2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

A Reinforcement Learning-based Control Strategy for Robust Interaction of Robotic Systems with Uncertain Environments

Diletta Sacerdoti, Federico Benzi and Cristian Secchi

Abstract— **In the context of interaction with unmodelled systems, it becomes imperative for a robot controller to possess the capability to dynamically adjust its actions in real-time, enhancing its resilience in the face of fluctuating environmental conditions. This adaptation process must be performed in a stability-preserving fashion, and resourcefully exploit the knowledge acquired during the interaction process. In this article, we propose a novel control strategy, based on the synergistic usage of state-of-the-art passivity-based control and Deep Reinforcement Learning (DRL). The concept of energy tank is used to provide stability guarantees for the interaction controller with uncertain environments, while an online learning policy allows to properly estimate the requirements of the task and adapt the controller accordingly, thus simultaneously achieving stability and performance. The proposed architecture is successfully validated through simulations and experiments with a collaborative manipulator in a surface polishing task.**

I. INTRODUCTION

In order to mirror human behavior and skills, robots have to be endowed with the capabilities of physically interacting with any environment, albeit uncertain or unmodelled, in a dexterous and adaptable way. In such scenarios, the robustness and flexibility of the robotic behavior are of paramount importance, especially in the presence of difficult-to-model phenomena such as friction [1] and body contact [2].

The robust stability of a robot is commonly ensured by resorting to state of the art interaction control strategies, such as impedance/admittance control [3]–[5]. The interaction process is modeled following an energy-centric perspective, which allows to ascertain the robust stability of the implemented behavior resorting to passivity-based control techniques [6], [7]. However, one of the main disadvantages of impedance/admittance control is its lack of flexibility, which leads to performance degradation when the environmental conditions diverge from the nominal ones. An effective solution to passively implement a variable impedance/admittance controller consists in leveraging energy tanks [8]–[10], a virtual energy reservoir that introduces an energy budget within the system for performing non-passive actions.

One of the longstanding issues in the formulation of the energy tank is the determination of the initial extent of this budget [11], a problem often addressed by providing conservative values *a priori* or by iterative approaches. This value determines the extent of the additional passivity margin of the system and, therefore, the amount of non-passive behaviors which can be implemented. Some recent works have made this initialization dependent on the nominal actions the robot has to perform [12], leveraging the concept of task energy introduced in [13]. This, however, requires the full knowledge of the model of the task, often unavailable, and cannot be adapted at run-time. Online refill approaches have been recently proposed in [14] and [15], which dynamically inject energy in the tank during task execution, instead of allocating the entire budget at the beginning of the task. However, the total energetic requirement is pre-calculated on the basis of nominal dynamics, and there is no online updating of this value. This renders the approach non-suitable for uncertain or time-varying scenarios. Additionally, this strategy foresees a technically non-passive approach according to traditional passivity theory [15]. In [16], the lower energy threshold value of the tank is updated online based on the difference between estimated and real power consumption, but the maximum storable energy value remains undefined, computable through some application-dependent guidelines. The authors of [17] introduce the concepts of energy freezing and melting to define a safety protocol capable of filtering dangerous actions, but the initialization energy value and the maximum flow rate still remain parameters to be tuned. Even the concept of power valves [18], [19], while improving robustness and safety by limiting the maximum extraction rate of energy from the tank, shifts the problem to the not trivial determination of a proper value for such power bound.

In general, attempting to calculate the value of the required energy *a priori* leads to inaccurate results, primarily due to the aforementioned modeling difficulty. In the context of system identification, Deep Learning (DL) and Deep Reinforcement Learning (DRL) based strategies are becoming increasingly popular. They treat systems as black boxes and, instead of deducing their dynamics from physical laws, they infer their behavior through data observation, identifying even highly nonlinear causal relationships. RL has found fertile ground in applications such as autonomous systems, both on-road vehicles [20] and AGVs and UGVs [21] [22], resource allocation [23], manipulation [24]–[26] and assembly [27] of objects with robotic arms. In many of these applications, the approach followed is entirely model-free. However, the "stand-alone" usage of these techniques has revealed some disadvantages, primarily the enormous amounts of data required and the complex parameter tuning.

The idea of making model-based and data-driven strategies coexist in the same control scheme to mutually reinforce each other is becoming increasingly promising and advantageous [28]–[31]. In particular, [32] and [33] introduce the first DRL algorithm with stability guarantees on a real robotic manipu-

D. Sacerdoti, F. Benzi and C. Secchi are with the Department of Sciences and Methods of Engineering, University of Modena and Reggio Emilia, Italy {name.surname}@unimore.it

lator. However, this approach, based on traditional Lyapunov analysis, lacks the flexibility and generality provided by the energy tank and requires new training from scratch for even minor task variations.

The goal of this paper is to combine model-based and data-driven methods for passively accomplishing interactive robotic tasks with uncertain environments via budgeting and periodic refilling of the energy tank. The synergy of these two approaches is realized by calculating online the energy needed to perform the task according to both modeled task parameters, as well as an additional factor estimated through DRL. This last term allows to compensate for any variation in the energy requirements compared to the nominal values due to uncertainties, disturbances, and unmodelled dynamics. The resulting energy "quanta" are periodically injected into the tank during task execution, constituting the budget to be used until the next energy update. Unlike other use in collaborative scenarios [34], RL is not employed here to directly generate robot control; rather, the learned policy is exploited to improve the flexibility of the underlying interaction controller.

This formulation provides flexibility and robustness to the tank, overcoming the necessity of correctly estimating the initial energy budget in design phase. This novel approach makes the tank proactively adapt to the current task necessities, thus further increasing the flexibility of the architecture.

