Robotics: Science and Systems 2012
Sydney, NSW, Australia, July 09-13, 2012

Tendon-Driven Variable Impedance Control Using Reinforcement Learning

Eric Rombokas, Mark Malhotra, Evangelos Theodorou, Emanuel Todorov, and Yoky Matsuoka

***Abstract*—Biological motor control is capable of learning complex movements containing contact transitions and unknown force requirements while adapting the impedance of the system. In this work, we seek to achieve robotic mimicry of this compliance, employing stiffness only when it is necessary for task completion. We use path integral reinforcement learning which has been successfully applied on torque-driven systems to learn episodic tasks without using explicit models. Applying this method to tendon-driven systems is challenging because of the increase in dimensionality, the intrinsic nonlinearities of such systems, and the increased effect of external dynamics on the lighter tendon-driven end effectors.**

We demonstrate the simultaneous learning of feedback gains and desired tendon trajectories in a dynamically complex sliding-switch task with a tendon-driven robotic hand. The learned controls look noisy but nonetheless result in smooth and expert task performance. We show discovery of dynamic strategies not explored in a demonstration, and that the learned strategy is useful for understanding difficult-to-model plant characteristics.

I. Introduction

Though significant progress has been made toward controlling tendon-driven robotic systems, learning control of such systems remains a challenging task. Most of the difficulties arise from the existence of strong nonlinearities imposed by the tendon-hood structure, model uncertainty, and task complexity due to the need for simultaneous movement and force control.

Additionally, fine object manipulation using tendon-driven systems may require sudden changes in gains and/or tendon excursions. Smooth controls can be used when the mass of a manipulated object is small compared to that of the manipulator, but as the dynamics of the external object become significant, dextrous contact with the object requires more abrupt changes in control. Tendon-driven systems allow lightweight manipulators which accentuate this problem. Previous work has learned manipulation tasks by learning parameters of smooth trajectory basis functions [19] [3], but the imposed control smoothness limits the dynamic interaction that can occur with the manipulated object.

Optimal control provides a principled approach to determining control policies that are compliant and dynamic [2], and can also handle contact dynamics [27], but requires identified models of the robot and task. Instead, we use state-of-the-art model-free reinforcement learning to directly learn policies for movement control without having to explicitly model uncertain robot and task dynamics. The challenges that come with manipulators using tendon actuation constitute an important and previously unmet challenge for model-free reinforcement learning.

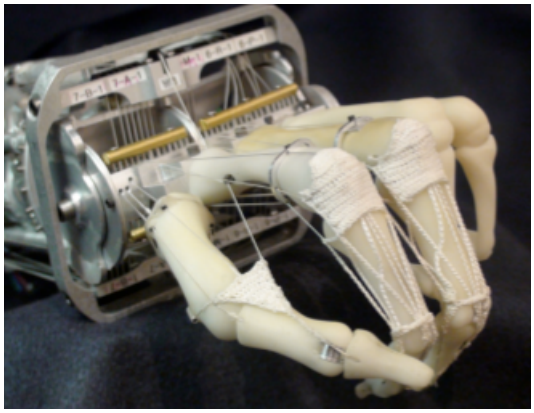

Fig. 1: The ACT hand is a tendon-driven robot designed to mimic the tendons and joints of the human hand.

In this work we use **P**olicy **I**mprovement with **P**ath **I**ntegrals (Section IV) [24, 22] for learning complex manipulation tasks with a tendon-driven biomimetic robotic hand (Section II). PI^2 concentrates sampling of the state space around a rough initial demonstration or previously learned strategy, making it effective in high dimensional problems. We introduce a structure by which PI^2 can learn a discontinuous variable impedance control policy that enables tasks requiring contact, motion, and force control during object interaction. With respect to previous results on variable stiffness control with reinforcement learning [3], here we are not using any policy parameterizations that are based on function approximation. Instead, we represent trajectories and control gains as markov diffusion processes. This choice expands the dimensionality of the controllable space and allows for better exploration of the nonlinear dynamics of the ACT hand and the task.

The learned strategies can yield insight into subtleties of the plant, showing how biomimetic robotics can not only use inspiration from nature to achieve robotic goals, but can provide insights into the systems which they mimic (Section VII-D).

Videos of the experiments can be found at http://tendondriven.pbworks.com/.

II. Tendon-Driven Biomimetic Hand

The robotic hand mimics the interaction among muscle excursions and joint movements produced by the bone and tendon geometries of the human hand, as in [6]. The index finger has the full 4 degree-of-freedom joint mobility and is controlled by six motor-driven tendons acting through a

crocheted tendon-hood. Two tendons, the FDS and FDP act as flexors; the EI, RI, and PI act as extensors and ab/aductors; the LUM is an abductor but switches from extensor to flexor depending on finger posture. By sharing the redundancies and nonlinearities of human hands [5], the system constitutes a challenging testbed for model identification, control, and task learning, while also providing a unique perspective for the study of biomechanics and human motor control.

The experiments presented here use only the index finger, with a silicon rubber skin on the palmar surface of the distal segment. The brushless DC motors that actuate the 6 tendons are torque-controlled at 200 Hz and measure tendon displacements at a resolution 2.30 μm; the tendon lengths alone are used for feedback control as there is no direct measurement of joint kinematics. Successfully performing manipulation tasks thus requires a control policy that can handle the nonlinear dynamics and high dimensionality of the robot as well as the dynamics of the task itself.

