TransferLab

Episode-Based Reinforcement Learning using Movement Primitives

appliedAI Seminar - Reinforcement Learning

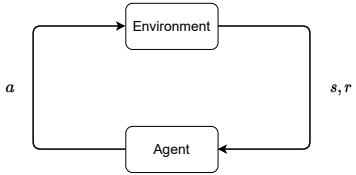
Maximilian Hüttenrauch

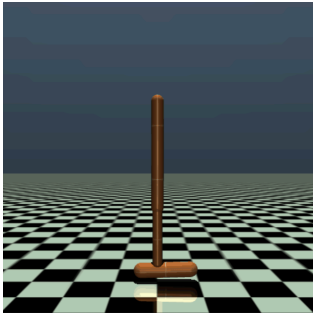
TransferLab@appliedAI

Introduction

Deep RL

- Widely used in sequential decision making problems
- Works well in broad set of tasks (e.g., forward motion)
- Step-based paradigm $\pi(a \mid s)$





Source: *https://gymnasium.farama.org/environments/mujoco/hopper/*

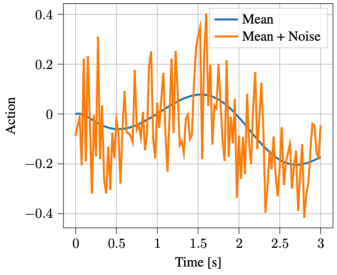
MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transition dynamics $P = p(s' \mid s, a)$
- Reward function $R = r(s, a)$
- Discount factor γ

Introduction

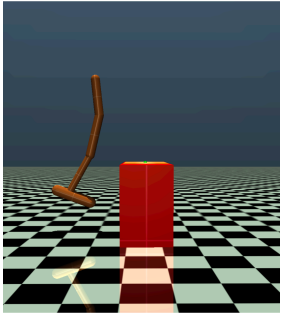
Why episode-based RL?

- Action noise bad for robotics
- Learned policies are not energy efficient [Otto et al., 2022]



Source: [Li et al., 2024]

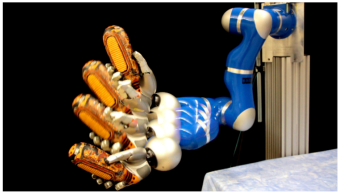
- Not all tasks can be modelled with dense/Markovian reward

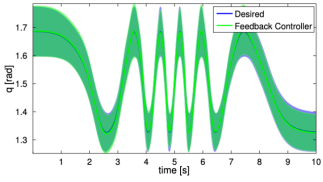


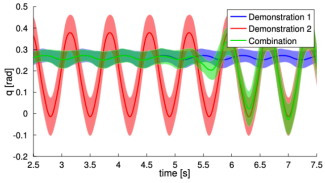
Source: [Hüttenrauch and Neumann, 2024]

Goal: Provide a different approach to learn episodic tasks

Movement Primitives







Source: [Paraschos et al., 2013]

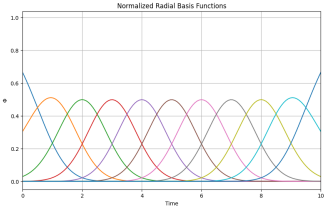
- Parameterized trajectory $\boldsymbol{\tau} = \boldsymbol{y}_{1:T} = \begin{bmatrix} q_{1:T} \\ \dot{q}_{1:T} \end{bmatrix} = f(\boldsymbol{w})$
- Modular movement representation and generation
- Linear basis function model as function approximator
- Execution: Trajectory tracking controller

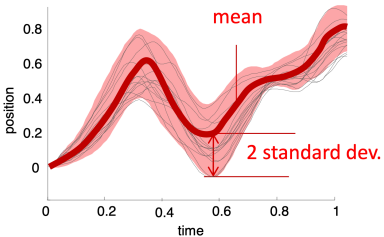
Probabilistic Movement Primitives [Paraschos et al., 2013]

- Distribution over trajectories $p(\boldsymbol{\tau} \mid \boldsymbol{w}) = \prod_t \mathcal{N}(\boldsymbol{y}_t \mid \boldsymbol{\Phi}_t \boldsymbol{w}, \boldsymbol{\Sigma}_y)$
- Trajectory defined directly in terms of position
- Start position cannot be enforced

ProMPs (mean trajectory)

$$\boldsymbol{y}_t = \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} = \boldsymbol{\Phi}_t \boldsymbol{w} + \epsilon_y$$





