The following publication A. Duan et al., "Safe Learning by Constraint-Aware Policy Optimization for Robotic Ultrasound Imaging," in IEEE Transactions on Automation Science and Engineering, vol. 22, pp. 2349-2360, 2025 is available at https://doi.org/10.1109/TASE.2024.3378915.

Safe Learning by Constraint-Aware Policy Optimization for Robotic Ultrasound Imaging

Anqing Duan, Chenguang Yang, *Senior Member, IEEE*, Jingyuan Zhao, Shengzeng Huo, Peng Zhou, Wanyu Ma, Yongping Zheng, *Senior Member, IEEE*, and David Navarro-Alarcon, *Senior Member, IEEE*

***Abstract*—Ultrasound-based medical examination usually requires establishing proper contact between an ultrasound probe and a human body that ensures the quality of ultrasound images. The scanning skills are quite challenging for a robot to learn primarily due to the complex coupling between the applied force profile and the resulting ultrasound image quality. While reinforcement learning appears as a powerful tool for learning complex robot skills, the deployment of these algorithms in medical robots demands special attention due to the evident safety concerns that arise from physical probe-tissue interactions. In this paper, we explicitly consider external constraints on the force magnitude when searching for the optimal policy parameters to enhance safety during ultrasound-guided robotic interventions. In particular, we study policy optimization under the framework of a constrained Markov decision process. The resulting gradient-based policy update is then subject to the involved constraints, which can be readily addressed by the primal-dual interior-point technique. In addition, upon the observation that policy update requires consecutive policies to be close to each other to have stable and robust performance with reinforcement learning algorithms, we design the learning rate of policy gradient from an imitation perspective. The performance of the proposed constraint-aware policy optimization method is validated with experiments of robotic ultrasound imaging for spinal diagnosis.**

***Note to Practitioners*—This paper was motivated by the problem of safely learning the optimal interaction force strategy to facilitate robotic ultrasound imaging. Existing approaches to robotic ultrasound imaging usually empirically set a constant value for the scanning force, despite the fact the force strategy plays an important role in the quality of the ultrasound images. This paper suggests the usage of reinforcement learning to identify the optimal interaction force due to the complex acoustic coupling between the force and the ultrasound image quality. Specifically, we propose constraint-aware reinforcement learning in view of the safety-critical issues as a result of physical human-probe interaction. We then conduct a theoretical analysis of the proposed safe reinforcement learning, including monotonic improvement and policy value bound under mild assumptions. Preliminary real experiments on ultrasound imaging of the spine of a phantom for scoliosis assessment suggest that the proposed approach can safely learn the optimal scanning force without violating the prescribed force threshold. In the future, we would like to apply our approach to learning the optimal scanning force on different organs of interest of human subjects.**

This research is funded in part by the Research Impact Fund of the HK Research Grants Council under grant R5017-18F, in part by PolyU through the Intra-Faculty Interdisciplinary Project under grant ZVVR, and in part by PolyU under grant G-UANS. *Corresponding author: D. Navarro-Alarcon.*

A. Duan, S. Huo, P. Zhou, W. Ma, Y. Zheng, and D. Navarro-Alarcon are with the Faculty of Engineering, The Hong Kong Polytechnic University, KLN, Hong Kong (e-mail: dna@ieee.org).

C. Yang is with Bristol Robotics Laboratory, University of the West of England, Bristol, BS16 1QY, UK (e-mail: cyang@ieee.org).

J. Zhao is with Nanyang Technological University, Singapore.

(a) (b)

Fig. 1. Illustration of medical diagnosis of scoliosis for adolescents through ultrasound scanning conducted by (a) a human operator and (b) a robotic arm.

***Index Terms*—Medical robotics, reinforcement learning, imitation learning, optimization, sequential decision making.**

I. Introduction

MEDICAL robots are promising in the healthcare industry as they can bring many benefits, such as freeing medical personnel from tedious jobs and standardizing the treatment procedure [1]. Yet, before medical assistive robots can widely penetrate people's daily life, a series of technical problems need to be addressed, with *safety* being the most critical one [2]. There are many examples of physical human-robot interactive systems in healthcare, e.g., robot-assisted stethoscopes, ultrasound scanning robots, prosthetics and orthotics, shock-wave therapy, automated massage systems, etc. [3]. To deploy these systems in the field, it is essential to guarantee their safe operation.

Due to the radiation-free, high portability, and non-invasiveness features, ultrasound imaging is extensively used in various types of diagnostic tasks and interventions [4]. In this paper, we particularly deal with ultrasound imaging of spines for *scoliosis* assessment, a procedure that evaluates the abnormal lateral curvature of spines [5]. Our focus is on safely learning the optimal force profile for ultrasound imaging of a spine with a robot arm [6].

The manual procedures for ultrasound-based scoliosis assessment are shown in Fig. 1(a), where a sonographer holds an ultrasound probe to capture images, which are used to reconstruct the spine's 3D structure [5]. During this task, the probe's motion is adjusted based on real-time ultrasound images; Fig. 2 shows the spine's anatomy with its different regions and corresponding ultrasound images. Learning the scanning skills for a robot arm (as depicted in Fig. 1(b))

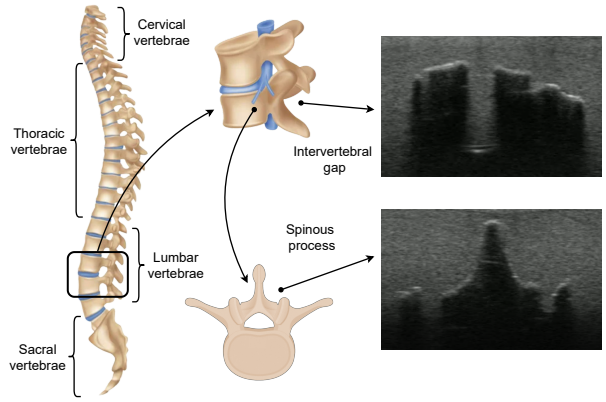


Fig. 2. Illustration of spinal anatomy with different regions labeled.

is a challenging task due to the complex coupling between the applied forces and the quality of the resulting ultrasound images [7]. The adequate control of these forces plays a decisive role in the quality of the reconstructed 3D spinal model [8]. Therefore, it is important to learn the optimal contact force strategies that can provide satisfying imaging results, which is precisely the goal of this paper.

From an algorithmic perspective, a popular technique for solving complex problems of robotic decision-making is policy optimization, a paradigm of reinforcement learning centered around the policy [9]. Policy optimization aims at finding an optimal set of control policy parameters based on the feedback of a specified cost or reward function. The generality of the reinforcement learning paradigm gradually makes it an attractive tool for autonomous ultrasound acquisitions, see e.g., [8], [10], [11]. Despite recent progress in this direction, the safety issue resulting from random exploratory policies (which may result in large contact forces) is usually overlooked. This situation requires special attention in our application scenario as it entirely relies on physical interactions between the robot-manipulated probe and the human patient [12].

We propose to complement the policy update procedure of vanilla policy gradient with user-specified force constraints. In addition, upon the observation that policy optimization restricts consecutive policies to be close to each other, we design the learning rate of policy gradient from an imitation principle. In brief, our contributions are outlined as follows:

- Safe policy learning for robotic ultrasound imaging with constraint-aware policy optimization;
- Imitation learning-informed design of the learning rate for updating policy parameters;
- Theoretical analysis of the properties of the proposed safe learning method;
- Experimental studies for validating the proposed method by safely scanning a phantom's spine.

The remainder of the paper is organized as follows. Related work is discussed in Sec. II. Subsequently, Sec. III reviews relevant preliminaries on the technical background. The proposed methods are presented in Sec. IV, analyzed in Sec. V, and evaluated in Sec. VI. Finally, Sec. VII concludes the paper.

II. Related Work

Robotic ultrasound image acquisition has received gradual research attention in recent years [4]. It has been observed that various robotic techniques have been applied to scan different parts of the human anatomy, such as limbs [13], breasts [14], livers [15], kidneys [11], etc. Despite the aforementioned successful cases, it is noted, however, that safely learning the optimal contact force still remains an under-explored topic. The aforementioned studies mainly prescribe the contact force between the tissue and the probe with a reference force profile, which is usually determined empirically and heavily depends on the expertise of a sonographer. To this end, in this paper, our focus lies on developing methods to safely search for the optimal contact force that guarantees the quality of the captured ultrasound images.