To the best of the Authors' knowledge, the only work in literature which resembles our approach is [35], which features the usage of RL in an energy tank-based control scheme. However, in [35] the tank only serves to impose a predefined and unchanging constraint used as an exploration exit condition, as the allocated tank energy is known beforehand. Moreover, the hereby presented work is the first to formulate the energy tank as an hybrid system, and to provide a formal proof of passivity accordingly.

In summary, the contributions of this paper are as follows:

- an energy tank-based control algorithm that not only ensures passivity at the theoretical level but also practical stability for a physically interactive controller;
- a robust and flexible online budgeting strategy that solves the initialization problem of the tank, combining model-based and data-driven information;
- a validation of the architecture in multiple polishing tasks, with both simulations and experimental testing using a collaborative manipulator.

The remainder of the paper is structured as follows: in Sec. II, the theoretical foundations of passivity-based control and energy tanks are briefly explained. Sec. III introduces the addressed problem, while Sec. IV presents the overall control architecture. Sec. V explains the new learning-based periodic refilling policy of the tank. In Sec. VI, the formal demonstration of system passivity is provided, and the advantages of the strategy over existing ones are highlighted. Sec. VII describes the simulations and experiments conducted to validate the formulation. In Sec. VIII, conclusions are drawn and potential future developments are discussed.

II. Theoretical Background

Consider a nonlinear affine system:

$$\Sigma = \begin{cases} \dot{\boldsymbol{\chi}}(t) = f(\boldsymbol{\chi}(t)) + g(\boldsymbol{\chi}(t))\boldsymbol{u}(t) \\ \boldsymbol{y}(t) = v(\boldsymbol{\chi}(t)), \end{cases} \tag{1}$$

where $\boldsymbol{\chi} \in \mathbb{R}^n$ represents the state of the system, the input-output pair $(\boldsymbol{u}, \boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^m$ is the power-port connecting it with the external environment, and $f(\boldsymbol{\chi})$ and $g(\boldsymbol{\chi})$ are locally Lipschitz-continuous functions.

Definition 1. *[6] A system Σ is passive w.r.t the power port $(\boldsymbol{u}, \boldsymbol{y})$ if there exists a state-dependent lower bounded function $V : \mathbb{R}^n \to \mathbb{R}^+$ belonging to class C^1, called storage function, such that for all $t \geq 0$ the following relation holds:*

$$\int_0^t \boldsymbol{y}^T(\phi)\boldsymbol{u}(\phi)d\phi \geq V(\boldsymbol{\chi}(t)) - V(\boldsymbol{\chi}(0)) \geq -V(\boldsymbol{\chi}(0)). \tag{2}$$

Passivity is to be considered an essential condition for control, necessary and sufficient to obtain robust and coupled stability. In fact, it has been demonstrated that if a system does not exhibit passive dynamics in the eyes of the power port through which it is connected to the environment, it is always possible to identify a passive environment capable of destabilizing the interaction [7]. Non-passive actions can be passivized by leveraging energy tanks.

The energy tank is a virtual system formulated as:

$$\begin{cases} \dot{x}_t(t) = u_t(t) \\ y_t(t) = \frac{\partial T}{\partial x_t} = x_t(t), \end{cases} \tag{3}$$

with x_t, u_t and $y_t \in \mathbb{R}$ representing respectively the state of the tank and the input and output that connect it to the rest of the system. The non-negative function $T(x_t) : \mathbb{R} \to \mathbb{R}$ represents the energy stored in the tank at a given moment:

$$T(x_t) = \frac{1}{2}{x_t}^2. \tag{4}$$

Combining (3) and (4) yields:

$$\dot{T} = u_t(t)y_t(t). \tag{5}$$

After being initialized with a finite energy amount $T(x_t(0))$, energy tanks can be interconnected with any system (1) in a power-preserving way and can then be employed to reproduce any desired (possibly non-passive) port behavior by exploiting the following modulation:

$$\begin{cases} u_t(t) = \boldsymbol{a}^T(t)\boldsymbol{y}(t) \\ \boldsymbol{u}(t) = -\boldsymbol{a}(t)y_t(t), \end{cases} \tag{6}$$

where $\boldsymbol{a}(t) \in \mathbb{R}^m$ is a modulation term formulated as

$$\boldsymbol{a}(t) = -\frac{\boldsymbol{\gamma}(t)}{x_t(t)}, \tag{7}$$

and $\boldsymbol{\gamma}(t) \in \mathbb{R}^m$ is the desired input. From (6), we have that

$$\boldsymbol{y}^T(t)\boldsymbol{u}(t) = -\boldsymbol{y}^T(t)\boldsymbol{a}(t)y_t = -u_t(t)y_t(t). \tag{8}$$

This equality implies that the power that flows through the input-output port $(\boldsymbol{u}, \boldsymbol{y})$ is the same as the one flowing

through the port (u_t, y_t) of the tank. By substituting (6) into (3), we obtain the formulation of the modulated tank [36]:

$$\begin{cases} \dot{x}_t(t) = \boldsymbol{a}^T(t)\boldsymbol{y}(t) \\ \boldsymbol{u}(t) = -\boldsymbol{a}(t)y_t(t) = \boldsymbol{\gamma}(t), \end{cases} \tag{9}$$

The modulated tank (9) implements a passive exchange of energy as long as it is non-empty, i.e., $x_t > 0$, as this would otherwise lead to a singularity in (7).

