Imitation Learning on Bimanual Robots

by

Ravi Swaroop Rayavarapu

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2023 by the
Graduate Supervisory Committee:

Heni Ben Amor, Chair
Nakul Gopalan
Ransalu Senanayake

ARIZONA STATE UNIVERSITY

December 2023

ABSTRACT

Bimanual robot manipulation, involving the coordinated control of two robot arms, holds great promise for enhancing the dexterity and efficiency of robotic systems across a wide range of applications, from manufacturing and healthcare to household chores and logistics. However, enabling robots to perform complex bimanual tasks with the same level of skill and adaptability as humans remains a challenging problem. The control of a bimanual robot can be tackled through various methods like inverse dynamic controller or reinforcement learning, but each of these methods have their own problems. Inverse dynamic controller cannot adapt to a changing environment, whereas Reinforcement learning is computationally intensive and may require weeks of training for even simple tasks, and reward formulation for Reinforcement Learning is often challenging and is still an open research topic.

Imitation learning, leverages human demonstrations to enable robots to acquire the skills necessary for complex tasks and it can be highly sample-efficient and reduces exploration. Given the advantages of Imitation learning we want to explore the application of imitation learning techniques to bridge the gap between human expertise and robotic dexterity in the context of bimanual manipulation.

In this thesis, an examination of the Implicit Behavioral Cloning imitation learning algorithm is conducted. Implicit behavioral cloning aims to capture the fundamental behavior or policy of the expert by utilizing energy-based models, which frequently demonstrate superior performance when compared to explicit behavior cloning policies. The assessment encompasses an investigation of the impact of expert demonstrations' quality on the efficacy of the acquired policies. Furthermore, computational and performance metrics of diverse training and inference techniques for energy-based models are compared.

ACKNOWLEDGMENTS

I am immensely grateful to my family and friends for their unwavering support throughout this journey. A special mention to Micheal Drolet for his contributions on this project. I would also like to extend my thanks to the members of the Interactive robotics Lab for their encouragement, both professional and personal level. I was fortuitous to be a part of discussions which enlightened me and supported in expanding domain of my knowledge. I would also like to extend my deepest gratitude to Dr. Heni Ben Amor for giving me the opportunity to be a part of the lab and for his insightful guidance in carrying out rigorous and meaningful research. I cannot overstate how much I appreciate the chance to work with him and I appreciate the advice he provided along the way.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

A bimanual robot is a robotic system engineered with the integration of two manipulator arms or end-effectors that operate in a harmonious, coordinated fashion. These robotic entities are designed to emulate the dexterity and adaptability inherent in human arms, thus equipping them with the capacity to adeptly tackle a spectrum of tasks. Bimanual robots exhibit remarkable versatility, not only excelling in activities that necessitate the collaborative synergy of both arms but also demonstrating prowess in handling intricate and multifaceted assignments. Their multifarious utility spans across a multitude of industries, encompassing manufacturing, healthcare, and research.

It's noteworthy that a predominant portion of current research endeavors has predominantly fixated on the domain of single-arm manipulation, leading to the establishment of well-recognized control methodologies for singular robotic arms. On the contrary, the domain of controlling and manipulating bimanual robots remains an open frontier within the research landscape. Historically, prior investigations into bimanual robots have predominantly leaned towards classical control-based approaches, exemplified by works like Mirrazavi Salehian *et al.* (2018) and reinforcement learning based policies like Bersch *et al.* (2011). However, a recent paradigm shift has witnessed a surge in research embracing learning-based approaches for the control and manipulation of bimanual robots, with Stepputtis *et al.* (2022) exemplifying one such innovative approach. This approach hinges upon the utilization of motor primitives to effectively manage the intricate manipulation aspects of contact-rich bimanual robots. While this methodology has demonstrated its effectiveness, our research endeavors are inclined towards eliminating the inherent inductive bias associated with motor primitives.

While this approach is shown to be effective, we seek to remove the inductive bias of using motor primitives and instead explore neural network-based approaches. In doing so, we seek to use a more general class of function approximation that can potentially capture correlations not expressed by the Radial Basis Function weights given by DMPs.

Imitation learning stands as a promising middle ground in the domain of robotic control strategies, positioned between traditional classical control-based approaches, which lack adaptability to the dynamic nuances of real-world environments, and the reinforcement learning paradigm. It offers the notable advantages of heightened sample efficiency and the removal of the exploration-exploitation trade-off. While simpler imitation learning techniques like behavior cloning are available, they are not without their drawbacks, particularly the issue of error accumulation.

The primary focus of this thesis is to conduct an in-depth assessment of Implicit Behavioral Cloning Florence *et al.* (2022), a distinctive approach within the realm of imitation learning. What sets Implicit Behavioral Cloning apart is its departure from the conventional direct mapping of actions from observed data. Instead, it harnesses the power of implicit models, particularly the composition of *argmin* in conjunction with a continuous energy function represented as E_θ During the inference phase, implicit regression takes center stage, characterized by the optimization of the optimal action a. This optimization process can be accomplished through techniques such as sampling or gradient descent, ensuring that the model adapts to the intricacies of the task at hand. To facilitate this evaluation, we employ the Mujoco advanced physics simulator, a powerful tool that enables us to model and simulate the performance of a bimanual robot in a variety of complex tasks. One such task involves the insertion of a male adaptor, featuring four holes, into a female adaptor equipped with four pegs, all with a level of precision measured in millimeters.

Chapter 2

METHODOLOGY

For the purpose of this experiment, we built a custom environment within the Gym interface, a choice made for its user-friendly features and compatibility with various other reinforcement learning frameworks. To achieve a high level of realism and precision in our simulations, we integrated the Mujoco physics engine into this environment, leveraging its exceptional capabilities in modeling contact dynamics and providing accurate simulations.