Due to the complex acoustic coupling between the interaction force and the ultrasound image quality, we resort to reinforcement learning to identify such an optimal force profile. Moreover, given the safety concerns inherent to any physical human-robot interaction task, learning of the optimal interaction force profile should take place whilst respecting external constraints which limit the maximum allowable force output. We refer to this kind of approach as *safe reinforcement learning* [16]. Compared with vanilla reinforcement learning algorithms, safe reinforcement learning, which remains a trending research topic, is thought to be more suitable in our case due to its attribute of *constraint awareness*.

For instance, constrained policy optimization is proposed to tackle the issue of safe reinforcement learning by addressing an optimization problem formulated in the form of a linear objective with linear and quadratic constraints [17]. Besides, Lyapunov-based approaches can be leveraged to guarantee safety by transforming reinforcement learning algorithms into their safe counterparts [18]. In addition to model-free approaches, model information can also be used to facilitate enhancing the feasibility of policy learning where a generalized control barrier function is defined to penalize the trends of approaching the constraint boundary [19]. Distinct from the aforementioned approaches on constrained reinforcement learning, our proposed new method addresses the safe learning problem under the mirror descent framework [20], where the learning rate for policy updates is devised from the principle of imitation learning. The resulting nonlinear optimization problem for policy update is addressed by the primal-dual interior-point method. Moreover, monotonic improvement can be presented under certain assumptions, and policy value bound is provided using the error propagation analysis.

Another closely related research topic on safe learning is constrained learning from demonstration, which concentrates on bounding policy representations [21], especially the well-established movement primitives, such as constrained DMP [22], constrained ProMPs [23], and constrained KMP [24]. While constrained learning from demonstration has shown favorable outcomes in dealing with the problem of limit-violation avoidance for robot trajectory generation, our proposed approach focuses instead on safe policy search, which applies to more general scenarios.

III. PRELIMINARIES

A. Markov Decision Process

A Markov decision process (MDP) is a tuple $M = (\mathcal{S}, \mathcal{A}, P, c, \gamma, d_0)$ where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a transition probability distribution, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a cost function, $\gamma \in [0, 1)$ is a discount factor, and $d_0 \to \mathbb{R}$ is the starting state distribution. We assume that the absolute value of the cost function is upper bounded by $C_{\max}$, i.e. $|c(s, a)| \leq C_{\max}$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$. A stochastic policy $\pi \in \Pi : S \to \mathcal{P}(A)$ returns a distribution over actions given the encountered state, where Π denotes the set of all stationary policies.

The Markov decision process works as follows. The agent starts from a state sampled by $s_0 \sim d_0$. Then at each time step t, the agent executes an action a_t according to $\pi(\cdot|s_t)$, receives an immediate cost $c(s_t, a_t)$, and observes the next state s_{t+1} according to the process transition dynamics $\mathcal{P}(\cdot|s_t, a_t)$. The above procedure repeats and finally results in a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ that records the agent-environment interaction history. Additionally, one useful notion for studying infinite-horizon MDP is defined, namely the discounted state visitation frequency:

$$d_\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi, d_0), \tag{1}$$

which characterizes visitation measure over states.

B. Reinforcement Learning

The goal of reinforcement learning is to find an optimal policy π^* such that it can minimize the expected cumulative discounted cost $J(\pi)$. Formally, reinforcement learning solves the following optimization problem:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmin}} J(\pi), \text{ with } J(\pi) = \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \right] \tag{2}$$

where $s_0 \sim d_0$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$.

Typically, in the context of robotics, (2) is tackled by policy gradient algorithms, where the learned policy π_θ belongs to a parametric policy set Π_Θ with $\theta \in \Theta$ and $\Theta \subset \mathbb{R}^d$. Specifically, the gradient of $J(\pi_\theta)$ with respect to policy parameter θ is given by [25]:

$$\nabla_\theta J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \Psi \right], \tag{3}$$

where $\Psi = \sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$. Akin to stochastic gradient descent, the gradient (3) can be estimated from a number $\mathcal{I}$ of trajectory samples $\{\tau_i\}_{i=1}^{\mathcal{I}}$,

$$g := \hat{\nabla}_\theta J(\pi_\theta) = \frac{1}{\mathcal{I}} \sum_{i=1}^{\mathcal{I}} \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) \right) \Psi_i. \tag{4}$$

Then at the n-th iteration, the parameters of the policy can be simply updated with policy gradient:

$$\theta_{n+1} = \theta_n - \eta_n g, \tag{5}$$

where $\eta_n > 0$ is a learning rate of the n-th iteration.

Furthermore, given any two policies π and π', their respective values are related by the well-known performance difference lemma [26]:

$$J(\pi') = J(\pi) + \mathop{\mathbb{E}}_{\tau \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t) \right], \tag{6}$$

where $A_\pi(s, a)$ is called the advantage function whose definition is given by:

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s), \tag{7}$$

where the state-action value function $Q_\pi(s, a)$ and the value function $V_\pi(s)$ are respectively given by:

$$Q_\pi(s, a) = \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s, a_0 = a \right],$$

$$V_\pi(s) = \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s \right].$$

C. Problem Statement

As discussed in Sec. I, reinforcement learning emerges as a promising tool to identify the optimal contact force scheduling that generates high-quality ultrasound images of the spine. Nevertheless, it calls for extra attention on restricting contact force strength when directly applying (5) in search for the optimal parameters of the controller. To this end, we propose to take into account the external constraints on the force magnitude. Specifically, a Markov decision process associated with external constraints forms a constrained Markov decision process (CMDP) [27]. We denote the concerning cost functions as $\mathcal{C} = \{C_1, \ldots, C_m\}$ with $C_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ being the corresponding cost function and defined by:

$$J_{Ci}(\pi) = \mathop{\mathbb{E}}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t) \right]. \tag{8}$$

Each cost function needs to be bounded by its corresponding threshold d_i. Consequently, the policy optimization problem within the framework of CMDP is formulated as:

$$\pi_C^* = \underset{\pi_\theta \in \Pi_\Theta}{\operatorname{argmin}} J(\pi_\theta) \tag{9a}$$

$$\text{s.t.} \quad J_{Ci}(\pi_\theta) \leq d_i, \; i = 1, \ldots, m. \tag{9b}$$

Intuitively, the optimal policy π_C^* from (9) accumulates the minimum possible costs during interacting with the environment while respecting the hard constraints by expectation. Such policy optimization procedures featured by constraint awareness will enhance safety and thus be more suitable for medical robots interacting with human bodies.

IV. METHODOLOGY

To reconstruct a qualifying 3D spinal image for scoliosis assessment, the whole back of the test subject needs to be scanned with the ultrasound probe from the waist to the neck, as shown in Fig. 3. To safely learn the optimal interaction strategy, we first introduce the mirror descent framework in Sec. IV-A. We then illustrate the design of the learning rate in Sec. IV-B. Finally, constraint-aware policy optimization is elaborated in Sec. IV-C.

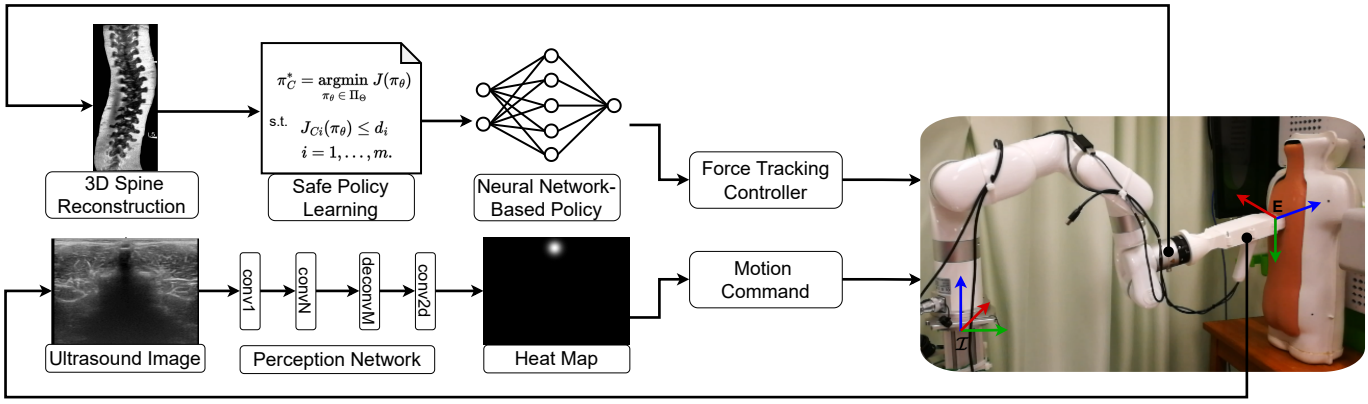


Fig. 3. Conceptual diagram of the proposed methodology for safely learning the optimal contact force in ultrasound imaging. The upper loop denotes the optimal force learning procedure and the lower loop denotes the motion control loop. The optimal force strategy is learned through the proposed constraint-aware policy optimization where the cost function is designed by the quality of ultrasound images.

