DS-GA 3001 Course Project Proposal

Raman Kumar Jha
NYU Tandon School of Engineering
Brooklyn, New York
ramanjha@nyu.edu

Amey Joshi
NYU Tandon School of Engineering
Brooklyn, New York
avj2036@nyu.edu

Meta-diff MPC [#4]

Abstract

Addressing the limitations of Trajectory-Dynamic Model Predictive Control 2 (TD-MPC2), this paper introduces enhancements focused on improving exploration, task adaptation, and reward signal utilization. While TD-MPC2 demonstrates strong performance in continuous control tasks by performing local trajectory optimization in the latent space of a learned implicit world model, its reliance on a policy prior and Model Predictive Path Integral (MPPI) planning limits exploration and adaptation to new tasks. Furthermore, TD-MPC2's performance is highly dependent on reward signals, with the method for leveraging those rewards for large-scale pretraining remaining an open problem. To overcome these limitations, this work explores the integration of a diffusion-based policy to enhance exploration by expressing complex, multimodal action distributions. Additionally, the method considers the incorporation of Hindsight Experience Replay (HER) principles to address reward sparsity and improve learning efficiency, and discusses the adoption of generalized reward signals to improve task completion. By addressing these key limitations, the proposed approach aims to broaden the applicability and improve the robustness of TD-MPC2 across a wider range of robotic tasks and environments.

1 Introduction

Recent advancements in reinforcement learning(RL) have demonstrated remarkable capabilities in solving complex control tasks. However, the challenge of developing agents that can quickly adapt to new environments remains a key focus in the field. This project proposal aims to extend the TD-MPC2 [1] algorithm, a state-of-the-art model-based reinforcement learning method, by incorporating meta-learning techniques and diffusion models to enhance its generalization and exploration capabilities.

Our approach builds upon the robust world model and planning framework of TD-MPC2,[1] introducing two key innovations:

1. A diffusion-based policy prior that leverages the power of Denoising Diffusion Probabilistic Models (DDPMs) [2] to generate diverse and adaptive action sequences.
2. A meta-learning framework for task embeddings, enabling rapid adaptation to new environments through efficient task inference [3] and representation [4].

By combining these elements, we aim to create an agent capable of quick adaptation and effective exploration across a wide range of tasks, potentially bridging the gap between single-task proficiency and multi-task generalization in reinforcement learning [5].

2 Related Work

Embodied learning and reinforcement learning (RL) are increasingly converging fields. They aim to develop agents capable of generalizing across diverse tasks and environments. Recent advancements in model-based RL, such as TD-MPC2 [1], and meta-learning frameworks for skill acquisition, provide a strong foundation for challenges multitask learning, generalization, and embodied intelligence.

2.1 Multitask Learning and Generalization

TD-MPC2 [1] introduces a scalable framework for training generalist agents across multiple tasks, domains, and embodiments. It is an extension of the previous TD-MPC [6] method, where it has shown several improvement from the TD-MPC. Diffusion models have shown remarkable success in generating diverse and high-quality samples across various domains. Janner et al. (2022) [7] demonstrated the potential of diffusion models for planning in RL, showcasing their ability to model complex action distributions1. Our proposal to incorporate a conditional Denoising Diffusion Probabilistic Model (DDPM) as a policy prior in TD-MPC2 aims to leverage these strengths for improved exploration and generalization.

2.2 Meta-Learning for Skill Acquisition

Meta-learning has enabled agents to adapt quickly to new tasks by leveraging prior knowledge. Gershman et al. (2017) [8] have explored the context based learning approach for this task. Finn et al. (2017) [9] introduced Model-Agnostic Meta-Learning (MAML), a seminal work that laid the foundation for gradient-based meta-learning in RL. Velazquez-Vargas et al. (2023) [10] explored the role of contextual cues in separating visuomotor mappings, showing that both humans and meta-learning agents benefit from contextual information for efficient skill acquisition. Eysenbach et at [2022] [11] have shown the use of contrastive learning, for goal-oriented reinforcement learning [12]. This aligns with the broader goal of designing agents that can dynamically adapt their representations based on task-specific requirements.

2.3 World Models and Planning

TD-MPC2 itself represents a significant advancement in model-based RL, building on the success of world models for planning and control [13]. The original work by Hansen et al. (2023) demonstrated the scalability and robustness of TD-MPC2 across diverse continuous control tasks [14]. Our proposed extensions aim to further enhance its capabilities in meta-learning and generalization scenarios.

3 Proposed Approach

We propose **DiffusionMPC** , a framework that combines the strengths of TD-MPC2 with diffusion models for enhanced planning and control. Our approach builds on recent advances in diffusion-based control policies and diffusion model predictive control, integrating them with TD-MPC2's robust world model architecture.