Proposition 1. *[37] If $T(x_t(t)) \geq \varepsilon > 0$ for all $t \geq 0$, then the modulated tank (9) is passive independently of the desired value $\boldsymbol{\gamma}(t)$.*

If no more energy is available in the tank, only an approximation of the desired behavior can be implemented [9], [36], jeopardizing task execution. Thus, an improper choice for the initial tank energy $T(x_t(0))$ can severely degrade the performance.

III. Problem Statement

Consider a gravity-compensated fully actuated n-DOF robotic manipulator operating in an m-dimensional space and modeled by the following Euler-Lagrange task space equation:

$$\boldsymbol{M}(\boldsymbol{x})\ddot{\boldsymbol{x}} + \boldsymbol{C}(\boldsymbol{x}, \dot{\boldsymbol{x}})\dot{\boldsymbol{x}} + \boldsymbol{D}(\boldsymbol{x})\dot{\boldsymbol{x}} = -\boldsymbol{F}_c + \boldsymbol{F}_{ext}, \tag{10}$$

where $\boldsymbol{x}, \dot{\boldsymbol{x}}, \ddot{\boldsymbol{x}} \in \mathbb{R}^m$ indicate the end-effector pose, twist and acceleration, the matrices $\boldsymbol{M}, \boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{m \times m}$ are the inertia term, the centrifugal and Coriolis term and the viscous damping one, respectively. Finally $\boldsymbol{F}_c, \boldsymbol{F}_{ext} \in \mathbb{R}^m$ are the control wrench and the external force.

In order to perform a physically interactive task, we employ a state of the art hybrid force/impedance control [13], where the control action $\boldsymbol{F}_c$ is given by the composition of a force tracking controller and an impedance controller:

$$\boldsymbol{F}_c = \boldsymbol{u}_{c,f} + \boldsymbol{u}_{c,imp}, \tag{11}$$

with the two force terms computed as:

$$\boldsymbol{u}_{c,f} = \boldsymbol{K}_P(\boldsymbol{F}_{ext} - \bar{\boldsymbol{F}}) + \boldsymbol{K}_I \int_0^t (\boldsymbol{F}_{ext} - \bar{\boldsymbol{F}})d\sigma, \tag{12}$$

[1]

$$\boldsymbol{u}_{c,imp} = \boldsymbol{B}_x(\dot{\boldsymbol{x}} - \dot{\bar{\boldsymbol{x}}}) + \boldsymbol{K}_x(\boldsymbol{x} - \bar{\boldsymbol{x}}) \tag{13}$$

where $\bar{\boldsymbol{x}}, \dot{\bar{\boldsymbol{x}}}, \bar{\boldsymbol{F}} \in \mathbb{R}^m$ are the desired pose, twist and interaction wrench for the end-effector, $\boldsymbol{K}_P, \boldsymbol{K}_I \in \mathbb{R}^{m \times m}$ are the proportional and integral force tracking gains, and $\boldsymbol{B}_x, \boldsymbol{K}_x \in \mathbb{R}^{m \times m}$ are the desired impedance task-space damping and stiffness positive definite matrices. The gains in (12) and (13) are chosen such that the impedance is inactive along the force tracking directions.

This type of controller can lead to non-passive behaviors being implemented (see, e.g., [13], [16]). We thus recover passivity by interconnecting the system (10) with the modulated energy tank (9) via the following:

$$\begin{pmatrix} \boldsymbol{F}_c \\ u_t \end{pmatrix} = \begin{bmatrix} 0 & -\boldsymbol{a} \\ \boldsymbol{a}^T & 0 \end{bmatrix} \begin{pmatrix} \dot{\boldsymbol{x}} \\ y_t \end{pmatrix}. \tag{14}$$

[1]We omit time dependancy in both equations for the sake of readability

The presence of uncertainties and/or disturbances in the environmental dynamics, however, makes it significantly difficult to properly estimate the initial energy budget $T(x_t(0))$ in the tank. As mentioned, an improper initial choice leads to sub-optimal performances. In order to address this issue, we hereby propose a new strategy for dynamic regulation of the tank energy. Instead of initially allocating the entire budget, we periodically refill the energy tank according to the the nominal task conditions, and exploit DRL to account for the effect of environmental uncertainties onto the energetic requirements of the task.

IV. Online Refilled Energy Tank-Based Control

In this Section, we present the two layers control architecture and introduce the online energy refill procedure. The overall architecture is depicted in Fig. 1.

First, a low frequency task planner defines high level specifications for task execution, such as the trajectory to be followed $\bar{\boldsymbol{x}}$ and the desired interaction force $\bar{\boldsymbol{F}}$. These nominal conditions are then exploited to compute the amount of energy to be periodically refilled into the tank $\Delta\hat{T}_{TB}$. This value is calculated assuming perfect velocity tracking ($\dot{\boldsymbol{x}}(t) \approx \dot{\bar{\boldsymbol{x}}}(t)$) and estimating the value of the control wrench $\hat{\boldsymbol{F}}_c$ in such conditions using the model of the robot (10) and of the controllers (12), (13), up until the successive refill time instant:

$$\Delta\hat{T}_{TB} = \int_{t_k}^{t_{k+1}} \dot{\bar{\boldsymbol{x}}}^T(\sigma)\hat{\boldsymbol{F}}_c(\sigma)d\sigma, \tag{15}$$

with t_k, t_{k+1} being the instants of energetic injection. In ideal conditions, this periodic refill procedure would provide the tank precisely the amount of energy necessary to implement the desired task in the upcoming time window, namely until the next refill instant t_{k+1}, thus guaranteeing both perfect performance and avoiding unnecessary energy storing.