A. Mirror Descent Framework

As a central topic in online optimization, mirror descent is increasingly employed to study decision-making algorithms due to its high analog to on-policy decision-making algorithms. Notably, several reinforcement learning algorithms and imitation learning algorithms can be recovered by the mirror descent framework [20]. We first cast the iterative update rule of policy optimization into the mirror descent framework.

Formally, given the first-order oracle g, policy update at the n-th iteration conforming to mirror descent is formulated as:

$$\theta_{n+1} = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \eta_n \langle g, \theta \rangle + D_h(\theta \| \theta_n), \tag{10}$$

where $D_h : \Theta \times \Theta \to \mathbb{R}$ is the Bregman divergence from θ_n to θ with respect to the function $h : \Theta \to \mathbb{R}$ [28]. Specifically, D_h is defined as:

$$D_h(\theta \| \theta_n) = h(\theta) - h(\theta_n) - \langle \nabla h(\theta_n), \theta - \theta_n \rangle. \tag{11}$$

By choosing $h(\theta) = \frac{1}{2}\|\theta\|^2$, vanilla policy gradient update (5) can then be recovered. It will be shown later that mirror descent provides a convenient framework to analyze policy search subject to constraints.

B. Imitation-Guided Learning Rate

In this section, we consider the design of the learning rate η_n in (10), which decides how far the policy parameters progress in one iteration and is often left as a constant hyper-parameter in vanilla policy gradient. Generally speaking, the value of a learning rate plays an important role in the algorithm's performance. For the reinforcement learning process to be stable, it has been observed that one common strategy is to additionally refrain the new policy from deviating too far away from the old one.

For example, in trust region policy optimization (TRPO), the maximum Kullback–Leibler (KL) divergence between the old and the new policy is constrained given any states [9]. Due to the intractability of the constraint for numerical optimization and estimation, the practical implementation of TRPO instead uses an average KL divergence between the old policy and the new policy averaged over the state distribution induced by the old policy, which is de facto behavior cloning [29].

Similarly, relative entropy policy search (REPS) constrains the KL divergence between the old and the new state-action distribution [30]. As there is a one-to-one correspondence between policy and state-action visitation frequency, given a state-action occupancy measure, its corresponding policy is unique [31]. In other words, the constraint of the observed state-action distribution and the state-action distribution induced by the new policy in REPS is also equivalent to making the new policy imitate the old one, sharing the same spirit of generative adversarial imitation learning [31].

From an imitation perspective to constraining the discrepancy between consecutive policies, we view the old policy π_{θ_n} as an expert and new policy $\pi_{\theta_{n+1}}$ as a learner. Our goal is to bound the value discrepancy between π_{θ_n} and $\pi_{\theta_{n+1}}$. To this end, the imitation loss between the expert and the learner can be quantified as[1]:

$$\mathop{\mathbb{E}}_{s \sim d\pi_{n+1}} \left[D_{\mathrm{KL}}(\pi_n(\cdot|s), \pi_{n+1}(\cdot|s)) \right], \tag{12}$$

where D_{KL} denotes the KL divergence between two probability distributions [32]. Due to the unknown and complex dynamics, the state distribution $d\pi_{n+1}$ induced by the learner policy π_{n+1} is typically hard to obtain. In practice, one common solution to this issue is the so-called behavior cloning where expert policy's state distribution $d\pi_n$ is employed in lieu of the learner policy distribution $d\pi_{n+1}$. The imitation loss that we would like to bound is then written as:

$$J_{IL}(\pi_{n+1}) = \mathop{\mathbb{E}}_{s \sim d_{\pi_n}} \left[D_{\mathrm{KL}}(\pi_n(\cdot|s), \pi_{n+1}(\cdot|s)) \right]. \tag{13}$$

Our designing principle on learning rate is that the policy parameters shall increment by a maximum allowable step along the first-order oracle within a given imitation loss δ. To this end, we first derive the Taylor expansions of (13) with respect to θ_{n+1} evaluated at θ_n to yield the relationship

[1]For simplicity, we write π_{θ_n} as π_n in the following.

between policy increment value $\eta_n g$ as well as imitation discrepancy δ:

$$J_{IL}(\pi) \approx J_{IL}(\pi_n) + \nabla J_{IL}^{\intercal}(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)^{\intercal} F(\theta_n)(\theta - \theta_n) + o(\|\theta - \theta_n\|^2), \quad (14)$$

where $J_{IL}(\pi_n) = 0$ since the imitation discrepancy is zero when two policies are identical. Also, as the Taylor expansion is evaluated at θ_n which serves as the optimal expert policy by definition, the first order derivative thus remains zero, namely $\nabla J_{IL} = 0$. Notably, the fact that both $J_{IL}(\pi_n)$ and ∇J_{IL} equal to zero shall be analytically verified by the properties of the KL divergence. Finally, $\nabla^2 J_{IL} = F(\theta_n)$ is the Fisher information matrix [26]:

$$F(\theta_n) = \mathop{\mathbb{E}}_{\substack{s\sim d_{\pi_n} \\ a\sim\pi_n}} \left[\nabla \log(\pi_n(a|s)) \nabla \log(\pi_n(a|s))^{\intercal}\right]. \quad (15)$$

As the largest learning step happens when the new policy results in the maximum allowable imitation loss δ compared with the previous policy, this implies $J_{IL}(\pi_{n+1}) = \delta$. By substituting (5) into (14), we then have:

$$\frac{1}{2}\eta_n g^{\intercal} F(\theta_n) \eta_n g = \delta. \quad (16)$$

Consequently, the imitation-guided learning rate η_n at n-th iteration is obtained as:

$$\eta_n = \sqrt{\frac{2\delta}{g^{\intercal} F(\theta_n) g}}. \quad (17)$$

Compared with the constant learning rate that is used in traditional policy gradient methods, our developed adaptive learning rate dictates the step size of the policy progress from a principled perspective of imitation learning.

C. Constraint-Aware Policy Optimization

Our ultimate goal is to search for the optimal policy parameters that minimize the cost function under external constraints as formulated in (9). Though performing vanilla reinforcement learning algorithms such as (10) can lead to a solution to (9a), the obtained optimal parameters could potentially violate the constraints as specified by (9b).

To guarantee that the policy search procedure respects the constraints over the course of task execution, we propose to augment (10) with (9b) such that each step of parameter update will take place within the allowed safety region, which is of critical importance in the case of physical human-robot interaction. Therefore, at each update step, a constrained optimization problem needs to be solved.

To make the constraints trackable, we propose to linearize the constraints following a similar treatment of (3) as shown in [17]. Specifically, the constraints at the n-th iteration are approximated as:

$$p_i^{\intercal}(\theta - \theta_n) + \phi_i \leq 0,\ i = 1, \ldots, m, \quad (18)$$

where $p_i = \nabla J_{Ci}$ represents the gradient of the corresponding constraint C_i and we denote:

$$\phi_i = J_{Ci}(\pi_n) - d_i. \quad (19)$$

By imposing (18) on (10), a constraint-aware version of the policy update is thus formulated as:

$$\min_{\theta\in\Theta} \quad f(\theta) \quad (20a)$$

$$\text{s.t.} \quad p_i^{\intercal}(\theta - \theta_n) + \phi_i \leq 0,\ i = 1, \ldots, m. \quad (20b)$$

For simplicity, we denote:

$$f(\theta) = \eta_n \langle g, \theta\rangle + D_h(\theta\|\theta_n).$$

It can be observed that (20) constitutes a constrained nonlinear optimization problem. In order to solve it, we employ a primal-dual interior-point method that is simple and efficient for solving constrained optimization problems [33].