Furthermore, as a pivotal component of our experimental setup, we developed an admittance controller to control bimanul robot and to regulate the interaction dynamics between the robot and the environment. This sophisticated control strategy is centered on the management of forces and motion at the precise points of contact between the robotic system and the environment. The objective is to ensure that the robot's behavior adheres to predefined criteria and desired performance benchmarks during interactions. We have chosen admittance controller for its compliance to external forces and promptly adapt to any fluctuations or irregularities encountered in the environment, such as variations in surface conditions or unexpected contact forces. Our choice of the admittance controller stems from its well-established effectiveness, particularly in applications that demand robots to engage with humans or delicately manipulate objects exhibiting a broad spectrum of characteristics. The compliance afforded by this controller is instrumental in ensuring the safety and seamless quality of interactions within such contexts.

Imitation Learning necessitates the initial provision of demonstrations by an expert to the robot. Subsequently, the robot engages in correspondence matching to

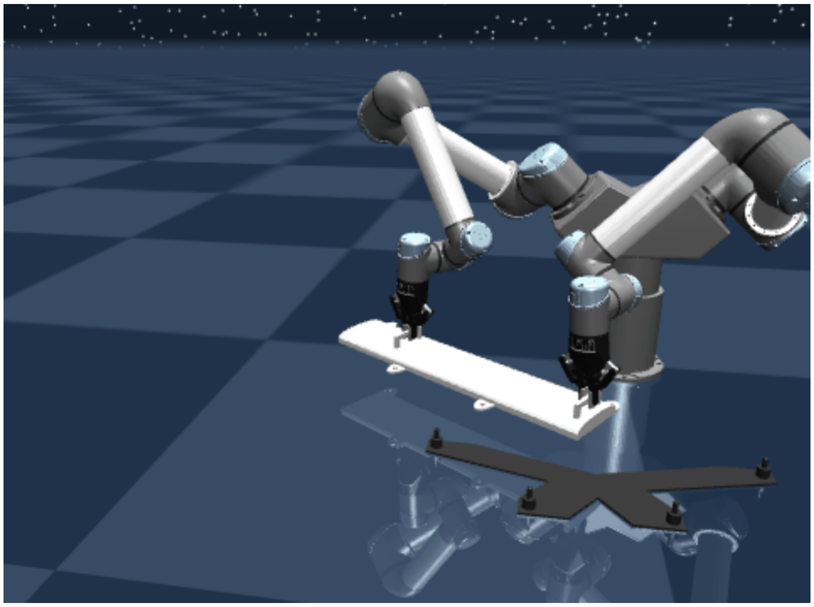

Figure 2.1: A bimanual robot placing a female adaptor onto a male counterpart in Mujoco engine

closely approximate the original trajectories of the expert, especially when the demonstrations are conveyed through kinesthetic teaching or performed by the original robot. It is assumed that the robot trajectories are derived from an arbitrary expert, involving the collection of right and left end-effector positions and orientations at each time step. To convert these original demonstrations into usable ones executed by the robot itself, an operational space controller is employed. This controller regulates the end effectors based on their task space errors, while simultaneously managing the robot's torso.

2.1 Operational Space Controller (OSC)

The operational space controller (OSC) shown in Peters and Schaal (2008) is used in this work to facilitate learning in the task space. This controller allows precise control of the end-effector's position and orientation, enabling the robot to perform tasks with accuracy and adaptability. It is particularly useful in applications where it is essential to control the end-effector's behavior in a specific operational space, regardless of the robot's joint configurations. Operational space controllers can adjust the compliance of each arm to accommodate such interactions, ensuring safety and adaptability. Given a desired 3D position and 3D rotation for each DoF j, we calculate the respective (scaled) task space errors as:

$$X_{task} = [\tilde{\rho}^1, \tilde{\phi}^1, \tilde{\rho}^2, \tilde{\phi}^2, ..., \tilde{\rho}^J, \tilde{\phi}^J] \tag{2.1}$$

We adopt the Jacobian pseudo-inverse method, using the projected influence of inertia in the configuration space:

$$\mathbf{M}_x(\mathbf{q}) = (\mathbf{J}(\mathbf{q})\mathbf{M}(\mathbf{q})^{-1}\mathbf{J}(\mathbf{q})^\top)^\dagger. \tag{2.2}$$

for each arm separately. Because we are controlling two UR5 arms, each with six DoF, and one base joint, to get the combined jacobian we concatenate the jacobians of both the arms horizontally to get the combined jacobian of the bimanual robot with Jacobian $\mathbf{J}(\mathbf{q}) \in \mathbb{R}^{12\times13}$ then $\mathbf{q} \in \mathbb{R}^{13}$, similarly the inertia matrix M of each arm is concatenated diagonally to get the combined inertia matrix $\mathbf{M}(\mathbf{q}) \in \mathbb{R}^{13\times13}$, and forces due to gravity $\mathbf{g}(\mathbf{q}) \in \mathbb{R}^{13}$. The system utilizes force/torque sensors on both end effectors, measured in the task space F_a. These sensors enable the robot to respond to external forces by adjusting its movement accordingly.

As a result, we obtain the following force vector used to control the robot:

$$\mathbf{u} = -\mathbf{J}(\mathbf{q})^\top\mathbf{M}_x(\mathbf{q})\mathbf{x} - \mathbf{K}_v\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) - \mathbf{F}_a \tag{2.3}$$

This force vector is used to control the bimanual robot.This approach offers the benefit of not necessitating the specification of the torso angle, enabling flexibility in the null space controller's adjustment to achieve its position in a manner that is kinematically feasible.