As this refill procedure is a purely feedforward approach, we still need to provide formal passivity guarantees for the low level, high frequency interaction controller (11). To this end, the desired force command $\boldsymbol{F}_{des}$ computed by (11) is filtered by an optimizer, scaling its values to respect the passivity constraint, i.e., avoiding tank depletion. The optimizer is formulated as follows:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \quad \|\boldsymbol{\alpha}(t) \odot \boldsymbol{F}_{des}(t) - \boldsymbol{F}_{des}(t)\|^2 \tag{16}$$

$$\text{s.t.} \quad T(t) - \tau\boldsymbol{\alpha}(t) \odot \boldsymbol{F}_{des}^T(t)\dot{\boldsymbol{x}}(t) \geq \varepsilon \tag{16a}$$

$$\mathbf{0} \preccurlyeq \boldsymbol{\alpha}(t) \preccurlyeq \mathbf{1} \tag{16b}$$

where $\odot, \preccurlyeq$ are component-wise operations, τ is the time step of the low level controller, while $\boldsymbol{\alpha} \in \mathbb{R}^m$ is the scaling factor. As showcased in [36] and [38], the solution of (16) provides the best passive approximation of the desired control action, i.e., the one that minimizes the deviation from the setpoint while respecting the passivity constraint. The implemented control input is then synthesized by applying the modulation term to the desired wrench, namely

$$\boldsymbol{F}_c(t) = \boldsymbol{\alpha}(t) \odot \boldsymbol{F}_{des}(t). \tag{17}$$

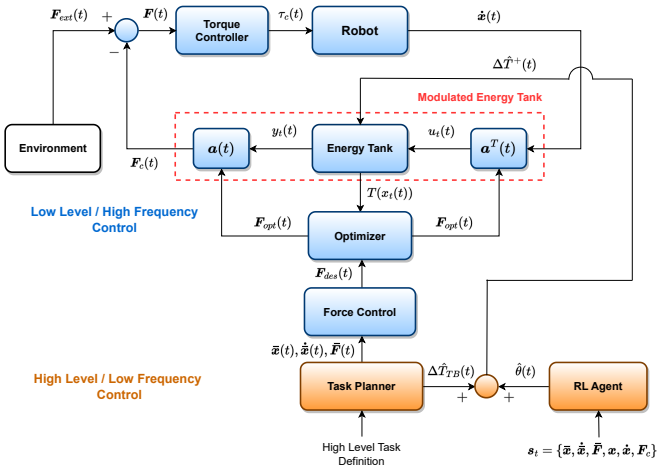


Fig. 1. Control architecture of the online tank refilling strategy. The total force command $\boldsymbol{F} = \boldsymbol{F}_c + \boldsymbol{F}_{ext}$ is then transmitted to the torque controller of the robot and applied via the control torque $\boldsymbol{\tau}_c$.

The discrete-time formulation of (16) is convex, thus can be solved in real-time with established state-of-the-art algorithms for quadratic programming (QP) (see, e.g., [36]). Differently from [36], [38], however, the hereby employed scaling factor used to filter the control input is an m-dimensional vector, rather than a scalar factor. In this way, we can act differently on each task direction, in order to better suit the nature of the hybrid force/impedance controller. Finally, since $\boldsymbol{\alpha} = \mathbf{0}$ is always an admissible solution, the problem is always feasible, even in case of tank depletion.

V. Learning How To Refill the Tank

The refilling procedure (15), as mentioned, can only function properly in nominal and ideal conditions, while failing to allocate the necessary budget in the presence of uncertainties, disturbances or unmodelled effects. To address this issue, we hereby exploit a DRL policy to estimate an additional quantity $\hat{\theta}$ to be added to the nominal energy budget $\Delta\hat{T}_{TB}$. The rationale behind the resort to RL is to allow the system to learn through experience how much energy is needed to perform a task in real scenarios, and specifically how this contributions differs from the nominal one. Thus, the amount of energy injected in the tank at each refill time t_k is

$$\Delta\hat{T}^{+}(t_k) = \Delta\hat{T}_{TB}(t_k) + \hat{\theta}(t_k). \tag{18}$$

A RL problem can be summarized as finding a policy to solve a Markov Decision Process (MDP), defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r)$. $\mathcal{S}$ and $\mathcal{A}$ represent respectively the continuous space states of the environment and the actions that the agent can take. At each time step t, the agent observes the current state of the environment $\boldsymbol{s}_t$ and, according to that, takes an action $\boldsymbol{a}_t$ dictated by a learnt policy $\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)$. We indicate with $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$ the probability density function quantifying the probability of each state transition $\boldsymbol{s}_t \to \boldsymbol{s}_{t+1}$, starting from state $\boldsymbol{s}_t$ and taking action $\boldsymbol{a}_t$. After a transition, the agent measures the optimality of its action through the calculation of the reward, given by the function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

In this work, we choose to adopt the Soft Actor-Critic algorithm (SAC) [39], an online off-policy actor-critic algorithm that aims to maximize not only the long-term reward but also the entropy $\mathcal{H}$ of the stochastic actor to promote stability and broad exploration, thus achieving asymptotic convergence to the optimal policy π^*. This policy is formulated as follows:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_t \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_t) \sim \rho_\pi}[r(\boldsymbol{s}_t, \boldsymbol{a}_t) + \alpha\mathcal{H}(\pi(\cdot|\boldsymbol{s}_t))], \tag{19}$$

where α is the temperature coefficient that regulates the relative priority of the entropic term $\mathcal{H}$ compared to the reward r, thereby determining the stochasticity of the policy.