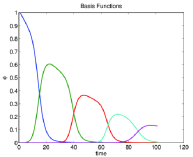
Source (right): [Neumann, 2018]

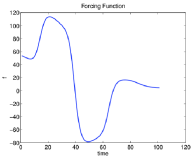
Dynamical Movement Primitives [Ijspeert et al., 2013]

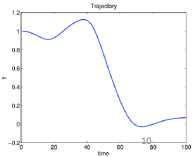
- Spring damper second order dynamical system resulting in smooth trajectory
- Trajectory defined in terms of acceleration profile
- Starting position built in, final position can be adapted

DMPs

$$\tau\ddot{y} = \alpha_z(\beta_z(g - y) - \dot{y}) + f(t)$$
$$f(t) = \mathbf{\Phi}_t \boldsymbol{w}$$







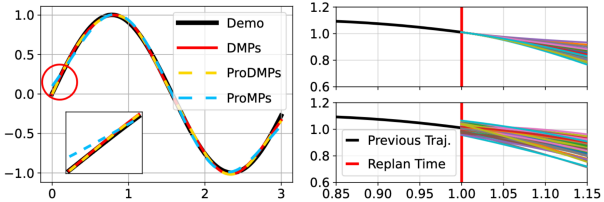
Source: [Neumann, 2018]

Probabilistic Dynamic Movement Primitives [Li et al., 2023]

- Combine advantages of ProMPs and DMPs
- Offline numerical integration
- Start position can be enforced, enables smooth replanning

ProDMPs

$$y(t) = c_1 y_1(t) + c_2 y_2(t) + \mathbf{\Phi}(t)\boldsymbol{w}$$
$$\dot{y}(t) = c_1 \dot{y}_1(t) + c_2 \dot{y}_2(t) + \dot{\mathbf{\Phi}}(t)\boldsymbol{w}$$



Source: [Li et al., 2023]

Learning from Demonstrations

A ProMP can be learned using linear regression

- Record trajectories of successful examples (kinesthetic teaching, tele-op)
- For each trajectory $\boldsymbol{\tau}_i$, obtain $\boldsymbol{w}_i$

$$\boldsymbol{w}_i = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma \boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\tau}_i$$

- Compute mean and variance

$$\boldsymbol{\mu}_w = \frac{1}{N}\sum_i w_i$$

$$\boldsymbol{\Sigma}_w = \frac{\sum_i (\boldsymbol{w}_i - \boldsymbol{\mu}_i)(\boldsymbol{w}_i - \boldsymbol{\mu}_i)^T}{N-1}$$

Reinforcement Learning by Direct Policy Search

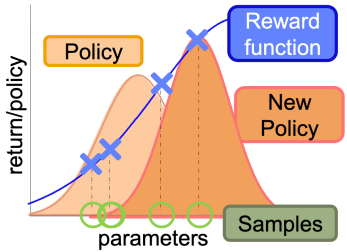
What if we don't have demonstrations?

- Find $\pi(\boldsymbol{w})$ with maximum return

$$\pi^* = \arg\max_{\pi} \int \pi(\boldsymbol{w}) \mathcal{R}(\boldsymbol{w}) d\boldsymbol{w}$$

$$\mathcal{R}(\boldsymbol{w}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t \mid \boldsymbol{w}\right]$$

- High level policy $\pi(\boldsymbol{w})$
  - Mean: Current best estimate of the maximum
  - Variance: Direction to explore
- Low level policy: Use open loop control



Source: [Neumann, 2017]

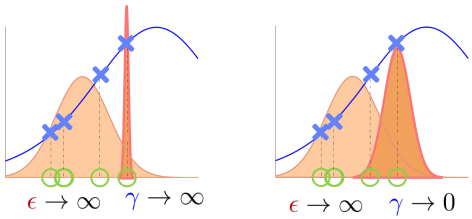
Information-Theoretic Policy Search

Towards an algorithm:

- Unconstrained optimization leads to pre-mature convergence
- Introduce constraints to tackle exploration-exploitation trade-off
  - Kullback Leibler Divergence: $KL(p\|q) = \int q(w) \log \frac{q(w)}{p(w)} dw$
  - Bound entropy loss: $H(p) = -\int p(w) \log p(w) dw$

Information-Theoretic Policy Update

$$\begin{aligned} \arg\max_{\pi} \quad & \int \pi(w)\mathcal{R}(w)dw \\ \text{s.t.} \quad & KL(\pi\|\pi_{old}) \leq \epsilon, \\ & H(\pi_{old}) - H(\pi) \leq \gamma \end{aligned}$$