2.2 Environment

The environment built using gym API, since its simple, pythonic and capable on representing general RL problems. We collect n demonstrations from the algorithm *generateExpertDemo* who's state space and action space can we set according to the evaluation the performance of different algorithms on all the environments, without changing the base environment. The environment gives a reward proportional to the experts trajectory, the RL algorithm has neither access to the reward nor the expert. It is only used to compare different policies. One of the design choices made to improve the performances of the algorithms is to use a 6D rotation matrix representation given by Hempel *et al.* (2022), it is used address problem of ambiguous rotation labels by introducing the rotation matrix formalism for our ground truth data, it reduces the state space of the environment.

2.2.1 6D rotation matrix representation

A commonly used and convenient way to represent orientation is through Euler angles. However, this representation has its limitations, such as susceptibility to gimbal lock, wherein multiple sets of rotation parameters can produce the same visual head pose. This can pose challenges for neural networks when attempting to accurately learn the pose. In contrast, while the quaternion representation avoids the gimbal lock issue, it introduces an ambiguity due to its antipodal symmetry. Instead, we adopt the approach as proposed by Zhou et al. and implement the Gram-Schmidt mapping directly within the representation by omitting the last column vector of the rotation matrix. This transformation reduces the original 3x3 matrix to a six-parameter rotation representation, which has been documented to yield reduced errors in direct regression.

$$g\left(\begin{bmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{bmatrix}\right) = \begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \tag{2.4}$$

The predicted 6D representation matrix can then mapped back into SO(3).

$$f\left(\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix}\right) = \begin{bmatrix} | & | & | \\ b_1 & b_2 & b_3 \\ | & | & | \end{bmatrix} \tag{2.5}$$

In this process, the retained column vector is determined through a cross product operation, which guarantees that the resulting 3x3 matrix adheres to the orthogonality constraint.

$$b_1 = \frac{a_1}{\|a_1\|}$$

$$b_2 = \frac{u_2}{\|u_1\|}, u_2 = a_2 - (b_1.a_2)b_1$$

$$b_3 = b_1 \times b_2$$

2.3 Algorithms

Behavioral cloning (BC) presented by Peters and Schaal (2008) stands as one of the most straightforward machine learning approaches for endowing real-world robots with the capability to acquire essential skills. BC formulates the process of imitating expert demonstrations as a supervised learning problem. Despite the presence of legitimate concerns, both empirical and theoretical, Tu *et al.* (2022); Ross *et al.* (2011) showed the limitations regarding its such as the issue of error accumulation—BC shown by has, in practice, demonstrated the remarkable potential to yield impressive outcomes. Real robots have been able to generalize complex behaviors to novel, unstructured scenarios, a feat that has garnered considerable attention and acclaim.

However, it is crucial to acknowledge that a substantial body of research has been dedicated to the development of novel imitation learning methods aimed at mitigating the known constraints associated with BC. Implicit Behavioral Cloning (IBC) looks to reformulate the fundamental aspect of imitation learning: the inherent structure of the policy itself. Diverging from conventional supervised learning algorithms, IBC policies do not adhere to the customary model representation of continuous feed-forward relationships, wherein actions a are directly mapped from input observations o through a function $F(o)$.Implicit Behavioral Cloning (IBC) is a supervised learning approach using Energy-Based Models. IBC policy is given by:

$$a = \underset{\mathbf{a} \in \mathcal{A}}{\operatorname{argmin}} \, E_\theta(\mathbf{o}, \mathbf{a}) \tag{2.6}$$

IBC is training using the Negative Counter Example (NCE) loss function, a technique that involves the generation of negative counter-examples derived from the expert's demonstrations to instruct the model. Within this framework, a distinctive aspect is the assignment of energies to pairs of observations and corresponding actions, and the policy selects the action associated with the lowest energy level. This approach distinguishes IBC by virtue of its capacity to effectively manage discontinuities that might emerge in the standard regression setting, in contrast to behavioral cloning, which may resort to mere interpolation.

With a trained energy model $E_\theta(x, y)$, implicit inference can be performed with stochastic optimization to solve $\hat{y} = argmin_y \; E_\theta(x, y)$, we present results with three different EBM training and inference methods discussed below

2.3.1 Derivative free optimization DFO

DFO (Conn *et al.* (2009)) does not rely on gradient information, making it suitable for optimizing functions that are non-smooth and discontinuous. Training with derivative free optimization is simple counter-examples are drawn from the uniform random distribution $\tilde{y} \sim U(y_{min}, y_{max})$. Training consists of drawing batches of data, sampling counterexamples for each sample in each batch, and applying $\mathcal{L}_{InfoNCE}$. Given a trained energy model $E_\theta(x, y)$, we use the following derivative-free optimization algorithm to perform inference:

```
Algorithm 1: Derivative-free optimization
  Output: $\hat{y}$
  $\{\tilde{y}_i\}_{i=1}^{N_{samples}} \sim U(y_{min}, y_{max}), \sigma = \sigma_{ini}$;
  for $iter$ in 1,2,...$N_{iter}$ do
    $\{E_i\}_{i=1}^{N_{samples}} = \{E_\theta(x, \tilde{y}_i)\}_{i=1}^{N_{samples}}$ ;    // compute energies
    $\{\tilde{p}_i\}_{i=1}^{N_{samples}} = \{\frac{e^{-E_i}}{\sum_{j=1}^{N_{samples}} e^{-E_j}}\}$ ;    // soft max
    if $iter < N_{iter}$ then
      $\{\tilde{y}_i\}_{i=1}^{N_{samples}} \leftarrow\sim \text{Multinomial}(N_{samples}, \{\tilde{p}_i\}_{i=1}^{N_{samples}}, \{\tilde{y}_i\}_{i=1}^{N_{samples}})$ ;
      // resample
      $\{\tilde{y}_i\}_{i=1}^{N_{samples}} \leftarrow \{\tilde{y}_i\}_{i=1}^{N_{samples}} + \mathcal{N}(0, \sigma)$ ;    // add noise
      $\{\tilde{y}_i\}_{i=1}^{N_{samples}} = \text{clip}(\{\tilde{y}_i\}_{i=1}^{N_{samples}}, y_{min}, y_{max})$ ;    // clip to y bounds
      $\sigma \leftarrow K\sigma$ ;    // shrink sampling variance
    end
  end
  $\hat{y} = argmax\{\tilde{p}_i\}, \{\tilde{y}_i\}$
```