In the specific context of this discussion, the observable state includes the vectors of robot desired and actual position, twist, and interaction wrench:

$$\boldsymbol{s}_t = \left\{\bar{\boldsymbol{x}}(t), \dot{\bar{\boldsymbol{x}}}(t), \bar{\boldsymbol{F}}(t), \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t), \boldsymbol{F}_c(t)\right\}. \tag{20}$$

The action corresponds to the amount of energy $\hat{\theta}$ recharged into the tank at each update in addition to the nominal refill ($a_t = \hat{\theta}(t)$) and the reward function is defined as follows:

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = 1 - |\Delta\hat{T}_{TB} + \hat{\theta} - \Delta T_{real}|. \tag{21}$$

The previous expression increases, approaching 1, as the energy budget recharged into the tank ($\Delta\hat{T}^{+}$) approaches the real energetic requirement (ΔT_{real}) for the given time interval. Thus, higher policy values indicate that the system has learned how to properly compensate for the effect of unmodelled environmental actions, resulting in a correct energy allocation at each time interval. For each energetic update, the flow of events develops as follows: *i.* the agent observes the current state (s_t) of the system and consequently takes action $\hat{\theta}$, which is added to ΔT_{TB} to refill the tank; *ii.* until the following update, the state of the robotic system evolves according to (10) exploiting the energy just recharged into the tank; *iii.* at the end of the current continuous interval, the reward is computed *a posteriori* with the value of ΔT_{real} calculated using the twist ($\dot{\boldsymbol{x}}$) and wrench ($\boldsymbol{F}$) vectors implemented on the real robot in the previous sample:

$$\Delta T_{real} = \int_{t_{k-1}}^{t_k} \dot{\boldsymbol{x}}^T(\sigma)\boldsymbol{F}_c(\sigma)d\sigma \tag{22}$$

. It is worth emphasizing that the use of rigorous physical models and analytical formulas for the evaluation of the major energy contribution $\Delta\hat{T}_{TB}$ significantly aids the convergence of the policy to its optimal form and increases the sample efficiency of the strategy, allowing for a substantial reduction in the dimension of the state to be observed and the duration of the neural networks training process. This aspect represents one of the main strengths of the formulation, capable of constructively combining the information contribution from the two sources.

Unlike [35], the passivity constraint (16a) is omitted during the learning phase because we want the agent to explore even non-passive policies, which will result in rewards lower than 1 in (21) since $\Delta\hat{T}^{+} < \Delta T_{real}$. Over time, the

agent will be discouraged from using non-passive policies as it experiences lower rewards, in favor of more accurate and task-adherent refilling. Although the passivity constraint (16a) is still included during the policy deployment phase for providing absolute passivity guarantees, ideally, i.e., with optimal policies and perfect training, its intervention should be minimal, if not null.

VI. Hybrid Tank Refill

Although the budgeting strategy (18) involves the repeated introduction of energy at run-time, we can still prove the passivity of the overall system. Since the state of the tank presents instantaneous variations, the proof of passivity requires recalling the theory of hybrid systems [40]–[42].

We apply the symbology of [42] to the tank subject to the refilling policy (18): $Q = \{q_1\}$, where Q represents the set of discrete states (or modes) in which the system assumes continuous dynamics, which in this case is always unique and represented by (3), with a single storage function T and a single continuous energy supply rate function defined as in (5); thereby, the set of possible transitions is also a singleton: $E = \{e_1\}$ with $e_1 = (q_1, q_1)$; t_k represents a generic switching instant as in (18). It is also assumed that the number of transitions is finite in a finite time (no Zeno behaviors). Additionally, since the value of $\Delta\hat{T}_{TB}$ computed via (15) is always bounded, and the policy is trained to only provide bounded $\hat{\theta}$, the refilled energy $\Delta\hat{T}^+$ is also bounded by an arbitrary value, namely $\Delta\hat{T}^+ < \Delta\overline{T}^+$.

Indeed, we can show that the following result holds:

Proposition 2. *The system* (10) *interconnected with the modulated tank via* (14)*, provided with the periodic energy refilling* (18) *is passive assuming* T *as a storage function.*

Proof. It is known (see, e.g., [43]) that (10) is passive w.r.t. $(-\boldsymbol{F}_c + \boldsymbol{F}_{ext}, \dot{\boldsymbol{x}})$ using $\mathcal{F} = \frac{1}{2}\dot{\boldsymbol{x}}^T\boldsymbol{M}\dot{\boldsymbol{x}}$, namely its kinetic energy, as a storage function. Since the interconnection (14) between the robot (10) and the tank is power-preserving, it is sufficient to demonstrate the passivity of the individual systems to prove that of the closed-loop system [4]. Thus, proving the passivity of the overall system (10) comes down to proving the passivity of the energy tank. For a hybrid system to be passive all the modes must be passive when active and the energy added to the system at discrete transitions must be bounded [42]. From (4), (18), we have:

