

Demonstration-Efficient Guided Policy Search via Imitation of Robust Tube MPC

Andrea Tagliabue, Dong-Ki Kim, Michael Everett, Jonathan P. How

Abstract—We propose a demonstration-efficient strategy to compress a computationally expensive Model Predictive Controller (MPC) into a more computationally efficient representation based on a deep neural network and Imitation Learning (IL). By generating a Robust Tube variant (RTMPC) of the MPC and leveraging properties from the tube, we introduce a data augmentation method that enables high demonstration-efficiency, being capable to compensate the distribution shifts typically encountered in IL. Our approach opens the possibility of zero-shot transfer from a single demonstration collected in a nominal domain, such as a simulation or a robot in a lab/controlled environment, to a domain with bounded model errors/perturbations. Numerical and experimental evaluations performed on a trajectory tracking MPC for a quadrotor show that our method outperforms strategies commonly employed in IL, such as DAgger and Domain Randomization, in terms of demonstration-efficiency and robustness to perturbations unseen during training.

I. INTRODUCTION

Model Predictive Control (MPC) [1] enables impressive performance on complex, agile robots [2]–[7]. However, its computational cost often limits the opportunities for onboard, real-time deployment [8], or takes away critical computational resources needed by other components governing the autonomous system. Recent works have mitigated MPC’s computational requirements by relying on computationally efficient deep neural networks (DNNs), which are leveraged to imitate task-relevant demonstrations generated by MPC. Such demonstrations are generally collected via Guided Policy Search (GPS) [9]–[12] and Imitation Learning (IL) [13]–[15] and are then used to train a DNN via supervised learning.

A common issue in existing IL methods (e.g., Behavior Cloning (BC) [16]–[18], Dataset-Aggregation (DAgger) [19]) is that they require collecting a relatively large number of MPC demonstrations, preventing training directly on a real robot and requiring a simulation environment that accurately represents the deployment domain. One of the causes for such demonstration inefficiency is the need to take into account and correct for the compounding of errors in the learned policy [19], which may otherwise create shifts (*covariate shifts*) from the training distribution, with catastrophic consequences [16]. These errors can be caused by: a) approximation errors in the learned policy; b) mismatches in the simulation dynamics due to modelling errors (i.e., sim2real gap); or c) model changes or disturbances that may not be present in a controlled training environment (e.g., simulation, or lab/factory when training on a real

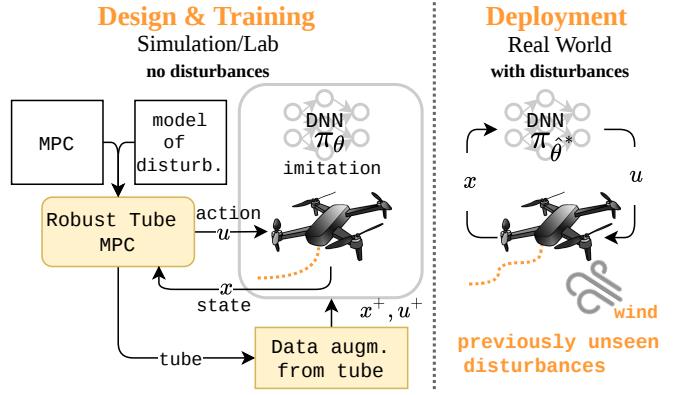


Fig. 1: Overview of the proposed approach. To compress a computationally expensive MPC controller in a DNN-based policy π_θ in a demonstration efficient way, we generate a Robust Tube MPC controller using a model of the disturbances encountered in the deployment domain. We use properties of the tube to derive a data augmentation strategy which generates extra state-action pairs (x^+, u^+) , obtaining $\pi_{\hat{\theta}^*}$ via Imitation Learning. Our approach enables zero-shot transfer from a single demonstration collected in simulation (*sim2real*) or a controlled environment (lab, factory, *lab2real*).

robot), but do appear during deployment in the real-world (i.e., lab2real gap). Approaches employed to compensate for such *gaps*, such as Domain Randomization (DR) [20], [21], introduce further challenges, for example, requiring to apply disturbances/model changes during training.

In this work, we address the problem of generating a compressed MPC policy in a demonstration-efficient manner by providing a data augmentation strategy that systematically compensates for the effects of covariate shifts that might be encountered during real-world deployment. Specifically, our approach relies on a model of the perturbations/uncertainties encountered in a deployment domain, which is used to generate a Robust Tube version of the given MPC (RTMPC) to guide the collection of demonstrations and the data augmentation. An overview of the strategy is provided in Figure 1. Our approach is related to the recent LAG-ROS framework [22], which provides a learning-based method to compress a planner in a DNN by extracting relevant information from the robust tube. LAG-ROS emphasizes the importance of nonlinear contraction-based controllers (e.g., CV-STEM [23]) to obtain robustness and stability guarantees. In a complementary way, our work emphasizes minimal requirements - namely a tube and a data augmentation strategy - to achieve demonstration efficiency and robustness to real-world conditions. By decoupling these aspects from the need for complex nonlinear models and control strategies, we greatly simplify the controller design and reduce the computational complexity, which enables *lab2real* transfers.

Contributions. Via numerical comparison with previous IL methods and experimental validations, we show that demon-

stration efficiency can be achieved in MPC-compression by generating a corresponding Robust Tube MPC (RTMPC) and using the tube to guide the data augmentation strategy for IL methods (DAgger, BC). To this end, we propose a data-sparse, computationally efficient (i.e., scales linearly in state size) adversarial Sampling Augmentation (SA) strategy for data augmentation. We highlight that the proposed approach, for example, can be used to train the robot in a low-fidelity simulation environment while achieving robustness to real-world perturbations unseen during the training phase. We validate the proposed approach by providing the first experimental (hardware) demonstration of zero-shot transfer of a DNN-based trajectory tracking controller for an aerial robot, learned from a *single* demonstration, in an environment (low-fidelity simulation) without disturbances, and transferred to an environment with wind-like disturbances.