To implement the primal-dual interior-point method, the inequality-constrained optimization problem (20) is first transformed into an unconstrained optimization problem with the help of barrier functions. The corresponding logarithmic barrier function associated with (20b) to convert the inequality constraints into a penalizing term is defined by:

$$-\sum_{i=1}^{m} \log(-p_i^{\intercal}(\theta - \theta_n) - \phi_i). \quad (21)$$

Consequently, the unconstrained optimization problem upon integrating (21) into the objective (20a) is given by:

$$\Phi(\theta, \mu) = -f(\theta) - \mu \sum_{i=1}^{m} \log(-p_i^{\intercal}(\theta - \theta_n) - \phi_i), \quad (22)$$

where $\mu > 0$ is called the barrier parameter, which usually makes the problem (22) a better approximation of (20) as it approaches zero.

To tackle the minimization problem (22), we equate its derivative with respect to the design variables to zero, which implies that:

$$\nabla\Phi = -\beta(\theta) - \mu \sum_{i=1}^{m} \frac{p_i}{p_i^{\intercal}(\theta - \theta_n) + \phi_i} \quad (23a)$$

$$= -\beta(\theta) - \sum_{i=1}^{m} \lambda_i p_i \quad (23b)$$

$$= -\beta(\theta) - P^{\intercal}\lambda = 0, \quad (23c)$$

where we denote

$$\beta(\theta) = \nabla f(\theta) = \nabla h(\theta) + (\eta_n g - \nabla h(\theta_n)).$$

And we have $P = \left[p_1, \ldots, p_m\right]^{\intercal}$. The dual variable $\lambda \in \mathbb{R}^m$ is introduced as:

$$\lambda_i = \frac{\mu}{-p_i^{\intercal}(\theta - \theta_n) - \phi_i}. \quad (24)$$

The zeros of (23), which represent the solution to (22), can be found with the help of root-finding algorithms. Here we employ the popular Newton's method to address (23c) and (24). The search directions can be determined by solving the following system of linear equations:

$$\begin{bmatrix} -\nabla^2 h(\theta) & -P^{\intercal} \\ \mathbf{I}\lambda P & C \end{bmatrix} \begin{bmatrix} \delta\theta \\ \delta\lambda \end{bmatrix} = \begin{bmatrix} -\beta(\theta) + P^{\intercal}\lambda \\ \mu\mathbf{1} - C\lambda \end{bmatrix}, \quad (25)$$

where $C = \texttt{diag}(-p_1^{\intercal}(\theta - \theta_n) - \phi_1, \ldots, -p_m^{\intercal}(\theta - \theta_n) - \phi_m)$ with $\texttt{diag}(\cdot)$ returning a diagonal matrix, $\mathbf{I}$ denotes the identity

Algorithm 1 Constraint-Aware Policy Optimization

Input: Initial policy parameters θ_0, trajectory samples number $\mathcal{I}$, max policy optimization iteration N, max interior point iteration K, imitation loss δ, and step size η_θ and η_λ;
for $n = 0, 1, \ldots, N$ **do**
  for $i = 1, 2, \ldots, \mathcal{I}$ **do**
    Roll out policy θ_n and collect trajectory τ_i;
  end for
  Compute the Fisher information matrix $F(\theta_n)$ as per (15);
  Calculate the policy learning rate η_n as per (17);
  Estimate the policy gradient g as per (4);
  Approximate the constraints C as per (18);
  for $k = 0, 1, \ldots, K$ **do**
    Form the system of linear equations as per (25);
    Update policy parameters as per (26);
  end for
end for
Output: Optimal parameters θ^*.

matrix, and $\mathbf{1}$ denotes a (column) vector of ones. The solution to (25) is then used to perform iterations as follows:

$$\begin{bmatrix}\theta_{k+1}\\ \lambda_{k+1}\end{bmatrix} = \begin{bmatrix}\theta_k\\ \lambda_k\end{bmatrix} + \begin{bmatrix}\eta_\theta \delta\theta_k\\ \eta_\lambda \delta\lambda_k\end{bmatrix} \tag{26}$$

where η_θ and η_λ denote the step size for θ and λ, respectively. Each step of the policy update procedure will undergo iteration as (26), which guarantees that the obtained policy will respect the specified constraints. The overall procedure for safe reinforcement learning of scanning skills is summarized in Algorithm 1.

V. Theoretical Analysis

This section analyzes the performance of the proposed constraint-aware policy optimization. In particular, we prove the monotonic improvement of the proposed algorithm (Sec. V-A) as well as the bound of value difference between consecutive policies (Sec. V-B).

To begin with, the following assumptions are made to facilitate theoretical analysis as in prior work on sequential decision making [34].

Assumption 1. We assume that h is α-strong convex, namely $\forall\theta_1, \theta_2 \in \Theta$, $\exists\alpha > 0$ such that:

$$(\nabla h(\theta_1) - \nabla h(\theta_2))(\theta_1 - \theta_2) \geq \alpha\|\theta_1 - \theta_2\|_2^2. \tag{27}$$

Assumption 2. We assume that J is l-Lipschitz continuous, namely $\forall\theta_1, \theta_2 \in \Theta$, $\exists l > 0$ such that:

$$J(\theta_2) - J(\theta_1) \leq \langle\nabla J(\theta_1), \theta_2 - \theta_1\rangle + \frac{l}{2}\|\theta_1 - \theta_2\|_2^2. \tag{28}$$

A. Monotonic Improvement

We first consider proving monotonic improvement for constraint-free policy optimization.

Proposition 1. In the case of $\eta_n l < 2\alpha$, monotonic improvement for (10) holds, i.e.

$$J(\theta_{n+1}) < J(\theta_n). \tag{29}$$

Proof. Using the fact of optimality and convexity properties for (10), we have:

$$\langle g + \frac{1}{\eta_n}(\nabla h(\theta_{n+1}) - \nabla h(\theta_n)), \theta - \theta_{n+1}\rangle \geq 0,\ \forall\theta \in \Theta. \tag{30}$$

By re-arranging (30), we then have:

$$\begin{aligned}\langle g, \theta_n - \theta_{n+1}\rangle &\geq \frac{1}{\eta_n}\langle\nabla h(\theta_n) - \nabla h(\theta_{n+1}), \theta_n - \theta_{n+1}\rangle\\ &\geq \frac{\alpha}{\eta_n}\|\theta_{n+1} - \theta_n\|_2^2.\end{aligned} \tag{31}$$

From the smoothness assumption on J made in Assumption 2, it can be obtained that:

$$\begin{aligned}&J(\theta_{n+1}) - J(\theta_n)\\ \leq&\langle\nabla J(\theta_n), \theta_{n+1} - \theta_n\rangle + \frac{l}{2}\|\theta_{n+1} - \theta_n\|_2^2 &\text{(32a)}\\ =&\langle g, \theta_{n+1} - \theta_n\rangle + \frac{l}{2}\|\theta_{n+1} - \theta_n\|_2^2 &\text{(32b)}\\ \leq&(-\frac{\alpha}{\eta_n} + \frac{l}{2})\|\theta_{n+1} - \theta_n\|_2^2. &\text{(32c)}\end{aligned}$$

Therefore, when $\eta_n l < 2\alpha$, (32c) becomes negative, which yields $J(\theta_{n+1}) < J(\theta_n)$. ■

This implies that the cumulative cost is reduced after each iteration and thus policy improves monotonically. Note that here we use g as an expectation of policy gradient. For a more general analysis that takes randomness of sampling g into account, interested readers are referred to [20].

In the case where the constraints are imposed, we consider studying the issue of monotonic improvement in terms of two situations, depending whether the constraints are triggered or not. If the constraints are not triggered during optimization, the property of monotonic improvement can be concluded following the same proof for the unconstrained case as in Proposition 1. We then prove that monotonic improvement also holds in the case where there are constraints triggered during policy optimization with $\eta_n l < 2\alpha$. To this end, we first show that (31) holds for the case of constraints triggered as well, as illustrated in the following lemma.