Source: [Neumann, 2017]

Model-Based Relative Entropy Stochastic Search [Abdolmaleki et al., 2015]

Idea:

- Replace return function with learned model
- Compatible function approximation
  $\mathcal{R}(w) \approx \hat{\mathcal{R}}(w) = -\frac{1}{2} w^T A w + w^T a + a$

Iterative algorithm:

- Estimate A, a, a from sampled returns
- Solve using method of Lagrange multipliers
  - Minimize Lagrange dual (closed form)
  - Update rules in terms of optimal L. multiplier
  - Natural gradient direction

MORE [Abdolmaleki et al., 2015]

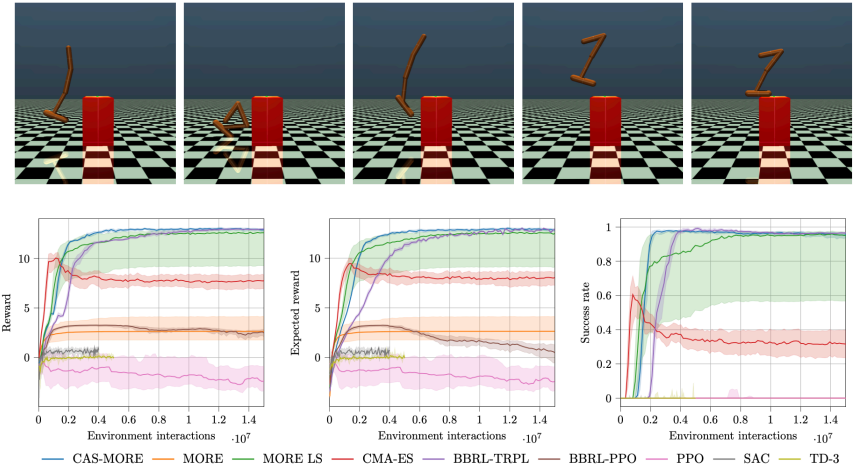
$$\begin{aligned} \arg\max_{\pi} \quad & \int \pi(w)\hat{\mathcal{R}}(w)dw \\ \text{s.t.} \quad & KL(\pi\|\pi_{old}) \leq \epsilon, \\ & H(\pi) \geq \beta \end{aligned}$$

CAS-MORE
[Hüttenrauch and Neumann, 2024]

$$\begin{aligned} \arg\max_{\boldsymbol{\mu}} \quad & \int \pi(w)\hat{\mathcal{R}}(w)dw\Big|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{old}} \\ \text{s.t.} \quad & KL(\pi\|\pi_{old})|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{old}} \leq \epsilon_{\mu} \end{aligned}$$

$$\begin{aligned} \arg\max_{\boldsymbol{\Sigma}} \quad & \int \pi(w)\hat{\mathcal{R}}(w)dw\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_{old}} \\ \text{s.t.} \quad & KL(\pi\|\pi_{old})|_{\boldsymbol{\mu}=\boldsymbol{\mu}_{old}} \leq \epsilon_{\Sigma} \end{aligned}$$

Model-Based Relative Entropy Stochastic Search



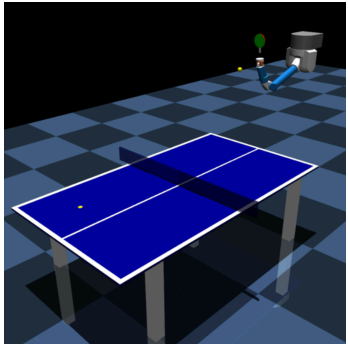
Source: [Hüttenrauch and Neumann, 2024]

Contextual Episode-Based Policy Search

- Conditions described by context $\boldsymbol{c}$
- Policy: $\pi(\boldsymbol{w} \mid \boldsymbol{c}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}(\boldsymbol{c}), \boldsymbol{\Sigma}(\boldsymbol{c}))$
- Find $\boldsymbol{w}$ that maximizes return given a context $\boldsymbol{c}$

$$\pi^* = \arg\max_{\pi} \int p(\boldsymbol{c}) \int \pi(\boldsymbol{w} \mid \boldsymbol{c}) \mathcal{R}(\boldsymbol{w}, \boldsymbol{c}) d\boldsymbol{w} d\boldsymbol{c}$$