II. RELATED WORK

MPC-like policy compression for mobile robots via IL and GPS. IL methods have found application in multiple robotics tasks. The works in [14], [24] use DAgger [19] to control aerial and ground robots, while [13] uses a combination of DAgger and DR to learn to perform acrobatic maneuvers with a quadrotor. Similarly, GPS methods have been demonstrated in simulation for navigation of a multirotor [10], [11], and to control a legged robot [12], [15]. These methods achieve impressive performance, but at the cost of requiring multiple demonstrations to execute a single trajectory, and do not explicitly take into account the effects of disturbances encountered in the deployment domain.

Sample efficiency and robustness in IL. Robustness in IL, required to deal with distribution shifts caused by the sim2real (model errors) or lab2real transfer (model changes, external disturbances), has been achieved a) by modifying the training domain so that its dynamics match the deployment domain, as done in DR [20], [21], [25], [26], or b) by modifying the actions of the expert, so that the state distribution during training matches the one encountered at deployment, as proposed in [27]–[30]. Although effective, these approaches do not leverage extra information available in the RTMPC, thus requiring a larger number of demonstrations.

Data augmentation in IL. Data augmentation is a commonly employed robustification strategy in IL. Most approaches focus on reducing overfitting in the high-dimensional policy input space (e.g., images), by applying noise [31], transformations [32] or adversarially-generated perturbations [33], [34], while maintaining the corresponding action label unchanged. Data augmentation is also employed to reduce covariate shift in self-driving by generating transformed observations [18], [35], [36] with the corresponding action label computed via a feedback controller. These approaches do not directly apply to our context, as they do not rely on RTMPC and we assume available state estimate. Aligned to our findings, [9], [12] observe that adding extra samples from the tube of an existing Iterative-Linear Quadratic Regulator (LQR) can achieve increased demonstration efficiency in GPS. Compared to these, thanks to RTMPC, we can additionally

consider the effects of disturbances encountered in the sim2real or lab2real transfer, providing additional robustness.

III. METHOD

This section explains the given MPC expert and its Robust-Tube variant, which we leverage to design a data augmentation strategy. We additionally cast the demonstration-efficiency challenge in IL as a robust IL problem in the context of transferring a policy between two different domains, and we present the SA strategy to improve demonstration efficiency and robustness of MPC-guided policies learned via IL.

A. MPC and Robust Tube MPC demonstrator

Model predictive trajectory tracking controller. We assume a trajectory tracking linear MPC [1] is given that controls a system subject to bounded uncertainty. The linearized, discrete-time model of the system is:

$$\ddot{x}_{t+1} = A\ddot{x}_t + B\ddot{u}_t + w_t \quad (1)$$

where $x \in \mathbb{X} \subseteq \mathbb{R}^{n_x}$ represents the state (size n_x), $u \in \mathbb{U} \subseteq \mathbb{R}^{n_u}$ are the commanded actions (size n_u), and $w \in \mathbb{W} \subseteq \mathbb{R}^{n_x}$ is an additive perturbation/uncertainty. $A \in \mathbb{R}^{n_x \times n_x}$ and $B \in \mathbb{R}^{n_x \times n_u}$ represent the nominal dynamics; $\ddot{\cdot}$ denotes internal variables of the optimization. At every discrete timestep t is given an estimate of the state of the actual system x_t and a reference trajectory $x_{0,t}^{\text{ref}}, \dots, x_{N,t}^{\text{ref}}$. The controller computes a sequence of N actions $\ddot{u}_0(x_t), \dots, \ddot{u}_{N-1}(x_t)$, subject to state and input constraints \mathbb{X}, \mathbb{U} , and executes the first computed optimal action $u_t = \ddot{u}_0^{\text{MPC},*}(x_t)$ from the sequence. The optimal sequence of actions is computed by minimizing the value function (dropping x_t in the notation):

$$V_t(\ddot{u}_0, \dots, \ddot{u}_{N-1}, \ddot{x}_0) = \sum_{n=0}^{N-1} (e_n^T Q e_n + \ddot{u}_n^T R \ddot{u}_n) + e_N^T P e_N, \quad (2)$$

where $e_n = (\ddot{x}_n - x_{n,t}^{\text{ref}})$. Matrices Q (size $n_x \times n_x$) and R (size $n_u \times n_u$) are user-selected, positive definite weights that define the stage cost, while P (size $n_x \times n_x$, positive definite) represents the terminal cost. The prediction horizon is an integer $N \geq 1$. The system is additionally subject to constraints $\ddot{x}_0 = x_t$; the predicted states $\ddot{x}_1, \dots, \ddot{x}_N$ are obtained from the model in (1) assuming the disturbance $w = 0$. The optimization problem is solved again at every timestep, executing the newly recomputed optimal action.

