Trajectory-Based Off-Policy Deep Reinforcement Learning

Andreas Doerr[1,2,3] Michael Volpp[1] Marc Toussaint[3] Sebastian Trimpe[2] Christian Daniel[1]

[1] Bosch Center for Artificial Intelligence, Renningen, Germany

[2] Max Planck Institute for Intelligent Systems, Stuttgart/Tübingen, Germany

[3] Machine Learning & Robotics Lab, University of Stuttgart, Germany

BOSCH

Trajectory-Based Off-Policy Deep Reinforcement Learning
Fast & Efficient Model-Free Reinforcement Learning

How far can we push "model-free" RL?

$$\nabla_\theta J = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{H}\nabla_\theta \log \pi\left(a_t^{(i)}\middle|s_t^{(i)};\theta\right)R(\tau_i)\right]$$ [1]

[1] Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. 1992

Trajectory-Based Off-Policy Deep Reinforcement Learning
Problems with Policy Gradient Methods

How far can we push "model-free" RL?

$$\nabla_\theta J = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{H}\nabla_\theta \log\pi\left(a_t^{(i)}\middle|s_t^{(i)};\theta\right)R(\tau_i)\right]$$ [1]

Problems

Data inefficiency
- On-policy samples required
- No sample reuse

Gradient variance
- Stochastic policy
- Stochastic environment

Exploration vs. exploitation
- Step size control
- Policy (relative) entropy

[1] Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. 1992

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator

- Incorporation of all data (off-policy)
- Backtracking to good solutions

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator

- Incorporation of all data (off-policy)
- Backtracking to good solutions

Implementation:

- Importance sampling with empirical mixture distribution[1]

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_i(\theta) R(\tau_i) \qquad w_i(\theta) = \frac{p(\tau_i|\theta)}{\frac{1}{N}\sum_{j=0}^{N} p(\tau_i|\theta_j)}$$

[1] Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. NeurIPS 2010.

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator

- Incorporation of all data (off-policy)
- Backtracking to good solutions

Deterministic Policy

- Reduced rollout stochasticity
- Richer behaviors with parameter space exploration[2]

Implementation:

- Importance sampling with empirical mixture distribution[2]

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_i(\theta) R(\tau_i) \qquad w_i(\theta) = \frac{p(\tau_i|\theta)}{\frac{1}{N}\sum_{j=0}^{N} p(\tau_i|\theta_j)}$$

[1] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen,R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz,M. Parameter space noise for exploration. ICLR 2018.
[2] Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. NeurIPS 2010.

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator

- Incorporation of all data (off-policy)
- Backtracking to good solutions

Implementation:

- Importance sampling with empirical mixture distribution[2]

Deterministic Policy

- Reduced rollout stochasticity
- Richer behaviors with parameter space exploration[1]

Implementation:

- Model parameter Σ
- Length scale in action space

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_i(\theta) R(\tau_i) \qquad w_i(\theta) = \frac{N(a_t \mid \mu_\theta(s_t), \Sigma)}{\frac{1}{N}\sum_{j=0}^{N}\prod_{t=0}^{H} N(a_t \mid \mu_\theta(s_t), \Sigma)}$$

[1] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen,R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz,M. Parameter space noise for exploration. ICLR 2018.
[2] Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. NeurIPS 2010.

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator	Deterministic Policy	Distributional Policy Search
▶ Incorporation of all data (off-policy) ▶ Backtracking to good solutions	▶ Reduced rollout stochasticity ▶ Richer behaviors with parameter space exploration[1]	▶ Policy search leveraging lower bound

Implementation:

- Importance sampling with empirical mixture distribution[2]

Implementation:

- Model parameter Σ
- Length scale in action space

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} w_i(\theta) R(\tau_i) \qquad w_i(\theta) = \frac{N(a_t \mid \mu_\theta(s_t), \Sigma)}{\frac{1}{N}\sum_{j=0}^{N}\prod_{t=0}^{H} N(a_t \mid \mu_\theta(s_t), \Sigma)}$$

[1] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen,R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz,M. Parameter space noise for exploration. ICLR 2018.
[2] Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. NeurIPS 2010.

