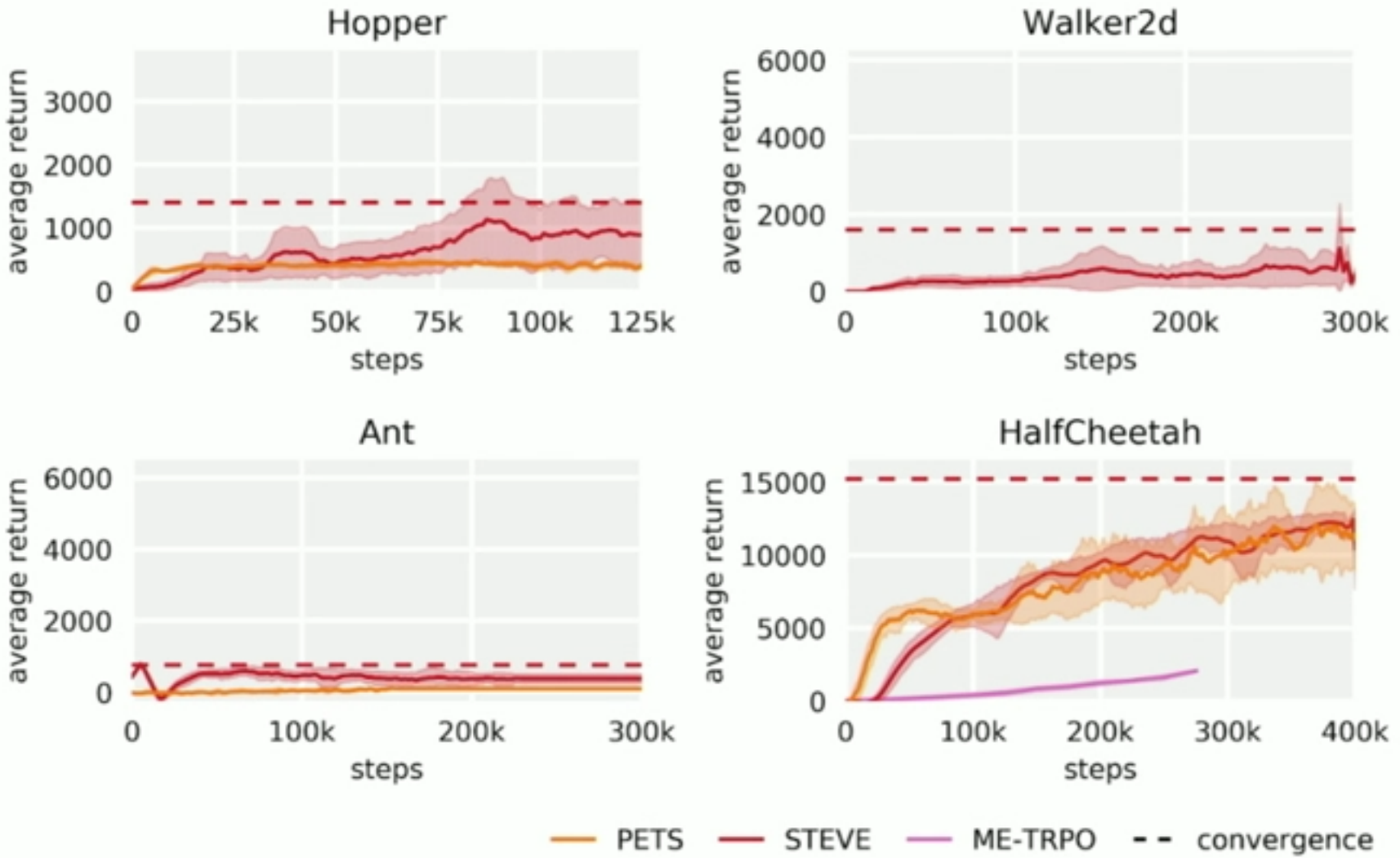
$$\epsilon_m + \frac{\mathrm{d}\epsilon_m}{\mathrm{d}\epsilon_\pi}$$