Robust Tube MPC. Given the MPC expert, we generate a Robust Tube variant using [37]. At every discrete timestep t , the RTMPC operates in a similar way as MPC, but it additionally accounts for the effects of w by introducing a feedback policy, called *ancillary controller*

$$u_t = \ddot{u}_0^*(x_t) + K(x_t - \ddot{x}_0^*(x_t)), \quad (3)$$

where u_t represents the executed action. The quantities $\ddot{u}_0^*(x_t)$ represents an optimal, feedforward action, and $\ddot{x}_0^*(x_t)$ is an optimal reference, and are computed by the RTMPC given the current state estimate x_t . The ancillary controller ensures that the state x of the controlled system remains inside a set (*tube*) $\mathbb{Z} \in \mathbb{R}^{n_x}$, centered around $\ddot{x}_0^*(x_t)$, for every possible

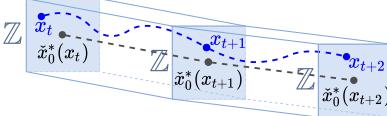


Fig. 2: Illustration of the robust control invariant tube \mathbb{Z} centered around the optimal reference $\check{x}_0^*(x_t)$ computed by RTMPC for each state of the system x_t .

realization of $w \in \mathbb{W}$. The quantities $\check{u}_0^*(x_t)$ and $\check{x}_0^*(x_t)$ are obtained by solving (dropping the dependence on x_t)

$$\check{x}_0^*, \check{u}_0^*, \dots = \arg \min_{\check{x}_0, \check{u}_0, \dots, \check{u}_{N-1}} V_t(\check{u}_0, \dots, \check{u}_{N-1}, \check{x}_0) \quad (4)$$

under the constraint that $x_t \in \check{x}_0 + \mathbb{Z}$. As in the original MPC formulation, the optimization problem is additionally subject to the given input and actuation constraints, tightened by an amount that takes into account the effects of the disturbances. The gain matrix $K \in \mathbb{R}^{n_u \times n_x}$ in (3) is computed such that $A_K := A + BK$ is stable, for example by solving the infinite-horizon, discrete-time LQR problem using (A, B, Q, R) . The set \mathbb{Z} has constant size, and it determines the shape/width of the tube. It is defined as disturbance invariant set for the closed-loop system A_K , and satisfies the property that $\forall x_t \in \mathbb{Z}, \forall w_t \in \mathbb{W}, \forall t \in \mathbb{N}^+, x_{t+1} = A_K x_t + w_t \in \mathbb{Z}$. In practice, \mathbb{Z} can be computed offline using A_K and the model of the disturbance \mathbb{W} via ad-hoc algorithms [1], [37], or can be learned from data [38]. The set \mathbb{Z} and the ancillary controller in (3) ensure (see [37]) that, given a state $x_t \in \check{x}_0^*(x_t) + \mathbb{Z}$, the perturbed system in (1) will remain in the tube centered around the trajectory of \check{x}_0^* , no matter the disturbance realization $w \in \mathbb{W}$, as shown in Figure 2. This additionally implies that the tube represents a model of the states that the system may visit when subject to the disturbances in \mathbb{W} . The ancillary controller provides a computationally-effective way to generate a control action to counteract such perturbations.

B. Covariate shift in sim2real and lab2real transfer

This part describes the demonstration-efficiency issue in IL as the ability to efficiently predict and compensate for the effects of covariate shifts during real-world deployment. We assume that the causes of such distribution shifts can be modeled as additive state perturbations/uncertainties $w \in \mathbb{W}$ encountered in the deployment domains.

Policies and state densities. We model the dynamics of the real system as Markovian and stochastic [39]. The stochasticity with respect to state transitions is introduced by unknown perturbations, assumed to be additive (as in (1)) and belonging to the bounded set \mathbb{W} , sampled under a (possibly unknown) probability distribution. These perturbations capture the effects of noise, approximation errors in the learned policy, model changes and other disturbances acting on the system. Two different domains \mathcal{D} are considered: a training domain \mathcal{S} (*source*) and a deployment domain \mathcal{T} (*target*). The two domains differ in their transition probabilities, effectively representing the sim2real or lab2real settings. We additionally assume that the considered system is controlled by a deterministic policy $\pi_\theta : \mathbb{X} \times \mathbb{X}_{0,\dots,N}^{\text{ref}} \rightarrow \mathbb{U}$, where $\mathbb{X}_{0,\dots,N}^{\text{ref}}$ represents the reference trajectory. Given $(x_{0,t}^{\text{ref}}, \dots, x_{N,t}^{\text{ref}}) \in \mathbb{X}_{0,\dots,N}^{\text{ref}}$, the resulting transition probability is $p(x_{t+1}|x_t, u_t =$

$\pi_\theta(x_t, x_{0,t}^{\text{ref}}, \dots, x_{N,t}^{\text{ref}}, \mathcal{D})$, denoted $p_{\pi_\theta, \mathcal{D}}(x_{t+1}|x_t)$ to simplify the notation. The probability of collecting a T -step trajectory $\xi = (x_0, u_0, x_1, u_1, \dots, x_{T-1})$ given a generic policy π_θ in \mathcal{D} is $p(\xi|\pi_\theta, \mathcal{D}) = p(x_0) \prod_{t=0}^{T-1} p_{\pi_\theta, \mathcal{D}}(x_{t+1}|x_t)$, where $p(x_0)$ represents the initial state distribution.

Robust IL objective. Following [27], given an expert RTMPC policy π_{θ^*} , the objective of IL is to find parameters θ of π_θ that minimize a distance metric $J(\theta, \theta^*|\xi)$. This metric captures the differences between the actions generated by the expert π_{θ^*} and the action produced by the learner π_θ across the distribution of trajectories induced by the learned policy π_θ , in the perturbed target domain \mathcal{T} :

$$\hat{\theta}^* = \arg \min_{\theta} \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{T})} J(\theta, \theta^*|\xi) \quad (5)$$

A choice of distance metric that we consider in this paper is the MSE loss: $J(\theta, \theta^*|\xi) = \frac{1}{T} \sum_{t=0}^{T-1} \|\pi_\theta(x_t) - \pi_{\theta^*}(x_t)\|_2^2$.