Where $\text{Multinomial}(N_{samples}, \{\tilde{p}_i\}_{i=1}^{N_{samples}}, \{\tilde{y}\}_i^{N_{samples}})$ refers to sampling $N_samples$ times from the multinomial distribution with probabilities $\{\tilde{p}_i\}_{i=1}^{N_{samples}}$

2.3.2 Autoregressive DFO

In the autoregressive version (Nash and Durkan (2019)) we interleave training and inference with m models i.e. one model $E_\theta^j(x, y^{:j})$ for each dimension $j = 1, 2, ...m$. This isolates sampling to one degree of freedom at a time, and enables scaling to higher dimensional action spaces.

Algorithm 2: Autoregressive Derivative-free optimization

```
Output: $\hat{y}$
$\{\tilde{y}\}_{i=1}^{N_{samples}} \sim U(y_{min}, y_{max}), \sigma = \sigma_{ini}$;
for $iter$ in 1,2,...$N_{iter}$ do
    for $j$ in 0,1,...,$m$ do
        $\{E_i\}_{i=1}^{N_{samples}} = \{E_\theta^j(x, \tilde{y}_i^{:j})\}_{i=1}^{N_{samples}}$ ;          // compute energies
        $\{\tilde{p}_i\}_{i=1}^{N_{samples}} = \{\frac{e^{-E_i}}{\sum_{j=1}^{N_{samples}} e^{-E_j}}\}$ ;          // soft max
    end
    if $iter < N_{iter}$ then
        $\{\tilde{y}_i^{:j}\}_{i=1}^{N_{samples}} \leftarrow\sim \text{Multinomial}(N_{samples}, \{\tilde{p}_i\}_{i=1}^{N_{samples}}, \{\tilde{y}_i^{:j}\}_{i=1}^{N_{samples}})$ ;
        // resample
        $\{\tilde{y}_i^j\}_{i=1}^{N_{samples}} \leftarrow \{\tilde{y}_i^j\}_{i=1}^{N_{samples}} + \mathcal{N}(0, \sigma)$ ;          // add noise
        $\{\tilde{y}_i\}_{i=1}^{N_{samples}} = \text{clip}(\{\tilde{y}_i\}_{i=1}^{N_{samples}}, y_{min}, y_{max})$ ;          // clip to y bounds
        $\sigma \leftarrow K\sigma$ ;          // shrink sampling variance
    end
end
$\hat{y} = argmax\{\tilde{p}_i\}, \{\tilde{y}_i\}$
```

2.3.3 Langevin sampling

Langevin Monte Carlo (MCMC) (Du *et al.* (2020); Du and Mordatch (2019)) is a gradiant based opitmizing algorithm used for sampling from probability distributions. It's particularly effective for sampling from high-dimensional and complex distributions, for garduient based Langevin MCMC we use the approach, which uses stochastic gradient Langevin dynamics (SGLD):

$$ {}^{k}\tilde{y}_i^j = {}^{k}\tilde{y}_i^j - \lambda(\frac{1}{2}\Delta_y E_\theta(x_i, {}^{k-1}\tilde{y}_i^j) + \omega^k), \omega^k \sim \mathcal{N}(0, \sigma) \tag{2.7} $$

we initialize $\{{}^{0}\tilde{y}\}$ from the uniform distribution,but then optimize these contrastive samples with MCMC. We use a polynomially decaying schedule for the stepsize λ.

Chapter 3

RESULTS

3.1 Algorithm Comparison

A crucial aspect to consider when comparing these methods lies in the delicate balance between the elegance of simplicity and the complexities introduced by higher-dimensional action spaces. To comprehensively assess the performance of various approaches across varying dimensions, an extensive experiment was conducted. In this experiment, all the aforementioned methods were trained using multiple environments, each characterized by an incrementally expanding action space while keeping the observation space constant. It's noteworthy that this evaluation involved training all the methods with a limited set of only 200 expert demonstrations. Each model undergoes training for 12,000 steps, as the training losses typically stabilize, and the hyperparameters remain consistent across all models.