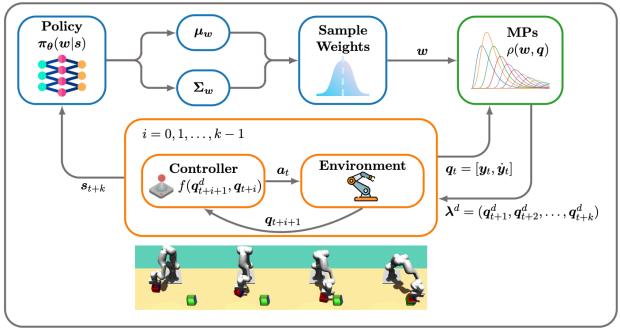
$$\mathcal{R}(\boldsymbol{c}, \boldsymbol{w}) = \mathbb{E}\left[\sum_{t=1}^{T} r_t \mid \boldsymbol{c}, \boldsymbol{w}\right]$$



Source: [Otto et al., 2023]

Deep Black-Box Reinforcement Learning [Otto et al., 2022]

- Neural network mapping from context to distribution parameters
- Differentiable trust region layers [Otto et al., 2020] for implicit constraint satisfaction



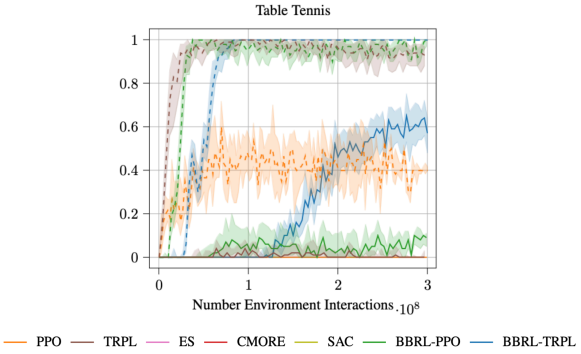
Source: [Otto et al., 2023]

BBRL TRPL

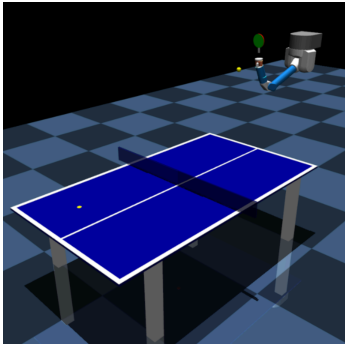
$$\arg\max_{\pi} \mathbb{E}\left[\frac{\pi(\boldsymbol{w} \mid \boldsymbol{c})}{\pi_{old}(\boldsymbol{w} \mid \boldsymbol{c})} A^{\pi_{old}}(\boldsymbol{c}, \boldsymbol{w})\right]$$

$$A^{\pi}(\boldsymbol{c}, \boldsymbol{w}) = \mathcal{R}(\boldsymbol{c}, \boldsymbol{w}) - V^{\pi}_{\phi}(\boldsymbol{c})$$

Deep Black-Box Reinforcement Learning



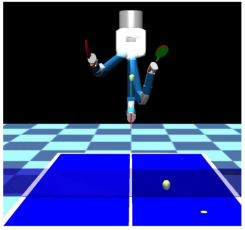
Source: [Otto et al., 2022]



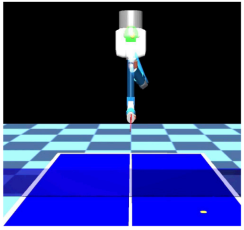
Source: [Otto et al., 2023]

Diverse Skill Learning

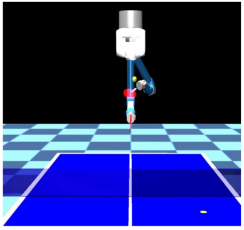
- Learn multiple solutions for a given $\boldsymbol{c}$
- Mixture of Experts policy $\pi(\boldsymbol{w} \mid \boldsymbol{c}) = \sum_o p(o \mid \boldsymbol{c}) p(\boldsymbol{w} \mid \boldsymbol{c}, o)$
  - Gating policy $p(o \mid \boldsymbol{c})$ assigns an expert o to a given context $\boldsymbol{c}$
  - Expert $p(\boldsymbol{w} \mid \boldsymbol{c}, o)$ movement primitive (Gaussian)



(a) Strike from left to right.



(b) Forehand counter strike.



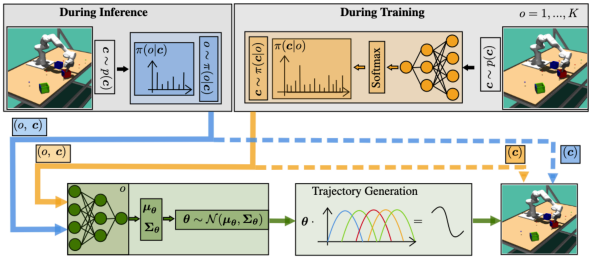
(c) Backhand backspin strike.