Lemma 1. Consider (20) under Assumption 1 and 2; It holds that:

$$\langle g, \theta_n - \theta_{n+1}\rangle \geq \frac{\alpha}{\eta_n}\|\theta_{n+1} - \theta_n\|_2^2. \tag{33}$$

Proof. Let the indices of the active constraints collectively expressed as $C_a = [C_{a1}, \ldots, C_{ak}]$. When considering only the active constraints, (20) shares the same solution as the following inequality-constrained optimization problem:

$$\begin{aligned}\min_{\theta\in\Theta}\quad &\eta_n\langle g, \theta\rangle + D_h(\theta\|\theta_n) &\text{(34a)}\\ \text{s.t.}\quad &A\theta = b &\text{(34b)}\end{aligned}$$

where A and b are defined as follows:

$$A = \begin{bmatrix}p_{C_{a1}}^\intercal\\ \vdots\\ p_{C_{ak}}^\intercal\end{bmatrix} \quad\text{and}\quad b = \begin{bmatrix}p_{C_{a1}}^\intercal\theta_n - \phi_{C_{a1}}\\ \vdots\\ p_{C_{ak}}^\intercal\theta_n - \phi_{C_{ak}}\end{bmatrix}. \tag{35}$$

By employing the Lagrangian multiplier $\lambda_E \geq 0$, the corresponding Lagrangian function of (34) can be written as:

$$\min_{\theta\in\Theta} \quad \mathcal{L}(\theta,\lambda_E) = \eta_n\langle g,\theta\rangle + D_h(\theta\|\theta_n) + \lambda_E^\intercal(A\theta - b). \quad (36)$$

Similarly to (31), from (36) we can obtain:

$$\langle g + A^\intercal\lambda_E^*, \theta_{n+1} - \theta_n\rangle \leq -\frac{\alpha}{\eta_n}\|\theta_{n+1} - \theta_n\|_2^2, \quad (37)$$

where λ_E^* represents the optimal value of the multiplier and can be expressed as

$$\lambda_E^* = \underset{\lambda_E\geq 0}{\operatorname{argmax}} \inf_{\theta} \mathcal{L}(\theta,\lambda_E). \quad (38)$$

Furthermore, we can show that:

$$\langle A^\intercal\lambda_E^*, \theta_{n+1} - \theta_n\rangle = \lambda_E^{*\intercal}(b - A\theta_n) \geq 0. \quad (39)$$

where we used the facts that $A\theta_{n+1} = b$ as (34b) and $A\theta_n \leq b$ as shown by (35). As a result, by leveraging the property of (39), we can further conclude from (37) that:

$$\begin{aligned}\langle g, \theta_{n+1} - \theta_n\rangle &\leq -\frac{\alpha}{\eta_n}\|\theta_{n+1} - \theta_n\|_2^2 - \langle A^\intercal\lambda_E^*, \theta_{n+1} - \theta_n\rangle \\ &\leq -\frac{\alpha}{\eta_n}\|\theta_{n+1} - \theta_n\|_2^2, \quad (40)\end{aligned}$$

which reveals (33). ■

Proposition 2. For $\eta_n l < 2\alpha$, monotonic improvement for (20) holds, i.e. $J(\theta_{n+1}) < J(\theta_n)$.

Proof. When there is no active constraint, the proof of monotonic improvement follows directly from Proposition 1. When there are active constraints, the proof can be derived by combining Lemma 1 and (32c). ■

Intuitively, such conclusion is as expected, with evidence from the unconstrained case that a smaller step increment will lead to a lower cost for the next policy. When external constraints are imposed, policy update will become more conservative and thus monotonic improvement can be admitted.

B. Policy Value Bound

In this part, we aim at providing the bound on the absolute value of the value difference between two consecutive policies to showcase the properties of the evolution of the learned policy. Specifically, we derive the bound by leveraging the error propagation analysis framework [35].

Lemma 2. The expected return of the updated policy π_{n+1} is given by[2]:

$$J(\pi_{n+1}) = \sum_s d_{\pi_n}(s)\sum_a \pi_{n+1}(a|s)c(s,a). \quad (41)$$

Proof. We first write the performance difference lemma (6) in terms of the state visitation frequency as:

$$J(\pi_{n+1}) = J(\pi_n) + \sum_s d_{\pi_{n+1}}(s)\sum_a \pi_{n+1}A_{\pi_n}(s,a). \quad (42)$$

[2]For compactness, the dependency on state and/or action is omitted when there is no ambiguity.

As pointed in [9], the second term of the right hand side can be locally approximated using the state visitation frequency induced by policy π_n, which implies that:

$$\begin{aligned}&\sum_s d_{\pi_{n+1}}(s)\sum_a \pi_{n+1}(a|s)A_\pi(s,a) \\ =&\sum_s d_{\pi_{n+1}}(s)\sum_a \pi_{n+1}(a|s)c(s,a) - J(\pi_n) \quad (43a)\\ \approx&\sum_s d_{\pi_n}(s)\sum_a \pi_{n+1}(a|s)c(s,a) - J(\pi_n). \quad (43b)\end{aligned}$$

Substituting (43b) into (42), we can obtain the expression for the value of π_{n+1} as:

$$J(\pi_{n+1}) = \sum_s d_{\pi_n}(s)\sum_a \pi_{n+1}(a|s)c(s,a), \quad (44)$$

which completes the proof. ■

Proposition 3. The upper bound of the absolute value of the value difference between two consecutive policies π_n and π_{n+1} is given by:

$$|J(\pi_{n+1}) - J(\pi_n)| \leq \frac{2\sqrt{2}C_{\max}}{1-\gamma}\sqrt{\mathop{\mathbb{E}}_{s\sim d_{\pi_n}}[D_{\mathrm{KL}}(\pi_{n+1},\pi_n)]}.$$

Proof. By resorting to the framework of the error propagation analysis, the policy value difference is upper bounded by:

$$\begin{aligned}&|J(\pi_{n+1}) - J(\pi_n)| \\ =&\frac{1}{1-\gamma}\left|\sum_s d_{\pi_n}(s)\sum_a(\pi_{n+1} - \pi_n)c(s,a)\right| \quad (45a)\\ \leq&\frac{C_{\max}}{1-\gamma}\sum_s d_{\pi_n}(s)\sum_a|\pi_{n+1} - \pi_n| \quad (45b)\\ =&\frac{2C_{\max}}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\pi_n}}[D_{\mathrm{TV}}(\pi_{n+1},\pi_n)] \quad (45c)\\ \leq&\frac{2C_{\max}}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{\pi_n}}\left[\sqrt{2D_{\mathrm{KL}}(\pi_{n+1},\pi_n)}\right] \quad (45d)\\ \leq&\frac{2\sqrt{2}C_{\max}}{1-\gamma}\sqrt{\mathop{\mathbb{E}}_{s\sim d_{\pi_n}}[D_{\mathrm{KL}}(\pi_{n+1},\pi_n)]} \quad (45e)\end{aligned}$$

wherein D_{TV} denotes the total variation between two probability distributions. Notably, the inequality (45d) is obtained due to Pinsker's inequality that bounds the total variation distance in terms of the KL divergence. The inequality (45e) is obtained following Jensen's inequality. ■

VI. EXPERIMENTS

In this section, we evaluate the effectiveness of our method by conducting ultrasound scanning experiments with a robotic manipulator and a tissue phantom. The overall experimental setup is first introduced in Sec. VI-A. Subsequently, we report the obtained results in Sec. VI-B.

A. Experimental Setup

An illustration of the hardware setup for evaluating the proposed method of safe robotic ultrasound imaging is shown in Fig. 4. The test subject involved in the experiments is a phantom model with realistic mechanical properties. The

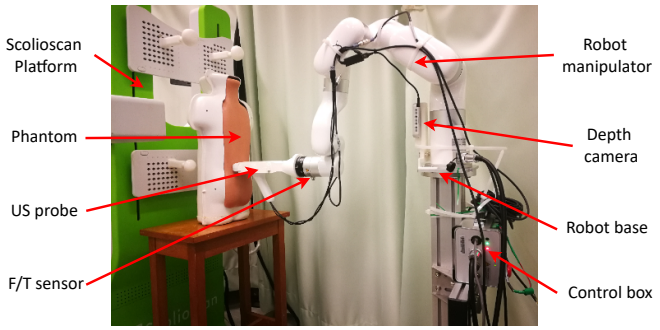


Fig. 4. Illustration of the experimental platform for scoliosis assessment.