Covariate shift due to sim2real and lab2real transfer. Since in practice we do not have access to the target environment, our goal is to try to solve (5) by finding an approximation of the optimal policy parameters $\hat{\theta}^*$ in the source environment:

$$\hat{\theta}^* = \arg \min_{\theta} \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{S})} J(\theta, \theta^*|\xi). \quad (6)$$

The way this minimization is solved depends on the chosen IL algorithm. The performance of the learned policy in target and source domain can be related via:

$$\begin{aligned} \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{T})} J(\theta, \theta^*|\xi) &= \\ &\underbrace{\mathbb{E}_{p(\xi|\pi_\theta, \mathcal{T})} J(\theta, \theta^*|\xi) - \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{S})} J(\theta, \theta^*|\xi)}_{\text{covariate shift due to transfer}} \\ &+ \underbrace{\mathbb{E}_{p(\xi|\pi_\theta, \mathcal{S})} J(\theta, \theta^*|\xi)}_{\text{IL objective}} \end{aligned} \quad (7)$$

which clearly shows the presence of a covariate shift induced by the transfer. The last term corresponds to the objective minimized by performing IL in \mathcal{S} . Attempting to solve (5) by directly optimizing (6) (e.g., via BC [16]) provides no guarantees of finding a policy with good performance in \mathcal{T} .

Compensating transfer covariate shift via Domain Randomization. A well known strategy to compensate for the effects of covariate shifts between source and target domain is Domain Randomization (DR) [20], which modifies the transition probabilities of the source \mathcal{S} by trying to ensure that the trajectory distribution in the modified training domain \mathcal{S}_{DR} matches the one encountered in the target domain: $p(\xi|\pi_\theta, \mathcal{S}_{\text{DR}}) \approx p(\xi|\pi_\theta, \mathcal{T})$. This is in practice done by sampling perturbations $w \in \mathbb{W}$ according to some knowledge/hypotheses on their distribution $p_{\mathcal{T}}(w)$ in the target domain [20], obtaining the perturbed trajectory distribution $p(\xi|\pi_\theta, \mathcal{S}, w)$. The minimization of (5) can then be approximately performed by minimizing instead:

$$\mathbb{E}_{p_{\mathcal{T}}(w)} [\mathbb{E}_{p(\xi|\pi_\theta, \mathcal{S}, w)} J(\theta, \theta^*|\xi)]. \quad (8)$$

This approach, however, requires the ability to apply disturbances/model changes to the system, which may be unpractical e.g., in the *lab2real* setting, and may require a large number of demonstrations due to the need to sample enough state perturbations w .

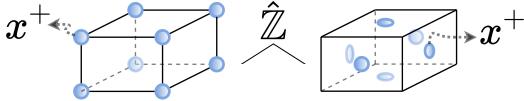


Fig. 3: The two alternative strategies evaluated to sample extra state/actions pairs from an approximation of the tube of the RTMPC expert: dense (left) and sparse (right).

C. Covariate shift compensation via Sampling Augmentation

We propose to mitigate the covariate shift introduced by the compression procedure not only by collecting demonstrations from the RTMPC, but by using additional information computed in the controller. Unlike DR, the proposed approach does not require to explicitly apply disturbances in the training phase. During the collection of a trajectory ξ in the source domain \mathcal{S} , we utilize instead the tube computed by the RTMPC demonstrator to obtain knowledge of the states that the system may visit when subjected to perturbations. Given this information, we propose a state sampling strategy, called Sampling Augmentation (SA), to extract relevant states from the tube. The corresponding actions are provided at low computational cost by the demonstrator. The collected state-actions pairs are then included in the set of demonstrations used to learn a policy via IL. The following paragraphs frame the tube sampling problem in the context of covariate shift reduction in IL, and present two tube sampling strategies.

RTMPC tube as model of state distribution under perturbations. The key intuition of the proposed approach is the following. We observe that, although the density $p(\xi|\pi_\theta, \mathcal{T})$ is unknown, an approximation of its support \mathfrak{R} , given a demonstration ξ collected in the source domain \mathcal{S} , is known. Such support corresponds to the tube computed by the RTMPC when collecting ξ :

$$\mathfrak{R}_{\xi^+|\pi_{\theta^*}, \xi} = \bigcup_{t=0}^{T-1} \{\check{x}_0^*(x_t) + \mathbb{Z}\}. \quad (9)$$

where ξ^+ is a trajectory in the tube of ξ . This is true thanks to the ancillary controller in (3), which guarantees that the system remains inside (9) for every possible realization of $w \in \mathbb{W}$. The ancillary controller additionally provides a way to easily compute the actions to apply for every state inside the tube. Let x^+ be a state inside the tube computed when the system is at x_t (formally $x^+ \in \check{x}_0^*(x_t) + \mathbb{Z}$), then the corresponding robust control action u^+ is simply:

$$u^+ = \check{u}_0^*(x_t) + K(x^+ - \check{x}_0^*(x_t)). \quad (10)$$

For every timestep in ξ , extra state-action samples (x^+, u^+) collected from within the tube can be used to augment the dataset employed for empirical risk minimization, obtaining a way to approximate the expected risk in the domain \mathcal{T} by only having access to demonstrations collected in \mathcal{S} :

$$\begin{aligned} & \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{T})} J(\theta, \theta^*|\xi) \approx \\ & \mathbb{E}_{p(\xi|\pi_\theta, \mathcal{S})} [J(\theta, \theta^*|\xi) + \mathbb{E}_{p(\xi^+|\pi_{\theta^*}, \xi)} J(\theta, \theta^*|\xi^+)]. \end{aligned} \quad (11)$$

The demonstrations in the source domain \mathcal{S} can be collected using existing techniques, such as BC and DAgger.