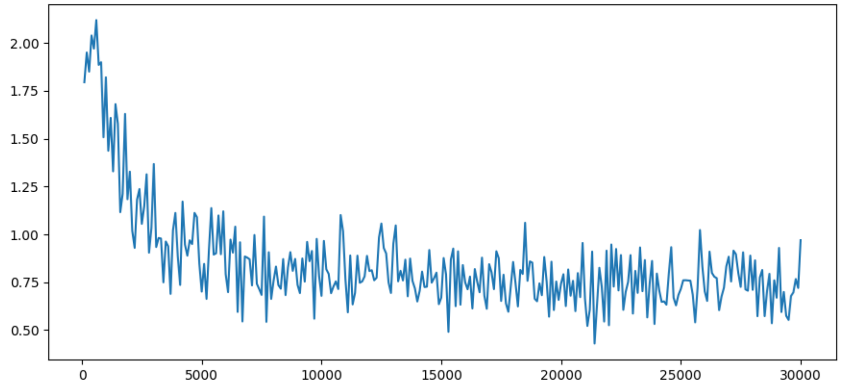

Figure 3.1: Training loss of langevin model

The results of this study unveiled a notable trend: the derivative-free optimization variant struggled to successfully solve the environments when the action space dimensionality exceeded a critical threshold, typically observed beyond N = 5 dimensions. This failure can be attributed to the well-documented ”curse of dimensionality,” which implies that as the dimensionality of the action space increases, the volume of the space grows exponentially, making it increasingly challenging to sample and explore effectively. This inherent challenge in naive sampling in high-dimensional spaces highlighted the need for more sophisticated approaches to tackle the complexity introduced by the curse of dimensionality. In contrast, both the autoregressive DFO and Langevin variants demonstrated a remarkable degree of resilience and proficiency.

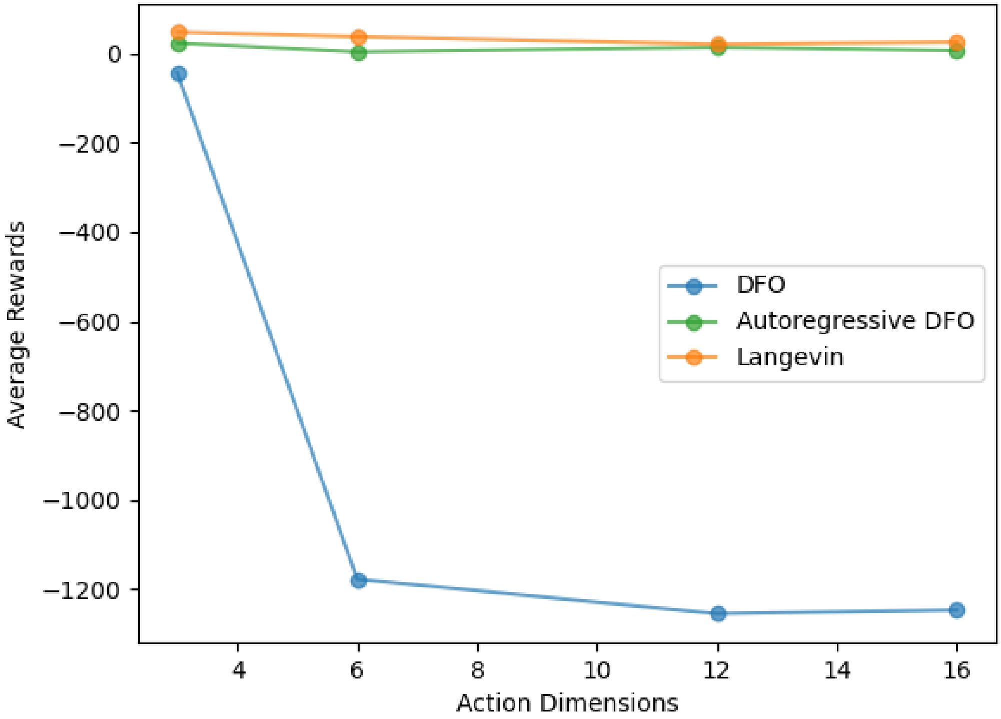

Figure 3.2: Average rewards across action dimensions

When comparing the computational efficiency of derivative-free optimization (DFO), autoregressive DFO, and Langevin methods, a clear trade-off becomes evident. DFO, characterized by its speed, emerges as the fastest among the three. However, this advantage comes at a cost, as DFO's performance tends to deteriorate significantly as the dimensionality of the action space increases, largely due to the inherent challenges posed by the curse of dimensionality and its reliance on naive sampling techniques. In contrast, the autoregressive DFO and Langevin methods exhibit commendable robustness, showcasing the ability to tackle complex high-dimensional action spaces while maintaining reliable performance. These two approaches prioritize the effective exploration of expansive state spaces, mitigating the adverse effects of dimensionality and offering a compelling solution to optimization challenges in settings where computational efficiency and performance must be balanced.