Trajectory-Based Off-Policy Deep Reinforcement Learning
Deep Deterministic Off-Policy Gradients (DD-OPG)

Core concepts in DD-OPG

Global Return Distribution Estimator

- Incorporation of all data (off-policy)
- Backtracking to good solutions

Implementation:

- Importance sampling with empirical mixture distribution[2]

Deterministic Policy

- Reduced rollout stochasticity
- Richer behaviors with parameter space exploration[1]

Implementation:

- Model parameter Σ
- Length scale in action space

Distributional Policy Search

- Policy search leveraging lower bound

Implementation:

- Estimation of empirical sample size and variance

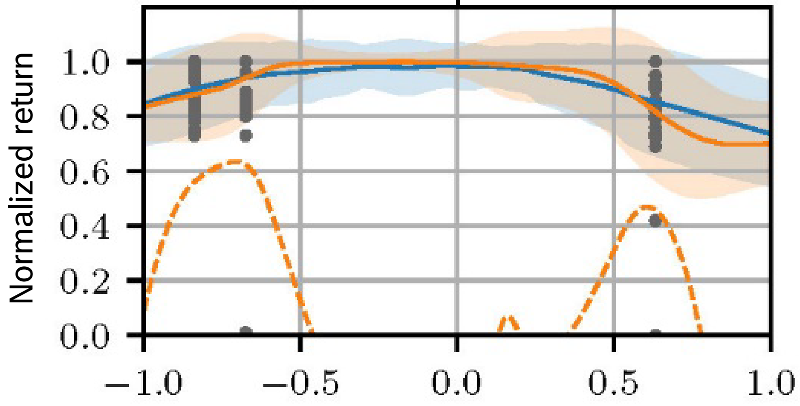
$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} w_i(\theta) R(\tau_i) \qquad w_i(\theta) = \frac{N(a_t \mid \mu_\theta(s_t), \Sigma)}{\frac{1}{N} \sum_{j=0}^{N} \prod_{t=0}^{H} N(a_t \mid \mu_\theta(s_t), \Sigma)}$$

[1] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen,R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz,M. Parameter space noise for exploration. ICLR 2018.
[2] Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. NeurIPS 2010.

BOSCH

Trajectory-Based Off-Policy Deep Reinforcement Learning
Return Distribution Estimator

Importance sampling estimate



Parameterspace

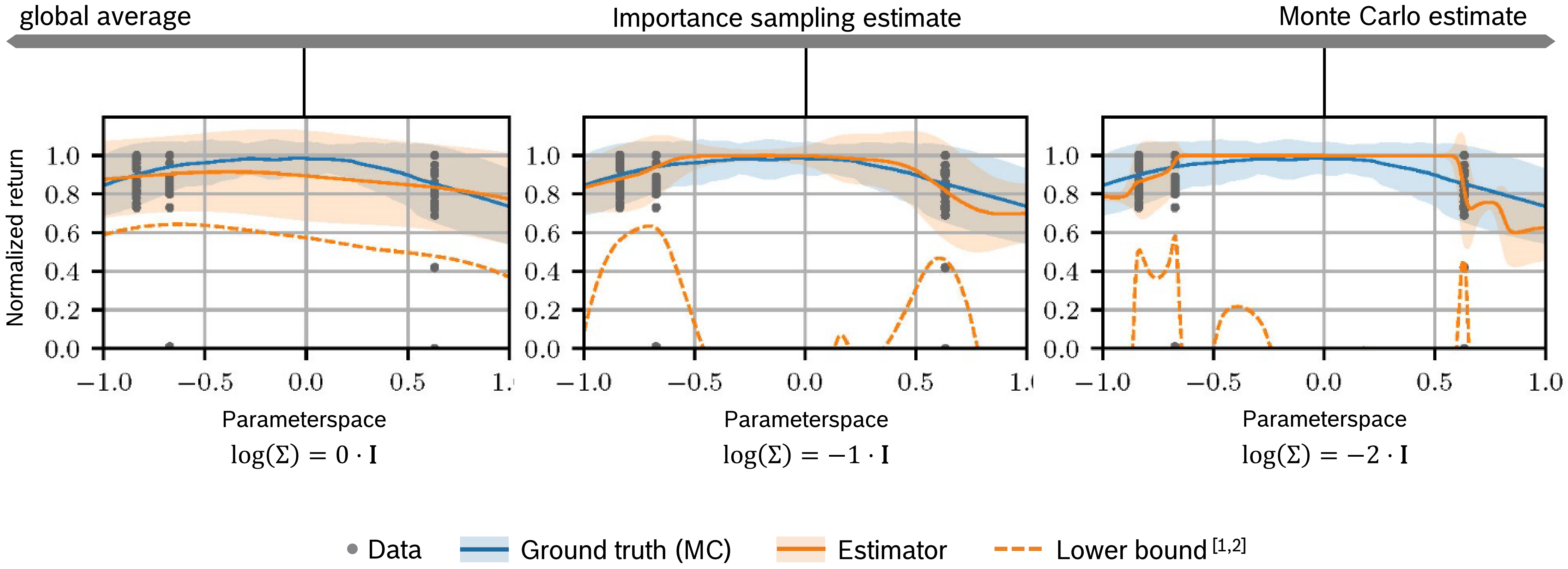
$$\log(\Sigma) = -1 \cdot \mathbf{I}$$

● Data — Ground truth (MC) — Estimator — Lower bound [1,2]

[1] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz,P. Trust region policy optimization. ICML 2015.
[2] Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. NeurIPS 2018.