Tube approximation and sampling strategies. In practice, the set \mathbb{Z} may have arbitrary shape (not necessarily politopic), and the density $p(\xi^+|\xi, \pi_{\theta^*})$ may not be available,

making difficult to establish where/which states to sample in order to derive a data augmentation strategy. We proceed by approximating \mathbb{Z} as an hyper-rectangle $\hat{\mathbb{Z}}$, outer approximation of the tube. We consider an adversarial approach to the problem by sampling from the states visited under worst-case perturbations. We investigate two strategies, shown in Figure 3, to obtain state samples x^+ at every state x_t in ξ : i) **dense sampling**: sample extra states from the vertices of $\check{x}_0^*(x_t) + \hat{\mathbb{Z}}$. The approach produces 2^{n_x} extra state-action samples. It is more conservative, as it produces more samples, but more computationally expensive. ii) **sparse sampling**: sample one extra state from the center of each facet of $\check{x}_0^*(x_t) + \hat{\mathbb{Z}}$, producing $2n_x$ additional state-action pairs. It is less conservative and more computationally efficient.

IV. RESULTS

A. Evaluation approach

MPC for trajectory tracking on a multirotor. We evaluate the proposed approach in the context of trajectory tracking control for a multirotor, using the controller proposed in [5], modified to obtain a RTMPC. We model \mathbb{W} under the assumption that the system is subject to force-like perturbations up to 30% of the weight of the robot (approximately the safe physical limit of the robot). The tube \mathbb{Z} is approximated via Monte-Carlo sampling of the disturbances in \mathbb{W} , evaluating the state deviations of the closed loop system A_K . The derived controller generates tilt (roll, pitch) and thrust commands ($n_u = 3$) given the state of the robot ($n_x = 8$ consisting of position, velocity and tilt) and the reference trajectory. The reference is a sequence of desired positions and velocities for the next 3s, discretized with sampling time of 0.1s (corresponding to a planning horizon of $N = 30$, and 180-dim. vector). The controller takes into account position constraints (e.g., available 3D flight space), actuation limits, and velocity/tilt limits.

Policy architecture. The compressed policy is a 2-hidden layers, fully connected DNN, with 32 neurons per layer, and ReLU activation function. The total input dimension of the DNN is 188 (position, velocity, current tilt expressed in an inertial frame, and the desired reference trajectory). The output dimension is 3 (desired thrust and tilt expressed in an inertial frame). We rotate the tilt output of the DNN in body frame to avoid taking into account yaw, which is not part of the optimization problem [5], not causing any relevant computational cost. We additionally apply the non-linear attitude compensation scheme as in [5].

Training environment and training details. Training is performed in a custom-built non-linear quadrotor simulation environment, where the robot follows desired trajectories, starting from randomly generated initial states, centered around the origin. Demonstrations are collected with a sampling time of 0.1s and training is performed for 50 epochs via the ADAM [40] optimizer, with a learning rate of 0.001.

Evaluation details and metrics. We apply the proposed SA strategies to DAgger and BC, and compare their performance against the two without SA, and the two combined with DR. Target and source domain differs due to

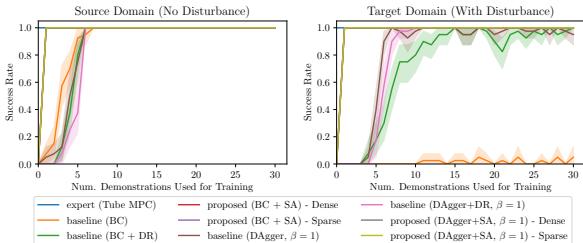


Fig. 4: Robustness (*Success Rate*) in the task of flying along an eight-shaped, 7s long trajectory, subject to wind-like disturbances (left) and without (right), starting from different initial states (Task T1). Evaluation is repeated across 8 random seeds, 20 times per demonstration per seed. We additionally show the 95% confidence interval.

Method	Training		Robustness		Performance		Demonstration Efficiency	
	Easy	Safe	succ. rate (%)	T1 T2	expert gap (%)	T1 T2	T1	T2
- BC	Yes	Yes	< 1	100	24.15	29.47	-	6
	Yes	No	98	100	15.79	1.34	7	6
DR	No	Yes	95	100	10.04	1.27	14	9
	No	No	100	100	4.09	1.45	10	6
SA-Dense	BC	Yes	100	100	25.64	1.34	1	1
	DAgger	Yes*	100	100	10.21	1.66	1	1
SA-Sparse	BC	Yes	100	100	4.23	1.13	1	1
	DAgger	Yes*	100	100	3.75	1.07	1	1

TABLE I: Comparison of the IL methods considered for RT-MPC compression. T1 refers to the trajectory tracking task under wind-like disturbances, T2 under model errors (drag coefficient mismatch). At convergence (iteration 20-30) we evaluate, in the target domain, robustness (success rate) and performance (relative percent error between actions of the expert and of the compressed policy). Demonstration-Efficiency represents the number of demonstrations required to achieve for the first time full success rate. An approach is considered easy if it does not require to apply disturbances/perturbations during training (e.g., in *lab2real* transfer); an approach is considered safe if does not execute actions that may cause state constraints violations (crashes) during training. *Safe in our numerical evaluation, but not guaranteed.

perturbations sampled from \mathbb{W} in target. During training with DR we sample disturbances from the entire \mathbb{W} . In all the comparisons, we set the probability of using actions of the expert β , hyperparameter of DAgger [19], to be 1 at the first demonstration, and 0 otherwise (as this was found to be the best performing setup). We monitor: i) *robustness (Success Rate)*, as the percentage of episodes where the robot never violates any state constraint; ii) *performance (MPC Stage Cost)*, as $\sum_t e_t^T Q e_t + u_t^T R u_t$ along the trajectory.