Source: [Celik et al., 2023]

Di-SkilL [Celik et al., 2023]

Curriculum learning

- Generate preferable contexts during learning
- Introduce learned context distribution $\rho(\boldsymbol{c}) = \sum_o p(\boldsymbol{c} \mid o) p(o)$



Source: [Celik et al., 2023]

Di-SkilL

$$\max_{\pi(\boldsymbol{w} \mid \boldsymbol{c}), \rho(\boldsymbol{c})} \quad \mathbb{E}_{\rho(\boldsymbol{c})}[\mathbb{E}_{\pi(\boldsymbol{w}|\boldsymbol{c})}[\mathcal{R}(\boldsymbol{c}, \boldsymbol{w})] + \alpha H(\pi)] - \beta KL(\rho \| p)$$

Summary ERL

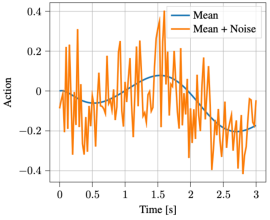
Depending on the setting different methods exist

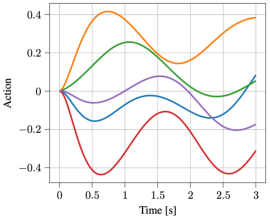
- Non-contextual: Black-box methods (MORE [Abdolmaleki et al., 2015, Hüttenrauch and Neumann, 2024], CMA-ES[Hansen, 2023])
- Contextual: Linear (C-MORE [Tangkaratt et al., 2017]) and non-linear (BBRL TRPL [Otto et al., 2022]) mapping from context to MP
- Diverse Skill Learning: Linear MoE (LADIPS [End et al., 2017])
- Diverse Skill and Curriculum Learning: Linear MoE (SVSL [Celik et al., 2022]) and non-linear MoE (Di-SkilL [Celik et al., 2023])

Episode-Based vs Step-Based RL

- Step-based
  - Standard random exploration unsuitable for robotics
  - Each transition (s, a, r, s') is a data point
- Episode-based
  - Smooth exploration
  - Each episode $(\boldsymbol{\tau}, R)$ is a data point

Can we be more sample efficient than pure episode-based RL while keeping smooth exploration?





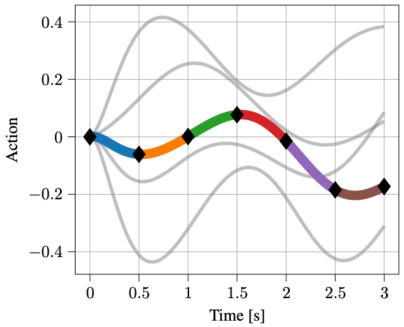
Source: [Li et al., 2024]

Temporally Correlated Episodic RL [Li et al., 2024]

- Similar setup to previous BBRL approach
- Trajectory segments as data points
- Policy update based on segment likelihood

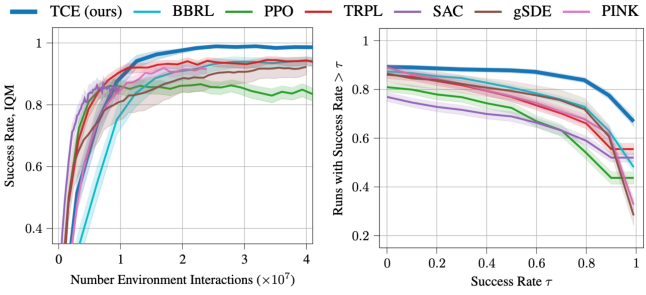
TCE

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{old}}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{p([\boldsymbol{y}_t]_{t=t_k:t'_k} \mid \boldsymbol{s})}{p_{old}([\boldsymbol{y}_t]_{t=t_k:t'_k} \mid \boldsymbol{s})} A^{\pi_{old}}(\boldsymbol{s}_{t_k}, [\boldsymbol{y}_t]_{t=t_k:t'_k})\right]$$

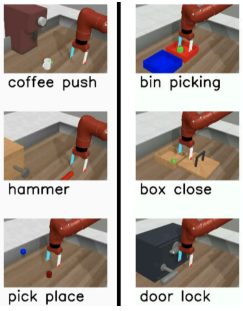


Source: [Li et al., 2024]