Trajectory-Based Off-Policy Deep Reinforcement Learning

Return Distribution Estimator



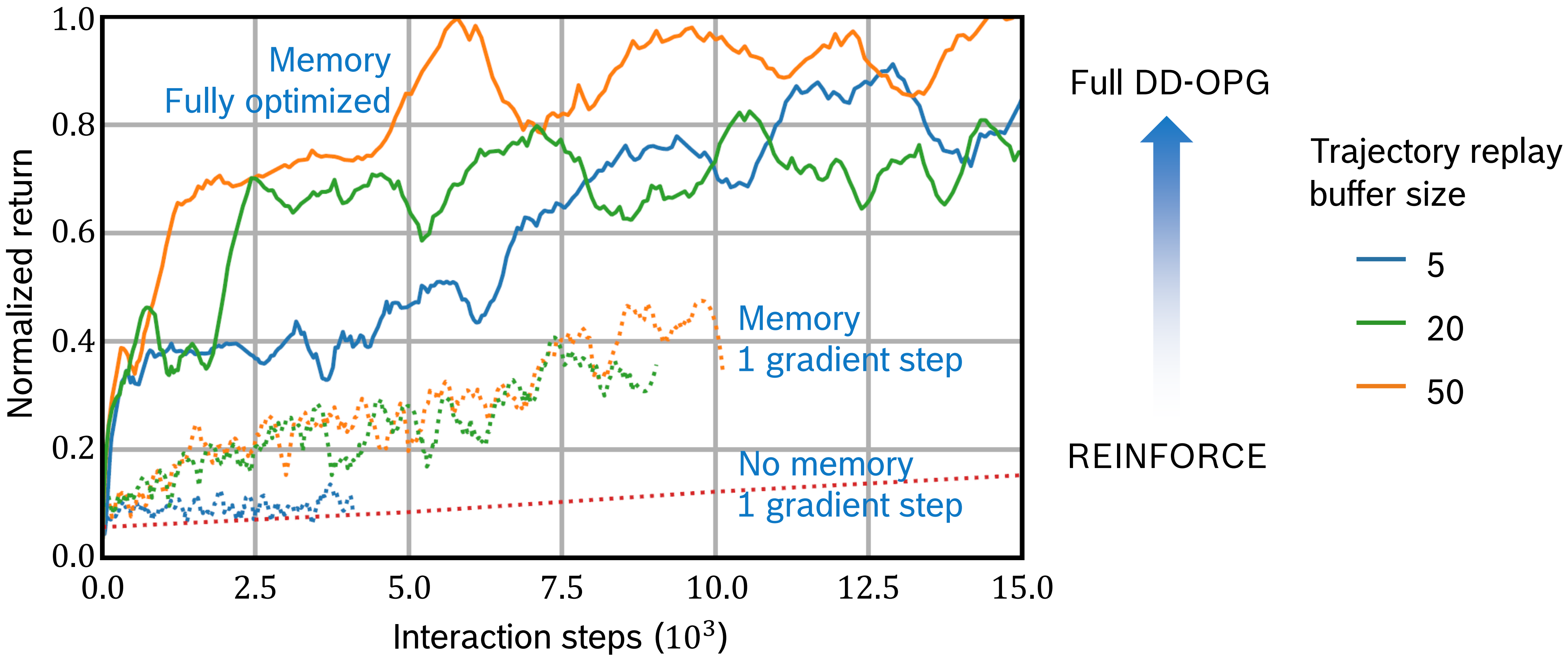
[1] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz,P. Trust region policy optimization. ICML 2015.
[2] Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. NeurIPS 2018.

Trajectory-Based Off-Policy Deep Reinforcement Learning
Algorithmic Choices

	DD-OPG	REINFORCE	TRPO	PPO
Memory selection	All available trajectories Prioritized trajectory replay	Only on-policy samples from current batch		
Exploration	Parameter space	Action space		
Objective $\mathcal{L}(\theta)$	Expected return lower bound	Expected return	Expected return with KL constraint	Expected return (lower bound)
Optimization	Fully optimized with backtracking	One gradient step	Locally optimized	