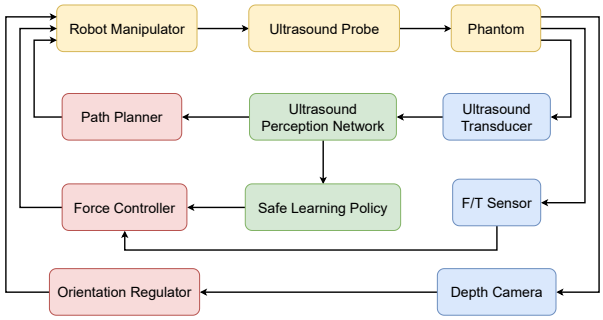


Fig. 5. Control block of ultrasound scanning for scoliosis diagnosis.

phantom has a deformed spine embedded inside its upper body. The high-fidelity phantom resembles a human upper body with mean stiffness of $195\,\mathrm{N/m}$. At the core of the setup is a six DoFs industrial robot manipulator UFACTORY xArm, which is connected via the TCP/IP protocol to the control computer. A USB ultrasound probe Sonoptek is attached to the end-effector of the robot manipulator, capturing ultrasound images at a frequency of $7.5\,\mathrm{MHz}$. To sense the interaction force between the phantom's back and the ultrasound probe, a six-axis force/torque sensor Robotiq FT300 is attached between the ultrasound probe and the robot manipulator. A depth camera RealSense is placed at the base link of the robot manipulator to have the point cloud of the phantom's back. The phantom is placed on the Scolioscan platform, which receives ultrasound images and coordinates of the probe for final 3D reconstruction of the spine [36].

The overall control system for our ultrasound scanning task is depicted in Fig. 5. The proposed robot control system is composed of three components, namely, a path planner, a force controller, and an orientation regulator. The path planner is responsible for making the ultrasound probe track the detected spinous process by sending the velocity command of the robot end-effector along the vertical direction. Detection of the spinous process is achieved by processing the ultrasound images with an ultrasound processing neural network, which uses ResNet18 as its backbone followed by three deconvolutional layers and one convolutional layer. The output of the network is a heatmap, composed of the probability of each pixel being a spinous process. The point with the maximum probability is then identified as the spinous process.

Once the spinous process is localized, the path planner commands the ultrasound probe such that the spinous process is maintained in the center of the probe's field of view. In addition to the bilateral movement command, the probe speed along the caudo-cranial direction is set to a constant value of $0.003\,\mathrm{m/s}$. The force controller is designed as a PI controller that regulates the optimal interaction force profile learned by the proposed safe reinforcement learning algorithm. The orientation regulator is responsible to adjust the probe's orientation such that the probe perpendicularly points towards the closest point on the phantom surface.

B. Experimental Results

The effectiveness of our proposed constraint-aware policy optimization is validated by safely learning the optimal interaction force profile, which will then be used for ultrasound scanning the phantom to clearly reconstruct its 3D spinal image. The corresponding learning mechanism should pertain to improving the policy parameters dictated by a cost function, which shall be properly designed to accurately reveal the learning goal. As our goal is to generate a clear 3D image of the spine, it is intuitive to design the cost function that can score the quality of 3D spinal images. Notwithstanding, the cost function for numerical evaluation of the quality of 3D spinal images is quite arduous to constitute. For the same spinal image, consensus on its quality can be hardly reached among sonographers as the evaluation can be very subjective [5]. Moreover, improving the contact force policy on the cost that evaluates the quality of the 3D spinal images could suffer from the credit assignment issue since the outcomes of the policy are delayed to the very end of 3D spinal image reconstruction.

As a workaround, our cost is designed in terms of the ultrasound image quality instead of the quality of the 3D spinal image. Traditional methods to evaluate ultrasound image quality, such as random walks, usually penalize shadow areas as they are undesirable [37]. This evaluation metric is not suitable in our case because the existence of the shadow areas is an inherent property of ultrasound imaging of bones due to the fact that ultrasound signals will be dramatically attenuated at the tissue-spine interface. Also, the random walks algorithm can be time-consuming to run. To this end, we propose to leverage the ultrasonic perception network that localizes the spinous process to evaluate the quality of the ultrasound images. More precisely, the confidence probability $P_c(t)$ of localizing the spinous process, which can be obtained in real-time, is used as a proxy of the ultrasound image quality at time step t. The immediate cost is then expressed as $c(t) = 1 - P_c(t)$.

To have better efficiency, we consider selecting several key locations containing the spinous processes on the phantom's back to separately identify the optimal contact strength rather than learning over the whole back. When finally assessing the spine conditions of the phantom, the force profile along the phantom's back is obtained by fitting the optimal force magnitude learned at different key locations. To make the selected locations representative, three key locations are selected in the

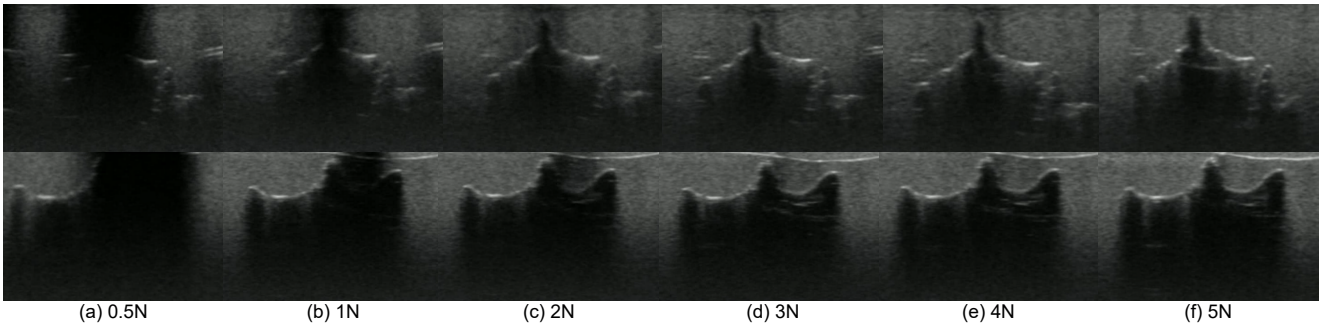


Fig. 6. Force versus ultrasound images quality for lumbar *top row* and thoracic *bottom row*

Fig. 7. Force strategy learning at different locations of the phantom's back. Three locations are located at the lumbar region (*left column*) and the other three locations are located at the thoracic region (*right column*).

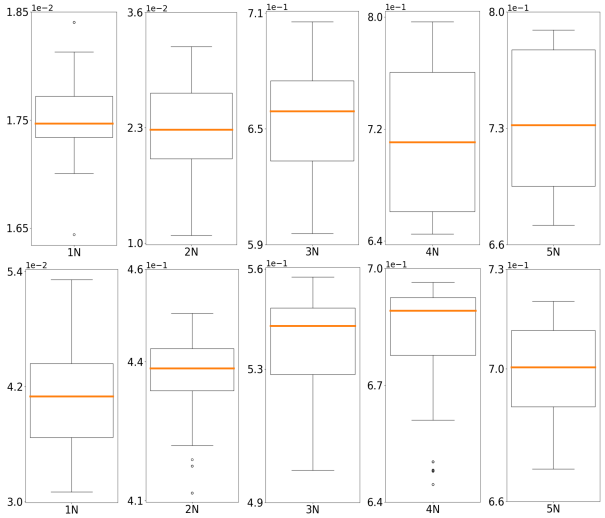


Fig. 8. Probabilities of identifying the spinous process with different force magnitudes at lumbar (*top row*) and thoracic (*bottom row*).

lumbar region and the other three locations are placed in the thoracic region, as shown in Fig. 7. All the selected locations share a similar learning procedure. The state is chosen as the received ultrasound image. The action of the policy is a force magnitude given the current state. The policy is trained to gradually improve the quality of the ultrasound image within the constraint of the force limits.

Before applying the proposed safe reinforcement learning algorithm, we first show the effects of different force magnitudes on the resulting ultrasound image quality. We exert different constant force magnitudes ranging from $0.5\,\mathrm{N}$ to $5\,\mathrm{N}$ at one lumbar spinous process and one thoracic spinous process. It can be qualitatively observed that the clarity of the ultrasound images becomes higher as the force magnitude increases, as shown in Fig. 6. Quantitatively, Fig. 8 shows the confidence probability of identifying the spinous process with the ultrasound perception neural network with respect to different force strategies. It can be concluded that the probability of finding out the spinous process becomes higher when increasing the interaction force.

For the constraint of safe reinforcement learning, we set the maximum force limit as $10\,\mathrm{N}$. The number of the trajectory samples is $\mathcal{I} = 5$. The maximum iteration steps of policy optimization are $N = 100$ and the maximum interior point iteration is $K = 10$. The imitation loss for bounding the deviation of the step for policy update is set as $\delta = 0.05$. And the step size parameters for the primal-dual update are set as $\eta_\theta = \eta_\lambda = 0.01$. The policy of interaction force is parametrized with a five-layer fully-connected neural network with the ReLU activation. The starting policy is initialized as a constant of $0.5\,\mathrm{N}$. The learning results are shown in Fig. 9. It can be observed that the vanilla reinforcement learning algorithm violates the specified force limits in some cases. While the force limit is well respected using the proposed constraint-aware policy optimization, thus the prescribed safety constraint is respected.