B. Numerical evaluation of demonstration-efficiency, robustness and performance for tracking a single trajectory

Tasks description. Our objective is to compress an RTMPC policy capable to track a 7s long, eight-shaped trajectory. We evaluate the considered approaches in two different target domains, with wind-like disturbances (T1) or model errors (T2). Disturbances in T1 are sampled adversarially from \mathbb{W} ($\approx 25\text{--}30\%$ of the UAV weight), while model errors in T2 are applied via mismatches in the drag coefficients used between training and testing.

Results. We start by evaluating the robustness in T1 as a function of the number of demonstrations collected in the source domain. The results are shown in Fig 4, highlighting that: i) while all the approaches achieve robustness (full success rate) in the source domain, SA achieves full success rate after only a single demonstration, being 5-6 times more sample efficient than the baseline methods; ii) SA is able to achieve full robustness in the target domain, while baseline methods do not fully succeed, and converge at much lower rate. These results remark the presence of a distribution shift between the source and target, which is not fully compensated

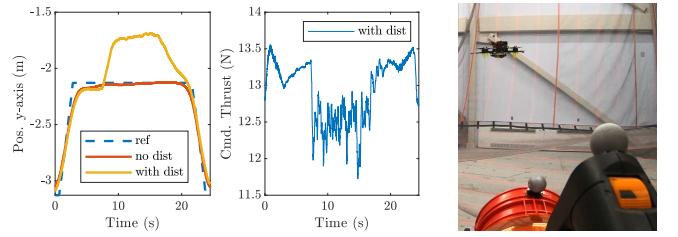


Fig. 5: Experimental evaluation performed by hovering with and without wind disturbances produced by a leaf blower. The employed compressed RTMPC policy is trained in simulation from a *single* demonstration of the desired trajectory. The wind-like disturbances produce a large position error, but do not destabilize the system. The thrust decreases due to the robot being pushed up by the disturbances. The state estimate (shown in the plot) is provided by onboard Visual-Inertial Odometry (VIO). The y-axis points approximately in the same direction as the wind.

by baselines methods such as BC, due to lack of exploration and robustness. The performance evaluation and additional results are summarized in Table I. We highlight that in the target domain, sparse SA combined with DAgger achieves closest performance to the expert. Dense SA suffers from performance drops, potentially due to the limited capacity of the considered DNN or challenges in training introduced by this data augmentation. Because of its effectiveness and greater computational efficiency, we use sparse SA for the rest of the work. Table I additionally presents the results for task T2. Although this task is less challenging (i.e., all the approaches achieve full robustness), the proposed method (sparse SA) achieves highest demonstration efficiency and lowest expert gap, with similar trends as T1.

Computation. The average latency (on i7-8750H laptop with NVIDIA GTX1060 GPU) for the expert (MATLAB) is 55.3ms, while for the compressed policy (PyTorch) is 0.25ms, achieving a two-orders of magnitude improvement. The average latency for the compressed policy on an Nvidia TX2 CPU (PyTorch) is 1.66ms.

C. Hardware evaluation for tracking a single trajectory

We validate the demonstration efficiency, robustness and performance of the proposed approach by experimentally testing policies trained after a *single* demonstration collected in simulation using DAgger/BC (which operate identically since we use DAgger with $\beta = 1$ for the first demonstration). The data augmentation strategy is based on the sparse SA. We use the MIT/ACL open-source snap-stack [41] for controlling the attitude of the MAV, while the compressed RTMPC runs at 100Hz on the onboard Nvidia TX2 (on its CPU), with the reference trajectory provided at 100Hz. State estimation is obtained via a motion capture system or onboard VIO. The first task considered is to hover under wind disturbances produced by a leaf blower. The results are shown in Figure 5, and highlight the ability of the system to remain stable despite the large position error caused by the wind. The second task is to track an eight-shaped trajectory, with velocities up to 3m/s. We evaluate the robustness of the system by applying a wind-like disturbance produced by an array of 3 leaf blowers (Figure 6). The given position reference and the corresponding trajectory are shown in Figure 6a. The effects of the wind disturbances are clearly visible in the altitude errors and changes in commanded thrust in Figure 6a (at $t = 11$ s and

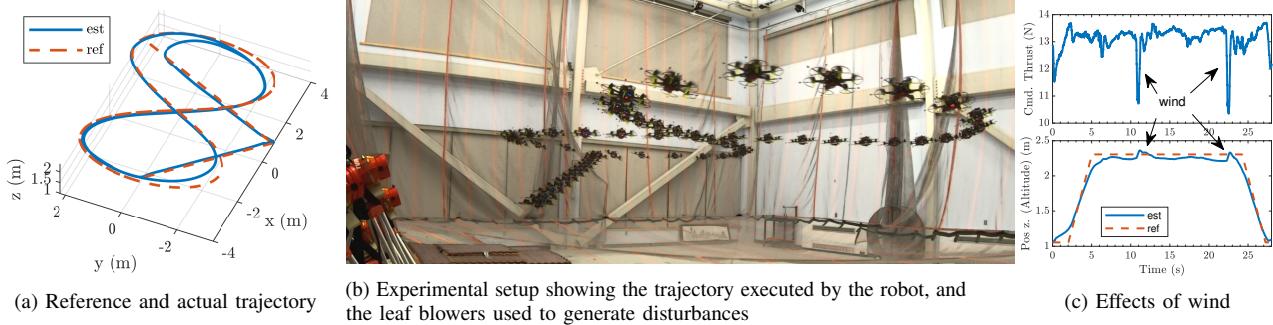


Fig. 6: Experimental evaluation of the proposed approach (Figure 6b) under wind-like disturbances. The robot learns to track an eight-shaped reference trajectory (Figure 6a) from a *single* demonstration collected in simulation, achieving zero-shot transfer. It is additionally able to withstand disturbances produced by an array of leaf-blowers, unseen during the training phase, and whose effects are clearly visible in the altitude errors (and change in commanded thrust) in Figure 6c.