Temporally Correlated Episodic RL



Source: [Li et al., 2024]



Source: [Yu et al., 2021]

Conclusion

- ERL better suited for robotics due to smooth exploration
- ERL good for sparse and non Markovian rewards
- Policy search objective robust to noise compared to classical BB optimizers
- Trade-off expressiveness vs complexity
- MP design may need domain knowledge and tuning

References i

[Abdolmaleki et al., 2015] Abdolmaleki, A., Lioutikov, R., Peters, J. R., Lau, N., Pualo Reis, L., and Neumann, G. (2015).
Model-based relative entropy stochastic search.
In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[Celik et al., 2023] Celik, O., Taranovic, A., and Neumann, G. (2023).
Reinforcement Learning of Diverse Skills using Mixture of Deep Experts.
In *Intrinsically-Motivated and Open-Ended Learning Workshop @NeurIPS2023*.

[Celik et al., 2022] Celik, O., Zhou, D., Li, G., Becker, P., and Neumann, G. (2022).
Specializing Versatile Skill Libraries using Local Mixture of Experts.
In *Proceedings of the 5th Conference on Robot Learning*, pages 1423–1433. PMLR.

References ii

[End et al., 2017] End, F., Akrour, R., Peters, J., and Neumann, G. (2017).
Layered direct policy search for learning hierarchical skills.
In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6442–6448.

[Hansen, 2023] Hansen, N. (2023).
The CMA Evolution Strategy: A Tutorial.

[Hüttenrauch and Neumann, 2024] Hüttenrauch, M. and Neumann, G. (2024).
Robust Black-Box Optimization for Stochastic Search and Episodic Reinforcement Learning.
JMLR (in submission).

References iii

[Ijspeert et al., 2013] Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., and Schaal, S. (2013).
Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors.
Neural Computation, 25(2):328–373.

[Li et al., 2023] Li, G., Jin, Z., Volpp, M., Otto, F., Lioutikov, R., and Neumann, G. (2023).
ProDMP: A Unified Perspective on Dynamic and Probabilistic Movement Primitives.
IEEE Robotics and Automation Letters, 8(4):2325–2332.

[Li et al., 2024] Li, G., Zhou, H., Roth, D., Thilges, S., Otto, F., Lioutikov, R., and Neumann, G. (2024).
Open the Black Box: Step-based Policy Updates for Temporally-Correlated Episodic Reinforcement Learning.
In *The Twelfth International Conference on Learning Representations*.

References iv

[Neumann, 2017] Neumann, G. (2017).
Seminar talk: Information-theoretic policy search methods for learning versatile, reusable skills.
https://www.robots.ox.ac.uk/~seminars/seminars/Extra/2017_01_23_GerhardNeumann.pdf.

[Neumann, 2018] Neumann, G. (2018).
Iros tutorial: Movement Primitives 2: Time-Dependent Primitives.
https://www.idiap.ch/project/iros2018-tutorial/iros2018-tutorial-part3.pdf.

[Otto et al., 2020] Otto, F., Becker, P., Ngo, V. A., Ziesche, H. C. M., and Neumann, G. (2020).
Differentiable Trust Region Layers for Deep Reinforcement Learning.
In *International Conference on Learning Representations*.

References v

[Otto et al., 2022] Otto, F., Celik, O., Zhou, H., Ziesche, H., Ngo, V. A., and Neumann, G. (2022).
Deep black-box reinforcement learning with movement primitives.
In *Conference on Robot Learning*, pages 1244–1265. PMLR.

[Otto et al., 2023] Otto, F., Zhou, H., Celik, O., Li, G., Lioutikov, R., and Neumann, G. (2023).
MP3: Movement Primitive-Based (Re-)Planning Policy.
(arXiv:2306.12729).

[Paraschos et al., 2013] Paraschos, A., Daniel, C., Peters, J. R., and Neumann, G. (2013).
Probabilistic movement primitives.
Advances in neural information processing systems, 26.

References vi

[Tangkaratt et al., 2017] Tangkaratt, V., Van Hoof, H., Parisi, S., Neumann, G., Peters, J., and Sugiyama, M. (2017).
Policy Search with High-Dimensional Context Variables.
Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).

[Yu et al., 2021] Yu, T., Quillen, D., He, Z., Julian, R., Narayan, A., Shively, H., Bellathur, A., Hausman, K., Finn, C., and Levine, S. (2021).
Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning.
(arXiv:1910.10897).