With the learned optimal force magnitudes at different locations, we obtain the optimal force profile by fitting these values along the phantom's back for vanilla reinforcement learning and constraint-aware policy optimization, respectively. Then we use both fitted force profiles to scan the phantom's back. The ultrasound probe begins from the sacrum region and slides the phantom's back in a caudo-cranial direction. The measured force evolution versus the phantom's back is shown in Fig. 10. Finally, the 3D spinal images are reconstructed using the volume projection imaging method for scoliosis assessment [38]. The volume projection imaging algorithm

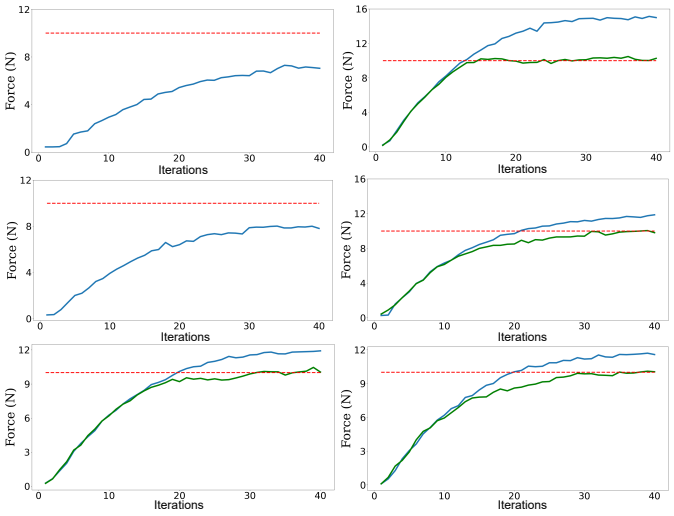


Fig. 9. Force evolution versus iterations at different locations of the phantom's back where blue curves denote the results from vanilla reinforcement learning, green curves denote the results from constraint-aware policy optimization, and red curves denote the maximum force limits.

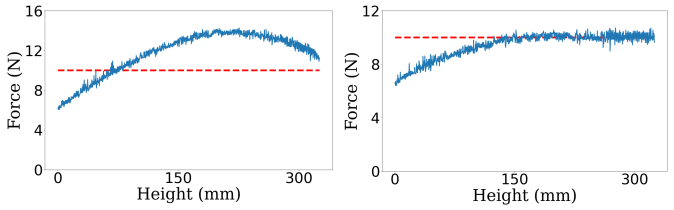


Fig. 10. Applying the optimal force profiles on the phantom's back, which are learned from vanilla reinforcement learning (*left*) and constraint-aware policy optimization (*right*), respectively.

visualizes spine anatomy by slicing the collected ultrasound images together with the corresponding 3D spatial information. Fig. 11 shows the evolution of the reconstructed spines along the vanilla reinforcement learning procedure. Fig. 12 shows the evolution of the reconstructed spines along the constraint-aware policy optimization learning procedure. It can be seen that the trend of clarity for obtained spines is gradually improved for both cases. Qualitatively, the transverse processes are becoming more and more identifiable and the black spots due to contact loss are becoming less and less prominent [5]. Also, the quality of the final spine image constructed by safe learning is comparable to the one constructed by vanilla reinforcement learning, which is sufficient to be used for scoliosis assessment.

VII. Conclusion

In this paper, we presented constraint-aware policy optimization to safely learn the optimal force profile when ultrasound scanning spines for scoliosis assessment. The effectiveness of the proposed algorithm is verified with real experiments of finding out the optimal force profile for 3D spinal reconstruction by ultrasound scanning a phantom. It is observed that the vanilla reinforcement learning algorithm exceeds the pre-specified threshold during policy search while our proposed approach can respect the force limit and thus guarantee safety.

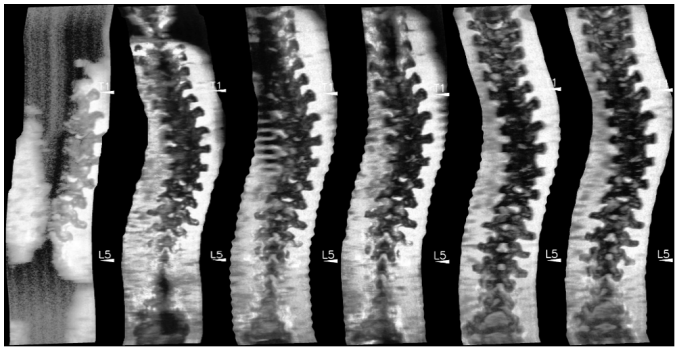


Fig. 11. Evolution of 3D reconstruction image of the phantom spine with vanilla reinforcement learning.

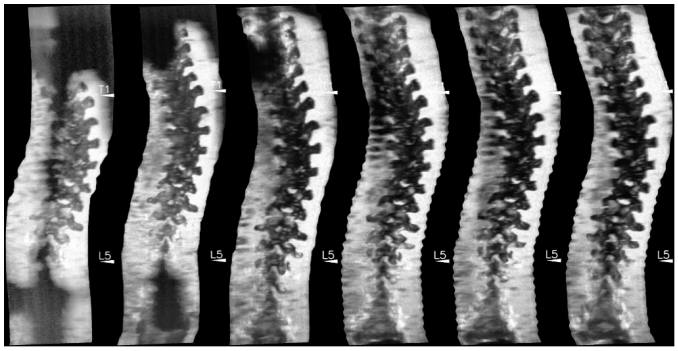


Fig. 12. Evolution of 3D reconstruction image of the phantom spine with constraint-aware policy optimization.

There are certainly a few limitations associated with our work. For example, the basis for our reward design shall be further investigated. In our current setting, we use the probability of the confidence of the spinous process as a proxy for quantifying the ultrasound image quality. As an indirect indicator is utilized to evaluate the quality of the whole ultrasound image. Another limitation in our framework lies in the lack of automatic spine region segmentation. With this information available, the robot could possess more flexible control policies for different regions.

Regarding future work, we would like to develop a force observer such that robotic ultrasound scanning can be conducted without F/T sensors by estimating the interaction force between the probe and the patient's back [39]. Furthermore, to make the interaction skill generalizable to real human testers, we would like to investigate the relationship between the interaction skill with respect to relevant features, e.g. body mass index, skin elasticity, muscle distribution, etc.

References

[1] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, "A survey of robots in healthcare," *Technologies*, vol. 9, no. 1, p. 8, 2021.

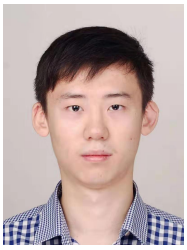
[2] C. Huang, H. Huang, J. Zhang, P. Hang, Z. Hu, and C. Lv, "Human-machine cooperative trajectory planning and tracking for safe automated driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 050–12 063, 2021.

[3] C. Esterwood and L. P. Robert, "A systematic review of human and robot personality in health care human-robot interaction," *Frontiers in Robotics and AI*, p. 306, 2021.