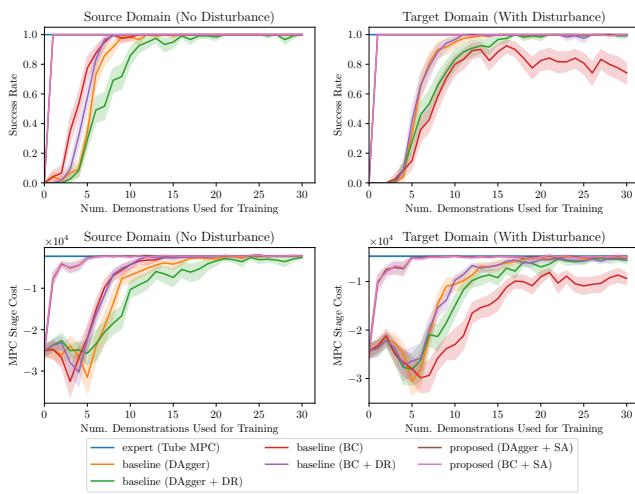


Fig. 7: Robustness (*Success Rate*, top row) and performance (*MPC Stage Cost*, bottom row) of the proposed approach (with 95% confidence interval), as a function of the number of demonstrations used for training, for the task of learning to track previously unseen circular, eight-shaped and constant position reference trajectories, sampled from the same training distribution. The left column presents the results in the training domain (no disturbance), while the right column under wind-like perturbations (with disturbance). The proposed RTMPC driven sparse SA strategy learns to track multiple trajectory and generalize to unseen ones requiring less demonstrations. At convergence (from demonstration 20 to 30), DAgger+SA achieves the closest performance to the expert (2.7% *expert gap*), followed by BC+SA (3.0% *expert gap*). Evaluation performed using 20 randomly sampled trajectories per demonstration, repeated across 6 random seeds, with prediction horizon of $N = 20$ to speed-up demonstration collection, and the DNN input size is adjusted accordingly.

$t = 23\text{s}$). These experiments show that the controller can robustly track the desired reference, withstanding challenging perturbations unseen during the training phase. The video submission presents more experiments, including lab2real transfer.

D. Numerical and hardware evaluation for learning and generalizing to multiple trajectories

We evaluate the ability of the proposed approach to track multiple trajectories while generalizing to unseen ones. To do so, we define a training distribution of reference trajectories (circle, position step, eight-shape) and a distribution for these trajectory parameters (radius, velocity, position). During training, we sample at random a desired, 7s long (70 steps) reference with randomly sampled parameters, generating a demonstration and updating the proposed policy, while testing on a set of 20, 7s long trajectories randomly sampled from

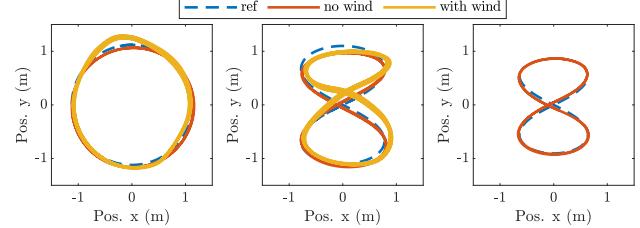


Fig. 8: Examples of different, arbitrary chosen trajectories from the training distribution, tested in hardware experiments with and without strong wind-like disturbances produced by leaf blowers. The employed policy is trained with 10 demonstrations (when other baseline methods have not fully converged yet, see Figure 7) using DAgger+SA (sparse). This highlights that sparse SA can learn multiple trajectories in a more sample-efficient way than other IL methods, retaining RTMPC’s robustness and performance. Prediction horizon used is $N = 20$, and DNN input size is adjusted accordingly.

the defined distributions. We monitor the robustness and performance of the different methods, with force disturbances applied in the target domain. The results of the numerical evaluation, shown in Figure 7, confirm that sparse SA i) achieves robustness and performance comparable to the one of the expert in a sample efficient way, requiring less than half the demonstrations than baseline approaches; ii) simultaneously learns to generalize to multiple trajectories randomly sampled from the training distribution. The hardware evaluation, performed with DAgger augmented via sparse SA, is shown in Figure 8. It confirms that the proposed approach is experimentally capable of tracking multiple trajectories under real-world disturbances/model errors.

V. CONCLUSION AND FUTURE WORK

This work has presented a demonstration-efficient strategy to compress a MPC in a computationally efficient representation, based on a DNN, via IL. We showed that greater sample efficiency and robustness than existing IL methods (DAgger, BC and their combination with DR) can be achieved by designing a Robust Tube variant of the given MPC, using properties of the tube to guide a sparse data augmentation strategy. Experimental results – showing trajectory tracking control for a multirotor after a *single* demonstration under wind-like disturbances – confirmed our numerical findings. Future work will focus on designing an adaptation strategy.

ACKNOWLEDGMENT

This work was funded by the Air Force Office of Scientific Research MURI FA9550-19-1-0386.