1) Since the storage function T is defined as (4), there always exist two class $\mathcal{K}$-functions [2] $\underline{\alpha}(x_t)$ and $\overline{\alpha}(x_t)$, such that the following holds:
$$\underline{\alpha}(x_t) \leq T(x_t) \leq \overline{\alpha}(x_t). \tag{23}$$
During continuous mode, the system evolves as in (3). Hence, between every transition $e_1 = (q_1, q_1)$ occurring at switching instants t_k, we have that for every t_1, t_2 such that $t_{k-1} \leq t_1 \leq t_2 \leq t_k$:
$$T(x_t(t_2)) \leq T(x_t(t_1)) + \int_{t_1}^{t_2} y_t(\phi)u_t(\phi)d\phi. \tag{24}$$
2) Since at each refill instant t_k the tank energy switches discretely as:
$$T(x_t(t_k^+)) = T(x_t(t_k^-)) + \Delta\hat{T}^+(t_k), \tag{25}$$
and since the instantaneous supply $\Delta\hat{T}^+$ is bounded, we can guarantee that there exists a discrete energy supply rate $\omega_d(x_t, e)$, bounded by a class-$\mathcal{K}$ function $W(|x_t|)$, $\omega_d(x_t, e) \leq W(|x_t|)$, s.t.:
$$T(x_t(t_k^+)) \leq T(x_t(t_k^-)) + \omega_d(x(t_k), e(t_k)). \tag{26}$$
In fact, since in continuous mode $T(x_t) = \frac{1}{2}x_t^2$, and since $\Delta\hat{T}^+$ is bounded, it is always possible to find such $W(|x_t|)$ function.

Thus, according to theory of hybrid system (see [42], Def. 1) we can conclude that the energy tank periodically refilled via (18) is a dissipative system. Moreover, since the continuous supply rate function of the tank is given by the product of the power variables u_t and y_t, we can conclude that it is also passive with respect to the port (u_t, y_t).

This allows to infer the passivity of the interconnected system, thus concluding the proof. ∎

It is worth emphasizing that the choice of any value of the refilling frequency does not compromise the guarantee of passivity. As f_{ref} tends to zero, it brings the tank back to its traditional continuous system formulation. Its specific value can be tuned even online based on the task to perform and will be proportional to the level of uncertainties and disturbances involved in its execution. The higher the value of f_{ref}, the closer the energy control approximates power control, regulating instantaneous energy, somehow ending up taking the place of power valves [18] [19].

The strength of this strategy lies in the central role attributed to the energy tank as an effective control tool. It no longer only guarantees passivity on a mathematical level, but it also prevents instability by providing the system with exactly the energy it needs to perform a specific task, ensuring practical safety.

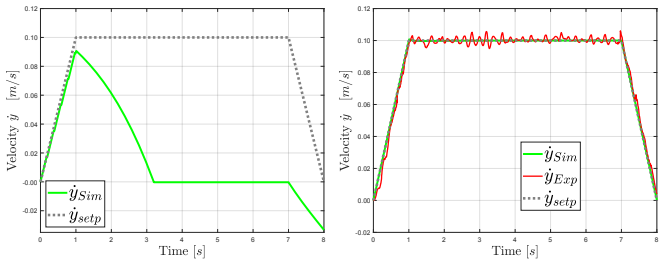
VII. Validation

The formulation was tested in the execution of a polishing task. The purpose of the validation is to demonstrate that the proposed strategy is capable of inferring the correct amount of energy required by the robot to complete the task, automatically adapting the varying environmental conditions. In particular, during learned policy deployment, the robot does not know in advance in which configuration it will have to perform the work, in terms of the type of tool used (brush or sponge) and the surface to polish (smooth wood, sandpaper or polystyrene), but, during the learning phase, it was able to experience with each of the 6 combinations.

The results of the performance of the policy applied to the sponge-polystyrene case are reported in this Section. Another experimental test is shown in the accompanying video.

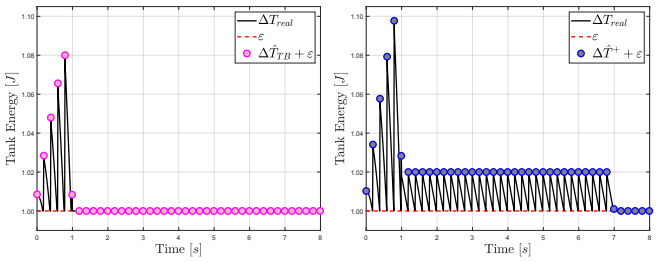
The task involves cleaning a strip of material by exerting a normal force $\bar{F}_3 = 10\,N$ along the z axis while simultaneously moving along a y-axis trajectory with a trapezoidal

[2] A class-$\mathcal{K}$ function is a function $\kappa : \mathbb{R}^+ \to \mathbb{R}$ such that κ is strictly increasing and $\kappa(0) = 0$



(a) (b)

Fig. 2. End-effector twist with only nominal refilling (a) and with combined nominal and learning-based refilling (b). We use the $\cdot_{setp}$ notation to indicate the velocity setpoint value, while $\cdot_{Sim}$ and $\cdot_{Exp}$ indicate the simulated and real experiment values of it.



(a) (b)

Fig. 3. Tank energy evolution with only nominal refilling (a) and with combined nominal and learning-based refilling (b)

velocity profile and a cruise velocity of $\dot{y}_{setp} = 0.1\,m/s$, for a total duration of 8 s. The nominal energy refilled in the tank is calculated as in (15) based on the setpoint to track, taking into account only the energy needed to accelerate the robot. All the energy required to overcome friction is estimated through RL. The refilling frequency f_{ref} was set to 5 Hz for graphical clarity (resulting in 5 tank updates per second).

A. Simulations

The validation has been first conducted in simulation within MATLAB environment using the Reinforcement Learning Toolbox and the Robotics System Toolbox. To reproduce the presence of the dynamic friction force $\boldsymbol{F}_{frc}$ exerted by the surface, the basic Coulomb model was employed, setting $F_{frc} = \mu F_z$, where $\mu = 0.1$ is the dynamic friction coefficient for the sponge-polystyrene interface. Static friction and velocity-dependent effects were neglected. Although this model is oversimplified, it allows taking into account the presence of friction during training and performing an initial coarse tuning of the parameters of the SAC learning algorithm.