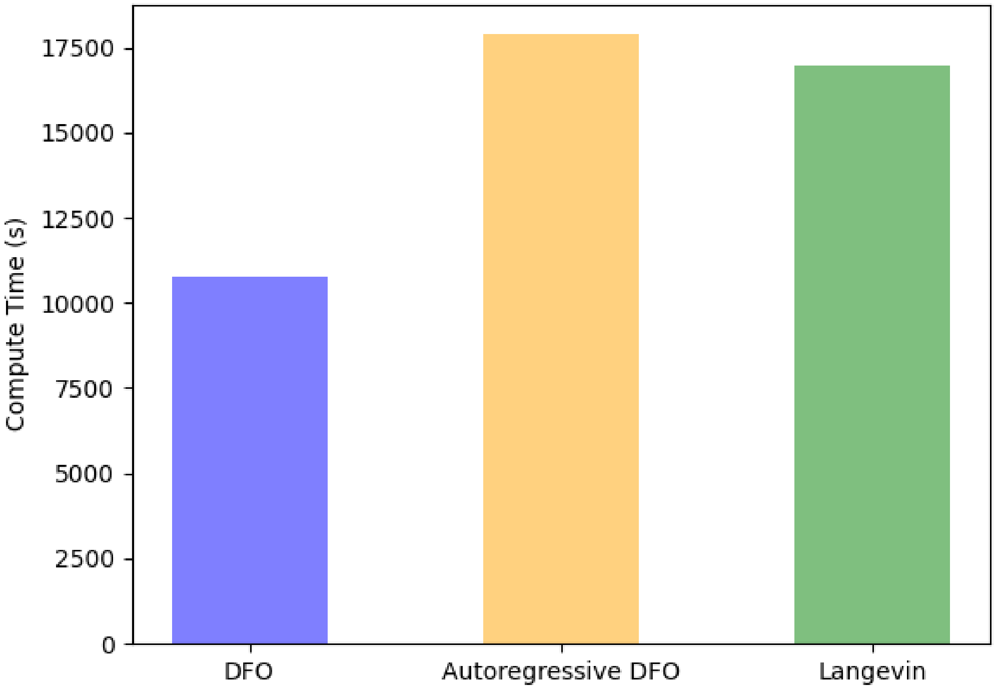

Figure 3.3: Average compute time

3.2 Expert Analysis

The success rates of a Langevin policy, trained with varying numbers of expert demonstrations (50, 100, and 200), provide an interesting insight into the algorithm's learning behavior, particularly when accounting for the introduction of randomized environmental noise. Notably, it becomes evident that the success rate of the policy improves as the number of expert demonstrations increases. The policy trained with 200 expert demonstrations achieves the highest success rate, underscoring the importance of a substantial volume of expert data in effectively training a proficient policy. This observation highlights the algorithm's demand for a maximum number of experts to capture the nuances and intricacies required for a successful policy, demonstrating the significance of data abundance in achieving optimal learning outcomes.

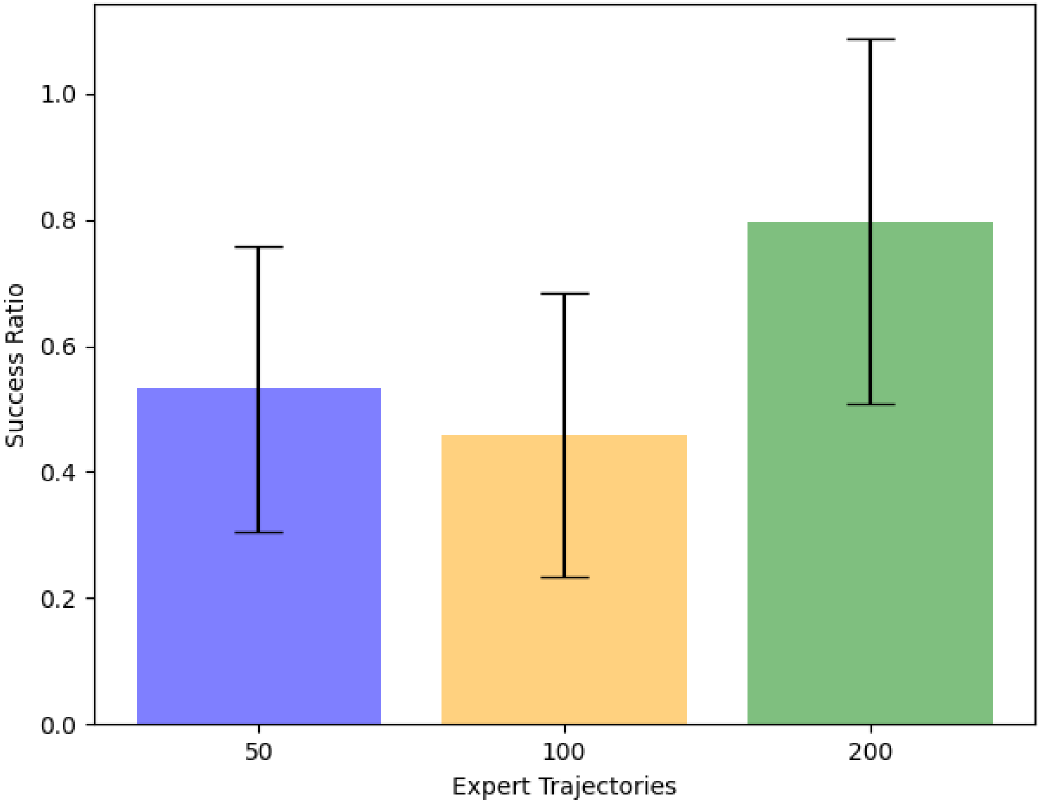

Figure 3.4: Average success ratio across expert demos

3.3 Noise Analysis

In the subsequent experiment, we do comprehensive comparative analysis, aimed at unraveling the nuanced ways in which environmental noise exerts its influence on policy performance. To conduct this investigation, we maintained uniformity in terms of expert demonstrations, with each policy receiving the same robust training dataset of 50,100,200 demonstrations. However, our manipulation came into play as we varied the levels of observation and action noises across three distinct sets. This approach allowed us to discern a compelling trend: the policy's resilience in the face of noise perturbations was notably more pronounced when it came to observation noise. Even as we escalated the magnitude of observation noise, the policy displayed a remarkable degree of stability, with its performance exhibiting only marginal deterioration. However, a striking and contrasting observation came to the fore when action noise was introduced. The policy's success rate took a substantial nosedive, and it struggled to perform with proficiency. This stark contrast underscores the heightened susceptibility of the policy to action noise perturbations, highlighting the pivotal role of noise management in the quest for robust and reliable policy execution within dynamic and stochastic environments.