1. **Diffusion-Based Trajectory Generation:** We leverage diffusion models [2, 7, 14], to learn multi-step action proposals and dynamics models from offline datasets. Unlike traditional approaches that use deterministic models, diffusion models can effectively capture multi-modal distributions in high-dimensional spaces [15]. This enables the generation of diverse, high-quality action sequences, capturing the full distribution of viable strategies for a given task, and avoiding compounding errors through trajectory-level modeling.
2. **Constraint-Aware Planning:** We incorporate explicit constraint handling into the diffusion process using techniques from Diffusion Predictive Control with Constraints (DPCC) [16]. This allows our system to adapt to novel constraints not seen during training, generate dynamically feasible trajectories [10], and balance task performance with constraint satisfaction.
3. **Hindsight Experience Replay Integration:** To address the challenge of sparse rewards, we integrate Hindsight Experience Replay (HER) [17] into our training pipeline. This approach, transforms failed trajectories into successful ones by retroactively changing the goal, enables

efficient learning from sparse, binary rewards without complex reward engineering, and creates an implicit curriculum as the agent naturally progresses from simple to complex goals [12].

Technical Approach:
DiffusionMPC combines four key components:

- **Multi-step Diffusion Models:** We train diffusion models [2, 7, 14] to generate both action proposals and predict dynamics over multiple timesteps, avoiding the compounding error problem faced by single-step models.
- **Latent World Model:** Following TD-MPC2, we maintain an implicit world model that enables efficient planning in a compressed latent space. [13]
- **Hindsight Relabeling:** We implement a modified HER [17] approach tailored to our latent world model, allowing for efficient learning even when rewards are sparse or delayed.
- **Constraint Projection:** We integrate model-based projections [10] into the denoising process to ensure constraint satisfaction while maintaining performance on the learned task.

The planning process follows this algorithm:

- Generate multiple trajectory candidates using the diffusion model.
- Project trajectories to satisfy constraints.
- Score trajectories using the learned world model and reward function.
- Execute the first action of the highest-scoring trajectory.

4 Expected Results, and Experiment Plans

Experimental Methods: We will evaluate DiffusionMPC through comprehensive experiments designed to test its core capabilities:

- Benchmark Performance: Evaluate against state-of-the-art methods (TD-MPC2 [1], D-MPC [18], SAC [19]) on standard benchmarks including D4RL [20] and dm_control [21] tasks.
- Novel Constraint Adaptation: Test the ability to satisfy constraints not seen during training, such as obstacle avoidance or joint limits.
- Novel Reward Optimization: Assess performance on tasks with reward functions different from those used during training.
- Dynamics Adaptation: Measure adaptation speed to changes in environment dynamics, such as simulated motor defects.
- Ablation Studies: Systematically evaluate the contribution of each component through controlled experiments.

Expected Impact and Applications: DiffusionMPC represents a significant advancement in robot learning by addressing key limitations in current approaches through a principled integration of diffusion models and model predictive control. The practical implications extend to:

- Robust Manipulation: Enabling robots to perform complex manipulation [22] tasks requiring multimodal[23] action distributions
- Safe Deployment: Ensuring constraint satisfaction even in novel environments.
- Flexible Adaptation: Allowing systems to quickly adapt to new tasks, constraints, and dynamics.
- Sample-Efficient Learning: Reducing the data requirements for learning complex skills in sparse reward settings

References

[1] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[3] Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

[4] Zichen Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. *Advances in Neural Information Processing Systems*, 37:33933–33961, 2025.

[5] Brandon Amos, Ivan Jimenez, Jacob Sacks, Byron Boots, and J Zico Kolter. Differentiable mpc for end-to-end planning and control. *Advances in neural information processing systems*, 31, 2018.

[6] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

[7] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[8] Samuel J Gershman. Context-dependent learning and causal structure. *Psychonomic bulletin & review*, 24:557–565, 2017.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[10] Carlos A Velazquez-Vargas, Isaac Ray Christian, Jordan A Taylor, and Sreejan Kumar. Learning to abstract visuomotor mappings using meta-reinforcement learning. *arXiv preprint arXiv:2402.03072*, 2024.

[11] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

[12] Kazuki Takahashi, Tomoki Fukai, Yutaka Sakai, and Takashi Takekawa. Goal-oriented inference of environment from redundant observations. *Neural Networks*, 174:106246, 2024.

[13] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[14] Michiel Nikken, Nicolò Botteghi, Wesley Roozing, and Federico Califano. Denoising diffusion planner: Learning complex paths from low-quality demonstrations. *arXiv preprint arXiv:2410.21497*, 2024.

[15] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[16] Ralf Römer, Alexander von Rohr, and Angela P Schoellig. Diffusion predictive control with constraints. *arXiv preprint arXiv:2412.09342*, 2024.

[17] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[18] F Cortes, D Linares, Diego Patiño, and Kamilo Melo. A distributed model predictive control (d-mpc) for modular robots in chain configuration. In *IX Latin American Robotics Symposium and IEEE Colombian Conference on Automatic Control, 2011 IEEE*, pages 1–6. IEEE, 2011.

[19] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[20] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[21] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

[22] Nolan Fey, Gabriel B Margolis, Martin Peticco, and Pulkit Agrawal. Bridging the sim-to-real gap for athletic loco-manipulation. *arXiv preprint arXiv:2502.10894*, 2025.

[23] Yingbai Hu, Fares J Abu-Dakka, Fei Chen, Xiao Luo, Zheng Li, Alois Knoll, and Weiping Ding. Fusion dynamical systems with machine learning in imitation learning: A comprehensive overview. *Information Fusion*, page 102379, 2024.