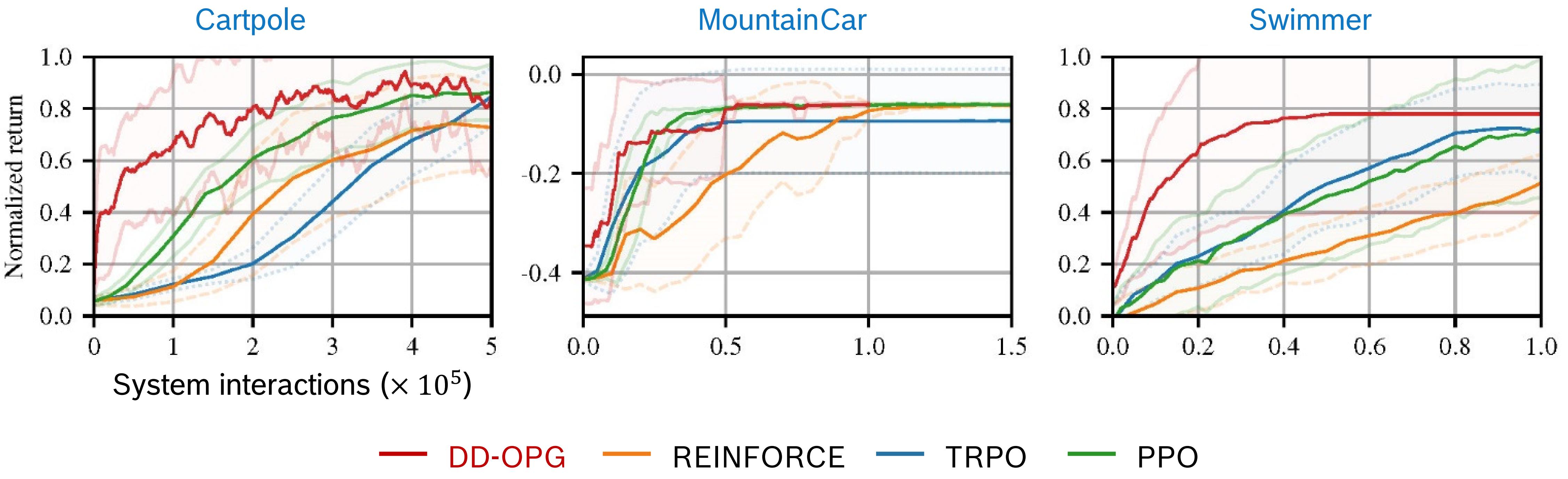
Trajectory-Based Off-Policy Deep Reinforcement Learning

Experimental Results – From REINFORCE to DD-OPG



Trajectory-Based Off-Policy Deep Reinforcement Learning

Experimental Results – Benchmark Results



- GARAGE: continuous control environments
- Gaussian MLP policy (16, 16)

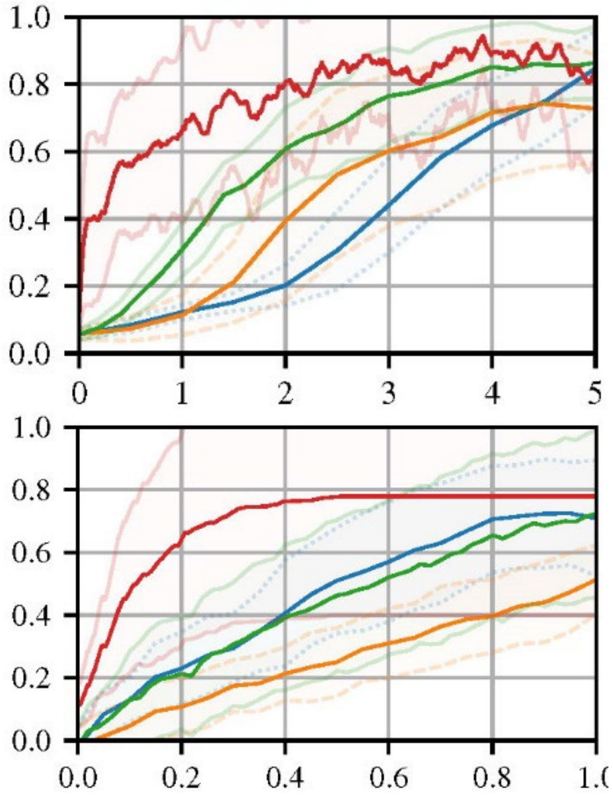
Trajectory-Based Off-Policy Deep Reinforcement Learning
Conclusion

- Novel off-policy policy gradient methods
- Enables data-efficient sample reuse
- Incorporation of low-noise deterministic rollouts
- Lengthscale in action space as only model assumption
- Promising benchmark results

Code available
https://github.com/
boschresearch/DD_OPG

Poster #44
Pacific
Ballroom

Bosch
booth



DD-OPG (red) benchmark results

BOSCH