[4] Z. Lu, M. Li, A. Annamalai, and C. Yang, "Recent advances in robot-assisted echography: combining perception, control and cognition," *Cognitive Computation and Systems*, vol. 2, no. 3, pp. 85–92, 2020.
[5] Y.-P. Zheng, T. T.-Y. Lee, K. K.-L. Lai, B. H.-K. Yip, G.-Q. Zhou, W.-W. Jiang, J. C.-W. Cheung, M.-S. Wong, B. K.-W. Ng, J. C.-Y. Cheng *et al.*, "A reliability and validity study for scolioscan: a radiation-free scoliosis assessment system using 3D ultrasound imaging," *Scoliosis and spinal disorders*, vol. 11, no. 1, pp. 1–15, 2016.
[6] A. Duan, M. Victorova, J. Zhao, Y. Sun, Y. Zheng, and D. Navarro-Alarcon, "Ultrasound-guided assistive robots for scoliosis assessment with optimization-based control and variable impedance," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8106–8113, 2022.
[7] Z. Zhou, Y. Guo, and Y. Wang, "Handheld ultrasound video high-quality reconstruction using a low-rank representation multipathway generative adversarial network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 575–588, 2021.
[8] M. Tirindelli, M. Victorova, J. Esteban, S. T. Kim, D. Navarro-Alarcon, Y. P. Zheng, and N. Navab, "Force-ultrasound fusion: Bringing spine robotic-us to the next "level"," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5661–5668, 2020.
[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.
[10] K. Li, A. Li, Y. Xu, H. Xiong, and M. Q.-H. Meng, "RL-TEE: Autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, pp. 1–13, 2023.
[11] X. Deng, Y. Chen, F. Chen, and M. Li, "Learning robotic ultrasound scanning skills via human demonstrations and guided explorations," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2021, pp. 372–378.
[12] G. Peng, C. L. P. Chen, and C. Yang, "Robust admittance control of optimized robot-environment interaction using reference adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.
[13] Z. Jiang, M. Grimm, M. Zhou, Y. Hu, J. Esteban, and N. Navab, "Automatic force-based probe positioning for precise robotic ultrasound acquisition," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11 200–11 211, 2020.
[14] J. Tan, B. Li, Y. Li, B. Li, X. Chen, J. Wu, B. Luo, Y. Leng, Y. Rong, and C. Fu, "A flexible and fully autonomous breast ultrasound scanning system," *IEEE Transactions on Automation Science and Engineering*, pp. 1–14, 2022.
[15] H. Huang, Y. Guo, G. Yang, J. Chu, X. Chen, Z. Li, and C. Yang, "Robust passivity-based dynamical systems for compliant motion adaptation," *IEEE/ASME Transactions on Mechatronics*, pp. 1–10, 2022.
[16] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.
[17] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.
[18] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A Lyapunov-based approach to safe reinforcement learning," *Advances in neural information processing systems*, vol. 31, 2018.
[19] H. Ma, J. Chen, S. Eben, Z. Lin, Y. Guan, Y. Ren, and S. Zheng, "Model-based constrained reinforcement learning using generalized control barrier function," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4552–4559.
[20] C.-A. Cheng, X. Yan, N. Wagener, and B. Boots, "Fast policy learning through imitation and reinforcemen," in *Proceedings of the 34th Conference on Uncertanty in Artificial Intelligence*, 2018, pp. 845–855.
[21] A. Duan, R. Camoriano, D. Ferigo, D. Calandriello, L. Rosasco, and D. Pucci, "Constrained DMPs for feasible skill learning on humanoid robots," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 1–6.
[22] Z. Lu, N. Wang, and C. Yang, "A constrained DMPs framework for robot skills learning and generalization from human demonstrations," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 3265–3275, 2021.
[23] M. Davoodi, A. Iqbal, J. M. Cloud, W. J. Beksi, and N. R. Gans, "Rule-based safe probabilistic movement primitive control via control barrier functions," *IEEE Transactions on Automation Science and Engineering*, pp. 1–15, 2022.
[24] Y. Huang and D. G. Caldwell, "A linearly constrained nonparametric framework for imitation learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4400–4406.
[25] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.
[26] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2002.
[27] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999, vol. 7.
[28] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.
[29] D. A. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," *Neural computation*, vol. 3, no. 1, pp. 88–97, 1991.
[30] J. Peters, K. Mulling, and Y. Altun, "Relative entropy policy search," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
[31] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.
[32] Y. Hu, G. Chen, Z. Li, and A. Knoll, "Robot policy improvement with natural evolution strategies for stable nonlinear dynamical system," *IEEE Transactions on Cybernetics*, pp. 1–13, 2022.
[33] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of computational and applied mathematics*, vol. 124, no. 1-2, pp. 281–302, 2000.
[34] J. N. Lee, M. Laskey, A. K. Tanwani, A. Aswani, and K. Goldberg, "Dynamic regret convergence analysis and an adaptive regularization algorithm for on-policy robot imitation learning," *The International Journal of Robotics Research*, vol. 40, no. 10-11, pp. 1284–1305, 2021.
[35] T. Xu, Z. Li, and Y. Yu, "Error bounds of imitating policies and environments for reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[36] K. K.-L. Lai, T. T.-Y. Lee, M. K.-S. Lee, J. C.-H. Hui, and Y.-P. Zheng, "Validation of scolioscan air-portable radiation-free three-dimensional ultrasound imaging assessment system for scoliosis," *Sensors*, vol. 21, no. 8, p. 2858, 2021.
[37] A. Karamalis, W. Wein, T. Klein, and N. Navab, "Ultrasound confidence maps using random walks," *Medical image analysis*, vol. 16, no. 6, pp. 1101–1112, 2012.
[38] C.-W. J. Cheung, G.-Q. Zhou, S.-Y. Law, T.-M. Mak, K.-L. Lai, and Y.-P. Zheng, "Ultrasound volume projection imaging for assessment of scoliosis," *IEEE transactions on medical imaging*, vol. 34, no. 8, pp. 1760–1768, 2015.
[39] W. He, Y. Sun, Z. Yan, C. Yang, Z. Li, and O. Kaynak, "Disturbance observer-based neural network control of cooperative multiple manipulators with input saturation," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1735–1746, 2019.

Anqing Duan received his Ph.D. degree in Bioengineering and Robotics from the Italian Institute of Technology and the University of Genoa, Italy, in 2021. He is currently Project Associate with The Hong Kong Polytechnic University. His research interest lies in robotic ultrasound imaging.

Chenguang Yang (M'10-SM'16) received the Ph.D. degree in control engineering from the National University of Singapore, Singapore, in 2010, and postdoctoral training in human robotics from the Imperial College London, London, U.K. He was awarded UK EPSRC UKRI Innovation Fellowship and individual EU Marie Curie International Incoming Fellowship. As the lead author, he won the IEEE Transactions on Robotics Best Paper Award (2012) and IEEE Transactions on Neural Networks and Learning Systems Outstanding Paper Award (2022). He is the Corresponding Co-Chair of IEEE Technical Committee on Collaborative Automation for Flexible Manufacturing (CAFM), and a Fellow of British Computer Society. His research interest lies in human robot interaction and intelligent system design.



Jingyuan Zhao received B.Eng. degree in Mechanical Engineering from The Hong Kong Polytechnic University of Hong Kong (PolyU), Kowloon, Hong Kong, in 2021. He is currently a M.Sc student at Nanyang Technological University, Singapore. His research interest includes perceptual robotics and dynamic consensus algorithm of multi-agent systems.

Shengzeng Huo received the B.S. degree in vehicle engineering from the South China University of Technology, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree in the Department of Mechanical Engineering from The Hong Kong Polytechnic University, Hong Kong. His research interests include bimanual manipulation, deformable object manipulation, and robot learning.

Peng Zhou received the M.Sc. degree in software engineering from Tongji University, Shanghai, China, in 2017. Since 2019, he is pursuing his Ph.D. degree in robotics at The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. In 2021, he was a visiting Ph.D. student at Robotics, Perception and Learning Lab, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include deformable object manipulation, robot learning, and interactive perception.

Wanyu Ma received the B.S. and MA.Eng degrees in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2016 and 2018, respectively. She is currently working toward the Ph.D. degree in robotics with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. Her research interests include robotics, visual servoing, and deformable objects manipulation.

Yongping Zheng (Senior Member, IEEE) received the B.Sc. and M.Eng. degrees in electronics and information engineering from the University of Science and Technology of China, and the Ph.D. degree in biomedical engineering from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 1997.

After a postdoctoral fellowship at the University of Windsor, Canada, he joined PolyU as an Assistant Professor and was promoted to Professor, in 2008, and Chair Professor, in 2019. In July 2017, he was appointed as the Henry G. Leong Professor in biomedical engineering. He is currently the Director of the Jockey Club Smart Ageing Hub, and the Research Institute for Smart Ageing, PolyU. His main research interests include biomedical ultrasound instrumentation, soft tissue elasticity measurement and imaging, 3D ultrasound imaging, ultrasound assessment of musculoskeletal tissues, ultrasound image and signal processing, and smart aging technologies.

Prof. Zheng is a fellow of The Hong Kong Institution of Engineers (HK), a Secretary of the World Association of Chinese Biomedical Engineers, from 2017 to 2019, the Past Chair of the Biomedical Engineering Division, HKIE, and an Honorary Advisor of the Hong Kong Medical and Healthcare Device Industry Association (HMHDIA). He serves as the President for the Guangdong Hong Kong Macau Chapter of the International Society of Gerontechnology.

David Navarro-Alarcon (Senior Member, IEEE) received the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong, in 2014. Since 2017, he has been with The Hong Kong Polytechnic University, where he is currently an Assistant Professor with the Department of Mechanical Engineering. His current research interests include perceptual robotics and control systems. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON ROBOTICS.