REFERENCES

- [1] F. Borrelli, A. Bemporad, and M. Morari, *Predictive control for linear and hybrid systems*. Cambridge University Press, 2017.
- [2] B. T. Lopez, J.-J. E. Slotine, and J. P. How, “Dynamic tube mpc for nonlinear systems,” in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1655–1662.
- [3] B. T. Lopez, “Adaptive robust model predictive control for nonlinear systems,” Ph.D. dissertation, Massachusetts Institute of Technology, 2019.
- [4] W. Li and E. Todorov, “Iterative linear quadratic regulator design for nonlinear biological movement systems.” in *ICINCO (1)*. Citeseer, 2004, pp. 222–229.
- [5] M. Kamel, M. Burri, and R. Siegwart, “Linear vs nonlinear mpc for trajectory tracking applied to rotary wing micro aerial vehicles,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3463–3469, 2017.
- [6] M. V. Minniti, F. Farshidian, R. Grandia, and M. Hutter, “Whole-body mpc for a dynamically stable mobile manipulator,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3687–3694, 2019.
- [7] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, “Aggressive driving with model predictive path integral control,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1433–1440.
- [8] J. B. Rawling and P. Kumar, “Industrial, large-scale model predictive control with deep neural networks,” http://helper.ipam.ucla.edu/publications/lco2020/lco2020_16406.pdf, (Accessed on 05/18/2021).
- [9] S. Levine and V. Koltun, “Guided policy search,” in *International conference on machine learning*. PMLR, 2013, pp. 1–9.
- [10] G. Kahn, T. Zhang, S. Levine, and P. Abbeel, “Plato: Policy learning using adaptive trajectory optimization,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3342–3349.
- [11] T. Zhang, G. Kahn, S. Levine, and P. Abbeel, “Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search,” in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 528–535.
- [12] J. Carius, F. Farshidian, and M. Hutter, “Mpc-net: A first principles guided policy search,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2897–2904, 2020.
- [13] E. Kaufmann, A. Loquercio, R. Ranftl, M. Müller, V. Koltun, and D. Scaramuzza, “Deep drone acrobatics,” *arXiv preprint arXiv:2006.05768*, 2020.
- [14] S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert, “Learning monocular reactive uav control in cluttered natural environments,” in *2013 IEEE international conference on robotics and automation*. IEEE, 2013, pp. 1765–1772.
- [15] A. Reske, J. Carius, Y. Ma, F. Farshidian, and M. Hutter, “Imitation learning from mpc for quadrupedal multi-gait control,” *arXiv preprint arXiv:2103.14331*, 2021.
- [16] D. A. Pomerleau, “Alvinn: An autonomous land vehicle in a neural network,” Carnegie-Mellon Univ Pittsburgh PA Artificial Intelligence and Psychology, Tech. Rep., 1989.
- [17] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, “An algorithmic perspective on imitation learning,” *arXiv preprint arXiv:1811.06711*, 2018.
- [18] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
- [19] S. Ross, 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*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
- [20] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Sim-to-real transfer of robotic control with dynamics randomization,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.
- [21] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, “Deep drone racing: From simulation to reality with domain randomization,” *IEEE Transactions on Robotics*, vol. 36, no. 1, pp. 1–14, 2019.
- [22] H. Tsukamoto and S.-J. Chung, “Learning-based robust motion planning with guaranteed stability: A contraction theory approach,” *IEEE Robotics and Automation Letters*, 2021.
- [23] ———, “Neural contraction metrics for robust estimation and control: A convex optimization approach,” *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 211–216, 2020.
- [24] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, “Agile autonomous driving using end-to-end deep imitation learning,” *arXiv preprint arXiv:1709.07174*, 2017.
- [25] A. Farchy, S. Barrett, P. MacAlpine, and P. Stone, “Humanoid robots learning to walk faster: From the real world to simulation and back,” in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, 2013, pp. 39–46.
- [26] Y. Chebotar, A. Handa, V. Makoviychuk, M. Macklin, J. Issac, N. Ratliff, and D. Fox, “Closing the sim-to-real loop: Adapting simulation randomization with real world experience,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8973–8979.
- [27] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, “Dart: Noise injection for robust imitation learning,” in *Conference on robot learning*. PMLR, 2017, pp. 143–156.
- [28] M. D. Laskey, “On and off-policy deep imitation learning for robotics,” Ph.D. dissertation, UC Berkeley, 2018.
- [29] J. Hanna and P. Stone, “Grounded action transformation for robot learning in simulation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [30] S. Desai, I. Durugkar, H. Karnan, G. Warnell, J. Hanna, P. Stone, and A. Sony, “An imitation from observation approach to transfer learning with dynamics mismatch,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [31] P. Florence, L. Manuelli, and R. Tedrake, “Self-supervised correspondence in visuomotor policy learning,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.
- [32] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, “Augmix: A simple data processing method to improve robustness and uncertainty,” *arXiv preprint arXiv:1912.02781*, 2019.
- [33] M. Shu, Y. Shen, M. C. Lin, and T. Goldstein, “Adversarial differentiable data augmentation for autonomous systems,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.
- [34] D. Antotsiou, C. Ciliberto, and T.-K. Kim, “Adversarial imitation learning with trajectory augmentation and correction,” *arXiv preprint arXiv:2103.13887*, 2021.
- [35] M. Toromanoff, E. Wirbel, F. Wilhelm, C. Vejarano, X. Perrotton, and F. Moutarde, “End to end vehicle lateral control using a single fisheye camera,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3613–3619.
- [36] A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus, “Learning robust control policies for end-to-end autonomous driving from data-driven simulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1143–1150, 2020.
- [37] D. Q. Mayne, M. M. Seron, and S. Raković, “Robust model predictive control of constrained linear systems with bounded disturbances,” *Automatica*, vol. 41, no. 2, pp. 219–224, 2005.
- [38] D. D. Fan, A.-a. Agha-mohammadi, and E. A. Theodorou, “Deep learning tubes for tube mpc,” *arXiv preprint arXiv:2002.01587*, 2020.
- [39] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [40] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [41] Aerospace Controls Laboratory, “snap-stack: Autopilot code and host tools for flying snapdragon flight-based vehicles,” <https://gitlab.com/mit-acl/fsw/snap-stack>, (Accessed on 02/23/2020).