- Accounting for empirical model generalization motivates model usage

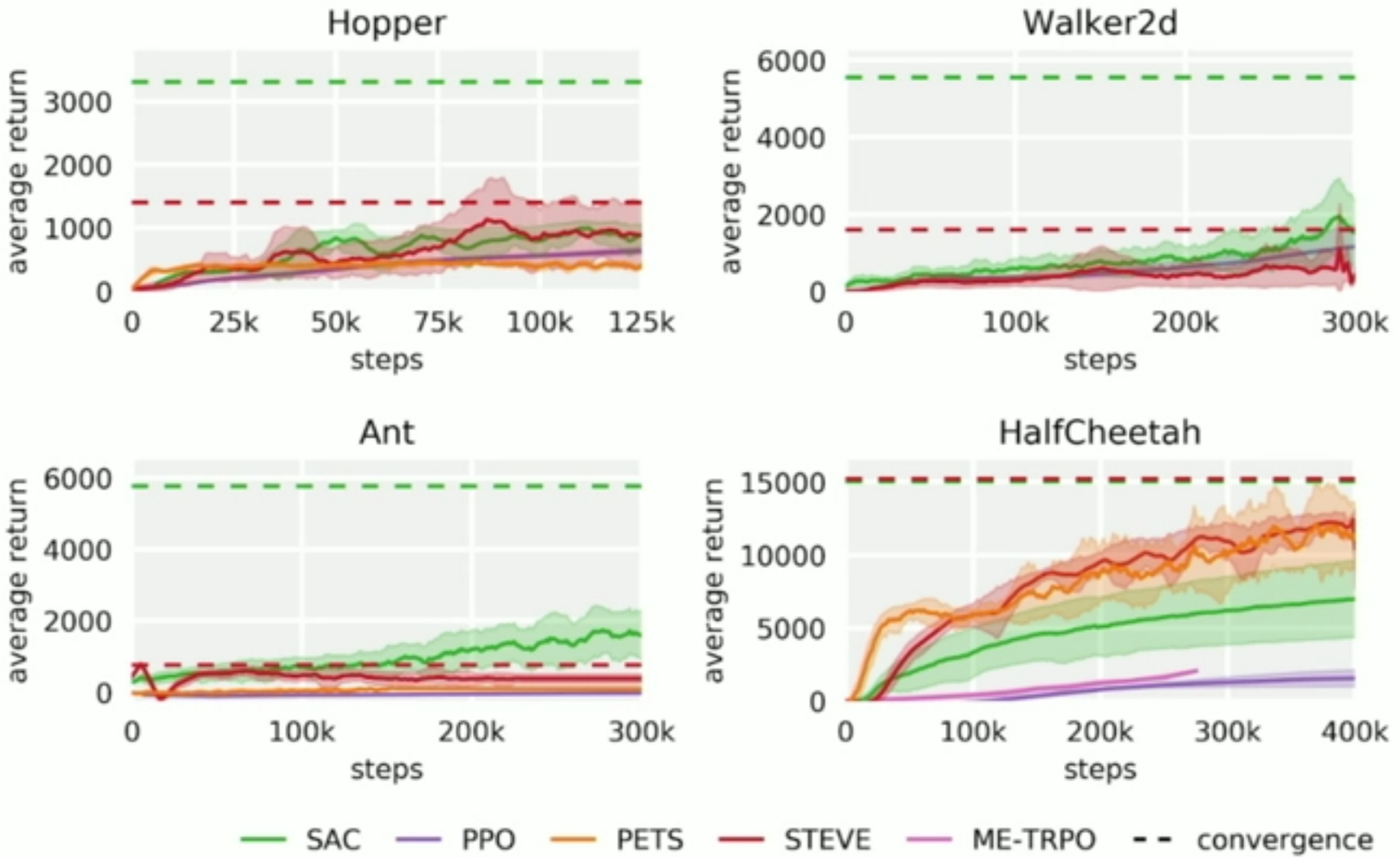
- If $\epsilon_\pi \approx \epsilon_m$, k will be small*

*When k=1, this corresponds to Dyna [Sutton, 1991]

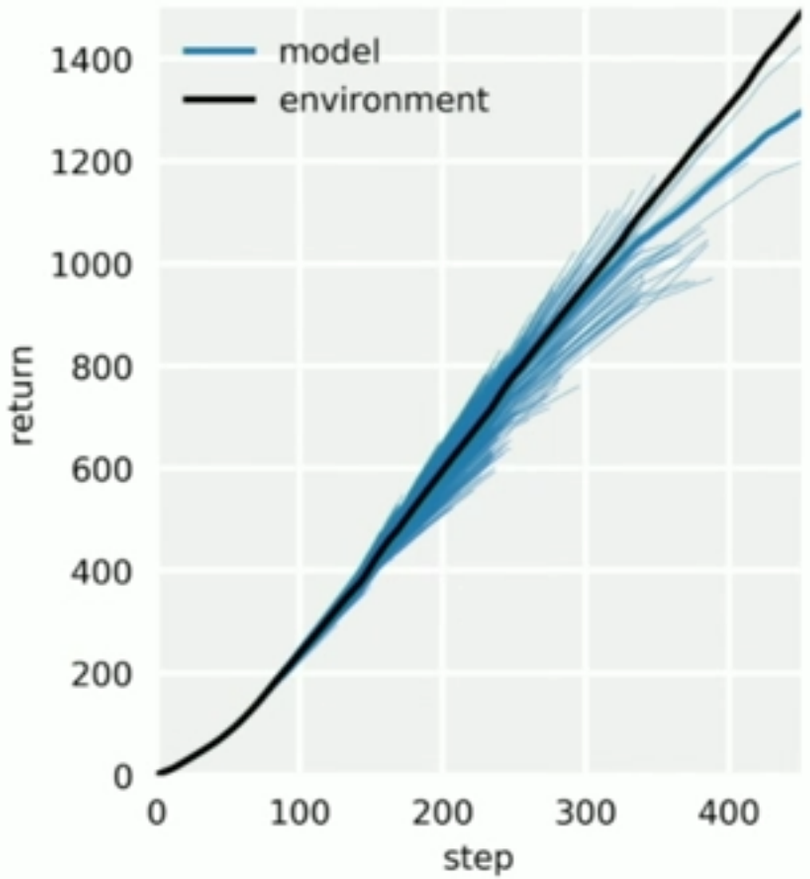
MBPO Results



MBPO Results



MBPO Results



added benefit: no model exploitation!

Summary

- Empirical model generalization motivates model usage
- Short model rollouts give large benefits to policy optimization
- MBPO avoids model exploitation and scales to long-horizon tasks

people.eecs.berkeley.edu/~janner/mbpo/