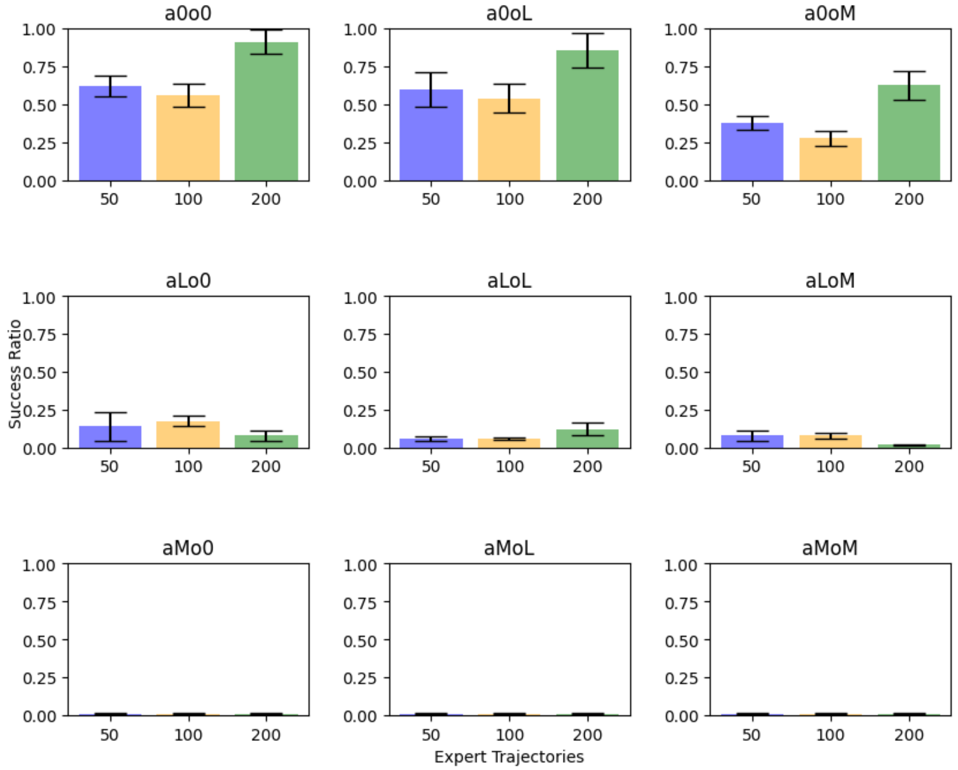

Figure 3.5: Average success ratio across environments

3.4 Sensitivity Analysis

In the forthcoming experiment, our aim is to delve into the algorithm's robustness concerning perturbations in hyperparameter tuning. We seek to evaluate the policy's performance when subjected to changes in hyperparameters compared to an ideal set of hyperparameters. To accomplish this, we imposed a constraint by introducing lower and upper bounds, effectively constituting a range of ±10 percent within which the hyperparameters were permitted to vary. Employing Latin hypercube sampling shown in Loh (1996), we randomly selected hyperparameters from this defined hypercube space and utilized these for training the model. Subsequently, we compared the performance of this model with the one trained using the meticulously tuned ideal hyperparameters. The following hyperparameters were used for this evaluation:

Hyperparameters	**Chosen Value**	**swept value**
EBM Variant	Langevin	
Train iterations	12000	
Batch size	512	460,512,560
Learning rate	0.001	0.0009,0.001,0.0011
Network width	256	128,256,512
Network depth	6	4,6,8
Counter examples	8	7,8,9

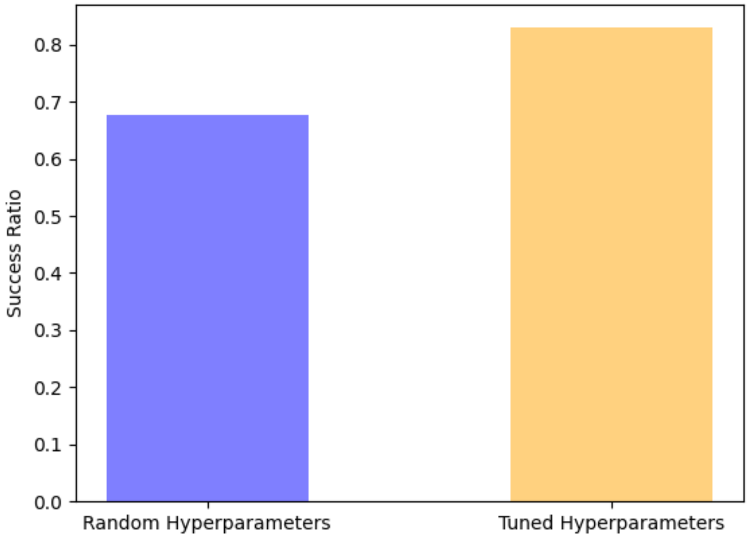

Figure 3.6: Success ratio

The results of this experiment clearly indicate that the policy's performance experiences a level of deterioration in the presence of hyperparameter perturbations. However, it's noteworthy that the policy remains serviceable even under these conditions. This experiment underscores the feasibility of selecting adequately performing hyperparameters without necessitating a perfectly fine-tuned configuration, emphasizing the practicality of achieving satisfactory policy outcomes through hyperparameter tuning that is sufficiently close to the ideal.

Chapter 4

DISCUSSION

In this paper, we have demonstrated the effectiveness reformulating a supervised imitation learning as a conditional energy-based modeling problem, with inference-time implicit regression, often greatly outperforms traditional explicit policy baselines in a contact rich bimanual robot manipulation, even when confronted with tasks characterized by high-dimensional action spaces. It is important to note that, in terms of computational requirements, explicit models typically demand more computing resources both during the training phase and for inference. However, our work also includes evidence showcasing the feasibility of deploying implicit policies for a Bimanual robot manipulation, and training time is modest compared to offline RL algorithms. To further motivate the use of implicit models, we presented an intuitive analysis of energy-based model characteristics, highlighting a number of potential benefits.