We compare the task executed with tank refilling performed only with the nominal energy (Fig. 3(a)) and with the addition of the $\hat{\theta}$ contribution to overcome friction (Fig. 3(b)). While during acceleration the energy exploited to increase velocity is predominant and energy demands are almost entirely met by $\Delta\hat{T}_{TB}$, during the constant velocity phase almost all the energy consumption is due to overcoming friction. In Fig. 2(a), a speed drop is observed because the recharged energy is insufficient, and the optimizer scales the commanded tangential wrench. In Fig. 2(b), the tracking remains accurate (green line) because the addition of $\hat{\theta}$ at each refill increases the available budget, making friction compensation a passive action. It should be noted that during deceleration, the robot dissipates energy, which is used to overcome friction; so the energy provided by the policy is zero, as it is not necessary.

B. Experiments

The refilling strategy has been experimentally validated using a 7-DOF KUKA LWR 4+ robotic manipulator, controlled at a frequency of 500 Hz. Along the y-axis, tangential to the surface, impedance control was implemented according to (13), choosing diagonal matrices $\boldsymbol{B}_x$ with elements equal to 10 $^{Ns}/_{m}$ (translations) and 1 $^{Nms}/_{rad}$ (rotations), and $\boldsymbol{K}_x$ with elements set to 300 $^{N}/_{m}$ (translations) and 30 $^{Nm}/_{rad}$ (rotations). On the z-axis, the desired contact force was imposed using (12), with gains of $\boldsymbol{K}_P$ equal to 1 and gains of $\boldsymbol{K}_I$ set to 1.5 s^{-1}. Agent training was fine-tuned on the real-world scenario implementing SAC algorithm in PyTorch, exploiting the OpenAI gym environment. The parameter values obtained from simulation significantly speeds up the fine-tuning process on the experimental setup and improves sample efficiency. The final parameter values are reported in Table I. Fig. 2(b) (red line) shows that the setpoint tracking is faithfully reproduced even in the time interval dominated by friction, where the end-effector maintains a constant velocity. This is achieved because the policy provides the robot with just the energy needed to perform the task correctly, maintaining its value pretty much always at the minimum threshold before each update (Fig. 3(b)). Notice, in fact, that once the policy correctly estimates the surplus energy to be recharged into the tank, the correct execution of the task coincides with restoring the energy level in the tank to the minimum value ε at the end of the refill sample. Thus, the tank provides a performance metric of the actual operation or the robot.

TABLE I
LEARNING PARAMETERS

Algorithm:	SAC	Discount Factor:	0.99	Neurons per Hidden Layer:	64
Optimizer:	Adam	N° of Epochs:	6000	Target Smoothing Coeff.:	0.005
Actor Learning Rate:	10^{-4}	Steps per Epoch:	1000	Activation Function:	ReLU
Critic Learning Rate:	10^{-4}	N° of Layers:	3	Replay Buffer Size:	10^{6}

VIII. CONCLUSIONS AND FUTURE WORKS

In this paper, a new online refilling strategy for the energy tank was proposed, combining model knowledge with data provided by DRL. The results have demonstrated that this approach is capable of estimating the exact amount of energy required for the robot to perform the task effectively, passively, and adapting to unknown dynamic environments. Future developments of this work include testing the formulation in interactive scenarios, and using energy estimation predictively for refining task execution.

REFERENCES

[1] Z. Liu and R. D. Howe, "Beyond coulomb: Stochastic friction models for practical grasping and manipulation," *IEEE Robotics and Automation Letters*, 2023.

[2] T. Pang, H. T. Suh, L. Yang, and R. Tedrake, "Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models," *IEEE Transactions on Robotics*, 2023.

[3] N. Hogan, "Impedance control of industrial robots," *Robotics and computer-integrated manufacturing*, vol. 1, no. 1, pp. 97–113, 1984.

[4] J. E. Colgate and N. Hogan, "Robust control of dynamically interacting systems," *International journal of Control*, vol. 48, no. 1, pp. 65–88, 1988.

[5] L. Villani and J. De Schutter, "Force control," *Springer handbook of robotics*, pp. 195–220, 2016.

[6] R. Ortega and P. J. Nicklasson, "Passivity-based control of euler-lagrange systems: Mechanical, electrical and electromechanical," *Mechanical, Electrical and Electromechanical Applications*, 2013.

[7] S. Stramigioli, "Energy-aware robotics," in *Mathematical control theory I: Nonlinear and hybrid control systems*. Springer, 2015, pp. 37–50.

[8] M. Franken, S. Stramigioli, S. Misra, C. Secchi, and A. Macchelli, "Bilateral telemanipulation with time delays: A two-layer approach combining passivity and transparency," *IEEE transactions on robotics*, vol. 27, no. 4, pp. 741–756, 2011.

[9] F. Ferraguti, N. Preda, A. Manurung, M. Bonfe, O. Lambercy, R. Gassert, R. Muradore, P. Fiorini, and C. Secchi, "An energy tank-based interactive control architecture for autonomous and teleoperated robotic surgery," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1073–1088, 2015.

[10] C. Secchi, S. Stramigioli, and C. Fantuzzi, "Position drift compensation in port-hamiltonian based telemanipulation," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 4211–4216.

[11] F. Califano, R. Rashad, C. Secchi, and S. Stramigioli, "On the use of energy tanks for robotic systems," in *International Workshop on Human-Friendly Robotics*. Springer, 2022, pp. 174–188.