4.1 Future Studies

The present study represents a focused examination of a singular imitation learning algorithm, complemented by a restricted set of Energy-Based Models (EBMs) for analysis. While this investigation provides a valuable foundation, there exists a promising avenue for future research expansion. Subsequent studies could be broadened to encompass a comprehensive comparative analysis involving a spectrum of imitation learning algorithms, thereby elucidating the nuances of their respective strengths and weaknesses. Furthermore, the scope of analysis can be enriched by incorporating a more extensive array of EBM models, which would allow for a deeper exploration of their influence on policy learning and, in turn, contribute to a more holistic understanding of the intricate dynamics underlying imitation learning in the field of robotics and artificial intelligence. Such future endeavors hold the potential to unveil novel insights and refine our approaches to policy optimization and learning.

Chapter 5

CONCLUSION

The core objective of this research project is to provide an in-depth examination and utilization of Implicit Behavioral Cloning techniques within the context of a bimanual robot that frequently engages with intricate, contact-rich environments. The primary focus of this study revolves around a comprehensive comparative analysis of various training methods. Furthermore, it explores the nuanced dynamics of expert noise and the sensitivity of hyperparameters, investigating their collective influence on the policy formation process. By scrutinizing the intricate interplay between these factors, this research seeks to offer a comprehensive understanding of how training methodologies, expert noise, and hyperparameter adjustments collectively mold the policies governing the behavior of the bimanual robot. This multifaceted exploration endeavors to shed light on the underlying complexities of policy development in the realm of advanced robotics and its interaction with challenging and dynamic surroundings.

The findings of this study provide valuable insights into the trade-offs associated with different optimization techniques. Notably, Derivative-Free Optimization (DFO) emerges as a computationally efficient approach, offering swiftness in lower-dimensional spaces. However, it becomes apparent that its performance encounters substantial challenges when dealing with higher-dimensional environments. On the other hand, both the autoregressive and Langevin versions of optimization techniques prove to be more robust in terms of solving complex environments. These methods, while exhibiting reliable performance, do come with the trade-off of higher computational demands. The Autoregressive DFO, in particular, stands out for its memory-intensive nature, necessitating N separate models for N dimensions. In contrast, the

Langevin version showcases scalability in high-dimensional spaces, requiring only a single model, which, while incurring higher compute times, offers a compelling advantage in terms of ease of implementation and resource efficiency. This analysis underscores the intricate balance between computational efficiency, scalability, and memory requirements in optimizing solutions for high-dimensional problems.

Additionally, our research has shed light on the critical role of high-quality expert demonstrations and hyperparameter sensitivity in Implicit Behavioral Cloning (IBC) and its pronounced dependence on them. It becomes evident that IBC thrives when furnished with a substantial volume of expert demonstrations, a factor that significantly impacts its performance. Our study has delved into the intricacies of expert demonstration noise, and implications. We showed that IBC is highy susceptible to action noises than the observation noises. This understanding carries noteworthy implications, particularly in the context of training real-world robots with real-world expert demonstrations, which are often characterized by inherent noise and uncertainties. The insights gleaned from our exploration offer valuable guidance in navigating the challenges posed by noisy expert demonstrations, providing a path toward more robust and effective policy learning for practical, real-world applications of robotics.

REFERENCES

Bersch, C., B. Pitzer and S. Kammel, “Bimanual robotic cloth manipulation for laundry folding”, in “2011 IEEE/RSJ International Conference on Intelligent Robots and Systems”, pp. 1413–1419 (2011).

Conn, A. R., K. Scheinberg and L. N. Vicente, *Introduction to derivative-free optimization* (SIAM, 2009).

Du, Y., S. Li, J. Tenenbaum and I. Mordatch, “Improved contrastive divergence training of energy based models”, arXiv preprint arXiv:2012.01316 (2020).

Du, Y. and I. Mordatch, “Implicit generation and modeling with energy based models”, Advances in Neural Information Processing Systems **32** (2019).

Florence, P., C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch and J. Tompson, “Implicit behavioral cloning”, in “Conference on Robot Learning”, pp. 158–168 (PMLR, 2022).

Hempel, T., A. A. Abdelrahman and A. Al-Hamadi, “6d rotation representation for unconstrained head pose estimation”, in “2022 IEEE International Conference on Image Processing (ICIP)”, pp. 2496–2500 (2022).

Loh, W.-L., “On latin hypercube sampling”, The annals of statistics **24**, 5, 2058–2080 (1996).

Mirrazavi Salehian, S. S., N. Figueroa and A. Billard, “A unified framework for coordinated multi-arm motion planning”, The International Journal of Robotics Research **37**, 10, 1205–1232 (2018).

Nash, C. and C. Durkan, “Autoregressive energy machines”, in “International Conference on Machine Learning”, pp. 1735–1744 (PMLR, 2019).

Peters, J. and S. Schaal, “Learning to control in operational space”, The International Journal of Robotics Research **27**, 2, 197–212 (2008).

Ross, S., G. Gordon and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning”, in “Proceedings of the fourteenth international conference on artificial intelligence and statistics”, pp. 627–635 (JMLR Workshop and Conference Proceedings, 2011).

Stepputtis, S., M. Bandari, S. Schaal and H. B. Amor, “A system for imitation learning of contact-rich bimanual manipulation policies”, in “2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)”, pp. 11810–11817 (IEEE, 2022).

Tu, S., A. Robey, T. Zhang and N. Matni, “On the sample complexity of stability constrained imitation learning”, in “Learning for Dynamics and Control Conference”, pp. 180–191 (PMLR, 2022).