[12] A. Pupa, P. Robuffo Giordano, and C. Secchi, "Optimal energy tank initialization for minimum sensitivity to model uncertainties (forthcoming)," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[13] C. Schindlbeck and S. Haddadin, "Unified passivity-based cartesian force/impedance control for rigid and flexible joint robots via task-energy tanks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 440–447.

[14] B. Gerlagh, F. Califano, S. Stramigioli, and W. Roozing, "Energy-aware adaptive impedance control using offline task-based optimization," in *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE, 2021, pp. 187–194.

[15] F. Califano, D. van Dijk, and W. Roozing, "A task-based post-impact safety protocol based on energy tanks," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8791–8798, 2022.

[16] F. Benzi, M. Brunner, M. Tognon, C. Secchi, and R. Siegwart, "Adaptive tank-based control for aerial physical interaction with uncertain dynamic environments using energy-task estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9129–9136, 2022.

[17] R. Rashad, D. Bicego, J. Zult, S. Sanchez-Escalonilla, R. Jiao, A. Franchi, and S. Stramigioli, "Energy aware impedance control of a flying end-effector in the port-hamiltonian framework," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3936–3955, 2022.

[18] E. Shahriari, L. Johannsmeier, and S. Haddadin, "Valve-based virtual energy tanks: A framework to simultaneously passify controls and embed control objectives," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 3634–3641.

[19] E. Shahriari, L. Johannsmeier, E. Jensen, and S. Haddadin, "Power flow regulation, adaptation, and learning for intrinsically robust virtual energy tanks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 211–218, 2019.

[20] J. Li, L. Yao, X. Xu, B. Cheng, and J. Ren, "Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving," *Information Sciences*, vol. 532, pp. 110–124, 2020.

[21] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient uav control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, 2018.

[22] D. Wang, M. Hu, and J. D. Weir, "Simultaneous task and energy planning using deep reinforcement learning," *Information Sciences*, vol. 607, pp. 931–946, 2022.

[23] H. Wang, Y. Wu, G. Min, J. Xu, and P. Tang, "Data-driven dynamic resource scheduling for network slicing: A deep reinforcement learning approach," *Information Sciences*, vol. 498, pp. 106–116, 2019.

[24] A. Koenig, Z. Liu, L. Janson, and R. Howe, "The role of tactile sensing in learning and deploying grasp refinement algorithms," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7766–7772.

[25] K. M. Oikonomou, I. Kansizoglou, and A. Gasteratos, "A hybrid spiking neural network reinforcement learning agent for energy-efficient object manipulation," *Machines*, vol. 11, no. 2, p. 162, 2023.

[26] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, "A review on reinforcement learning for contact-rich robotic manipulation tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023.

[27] M. Braun and S. Wrede, "Incorporation of expert knowledge for learning robotic assembly tasks," in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1. IEEE, 2020, pp. 1594–1601.

[28] S. McIlvanna, N. N. Minh, Y. Sun, M. Van, and W. Naeem, "Reinforcement learning-enhanced control barrier functions for robot manipulators," *arXiv preprint arXiv:2211.11391*, 2022.

[29] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3387–3395.

[30] X. Li, H. Liu, and M. Dong, "A general framework of motion planning for redundant robot manipulator based on deep reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5253–5263, 2021.

[31] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. Molinari Tosatti, and N. Pedrocchi, "Model-based reinforcement learning variable impedance control for human-robot collaboration," *Journal of Intelligent & Robotic Systems*, vol. 100, no. 2, pp. 417–433, 2020.

[32] S. Khader, H. Yin, P. Falco, and D. Kragic, "Learning stable normalizing-flow control for robotic manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1644–1650.

[33] S. A. Khader, H. Yin, P. Falco, and D. Kragic, "Learning deep energy shaping policies for stability-guaranteed manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8583–8590, 2021.

[34] A. Perrusquía, W. Yu, and A. Soria, "Position/force control of robot manipulators using reinforcement learning," *Industrial Robot: the international journal of robotics research and application*, vol. 46, no. 2, pp. 267–280, 2019.

[35] R. Zanella, G. Palli, S. Stramigioli, and F. Califano, "Passivizing learned policies and learning passive policies with virtual energy tanks in robotics," *arXiv preprint arXiv:2301.12759*, 2023.

[36] F. Benzi, F. Ferraguti, G. Riggio, and C. Secchi, "An energy-based control architecture for shared autonomy," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3917–3935, 2022.

[37] C. Secchi and F. Ferraguti, "Energy optimization for a robust and flexible interaction control," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1919–1925.

[38] F. Benzi, F. Ferraguti, and C. Secchi, "Energy tank-based control framework for satisfying the iso/ts 15066 constraint," *arXiv preprint arXiv:2304.14059*, 2023.

[39] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[40] M. S. Branicky, "Stability of switched and hybrid systems," in *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, vol. 4. IEEE, 1994, pp. 3498–3503.

[41] M. Zefran, F. Bullo, and M. Stein, "A notion of passivity for hybrid systems," in *Proceedings of the 40th IEEE conference on decision and control (Cat. No. 01CH37228)*, vol. 1. IEEE, 2001, pp. 768–773.

[42] E. Agarwal, M. J. McCourt, and P. J. Antsaklis, "Dissipativity of hybrid systems: Feedback interconnections and networks," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 6060–6065.

[43] C. Secchi, S. Stramigioli, and C. Fantuzzi, *Control of interactive robotic interfaces: A port-Hamiltonian approach*. Springer, 2007, vol. 29.