III. Sliding Switch Task

The kinematically simple task of sliding a switch is difficult to perform expertly with a tendon-driven finger; contact and task dynamics constitute a large part of the force required from the controlling tendons. An important research topic in neuromuscular control is how humans achieve such hybrid control, transitioning from motion to force control, as in tapping a finger on a rigid surface [29]. Even for isometric tasks it is nontrivial to decode the recorded activations of muscles and understand how these act through tendons to the end effector [28]. In this paper we examine the task of contacting a sliding switch and pushing it down (see Figure 3). The switch in our apparatus is coupled to a belt and motor which allow the imposition of synthetic dynamics. The position of the switch x is measured with a linear potentiometer. Importantly, the finger loses contact with the switch at x_{reach} before reaching the bottom of the possible range, denoted x_{min}.

We begin with a single demonstration of the desired task in which a human holds the finger and moves it through a motion of pushing the switch down. The tendon excursions produced by this externally-powered example grossly resemble those required for the robot to complete the task, but simply replaying them using a general-purpose PID controller would not result in successful task completion for two main reasons. Firstly, during demonstration the tendons are not loaded, which changes the configuration of the tendon network in comparison to when it is actively moving. Secondly, and more importantly, the tendon trajectories encountered during a demonstration do not impart any information about the necessary forces required to accommodate the dynamics of the task. For instance, at the beginning of the task, the finger must transition from moving through air freely, to contacting and pushing the switch. A feedback controller following a reference trajectory has no way of anticipating this contact transition, and therefore will fail to initially strike the switch with enough force to produce the desired motion.

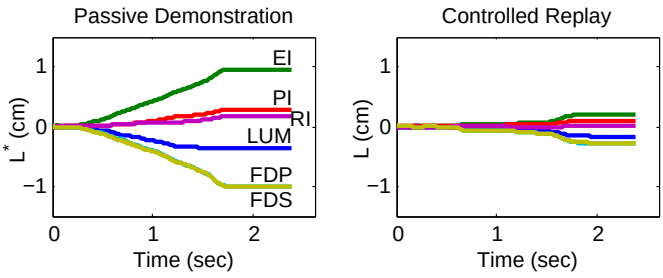

Fig. 2: Playback of the demonstration trajectories using a constant-gain proportional controller fails to achieve task-performing tendon trajectories, and does not exhibit compliance for task-irrelevant time periods. Acronyms are anatomical tendon names, eg Flexor Digitorum Profundis (FDP).

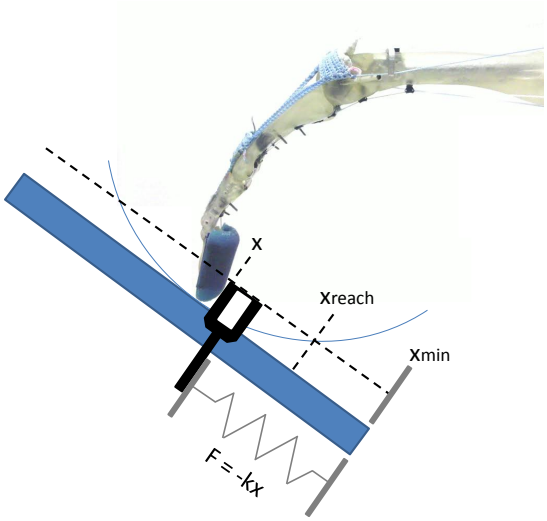

Fig. 3: Hybrid Switch Task. During the third experiment, spring dynamics are added to the natural switch dynamics, in the form of a virtual spring force F. X_{min} is the physical extent of the sliding switch, but the finger loses contact before reaching it, at a point X_{reach} which is dependent on finger movement.

Without any prior system identification of the robot or sliding switch, PI^2 can directly learn a control policy that minimizes a cost associated with the position of the switch, feedback gain, and tendon travel.

IV. Policy Improvement with Path Integrals: PI^2

In this section we review the framework of reinforcement learning based on path integrals. The stochastic optimal control is a constrained optimization problem formulated as follows:

$$V(\mathbf{x}) = \min_{\mathbf{u}(\mathbf{x},t)} J(\mathbf{x},\mathbf{u}) = \min_{\mathbf{u}} \int_t^{t_N} \boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t)dt \qquad (1)$$

subject to the nonlinear stochastic dynamics:

$$d\mathbf{x} = \boldsymbol{\alpha}(\mathbf{x})dt + \mathbf{C}(\mathbf{x})\mathbf{u}dt + \mathbf{B}(\mathbf{x})\delta\boldsymbol{\omega} \tag{2}$$

with $\mathbf{x} \in \Re^{n\times 1}$ denoting the state of the system, $\mathbf{u} \in \Re^{p\times 1}$ the control vector and $\delta\boldsymbol{\omega} \in \Re^{p\times 1}$ brownian noise. The function $\boldsymbol{\alpha}(\mathbf{x}) \in \Re^{n\times 1}$ is the drift, which can be a nonlinear function of the state $\mathbf{x}$. The matrix $\mathbf{C}(\mathbf{x}) \in \Re^{n\times p}$ is the control transition matrix and $\mathbf{B}(\mathbf{x}) \in \Re^{n\times p}$ is the diffusion matrix.

TABLE I: Policy Improvements with path integrals PI^2.

- **Given**:
  - An immediate state dependent cost function $q(\mathbf{x}_t)$
  - The control weight $\mathbf{R} \propto \boldsymbol{\Sigma}^{-1}$
- **Repeat** until convergence of the trajectory cost:
  - Create K roll-outs of the system from the same start state $\mathbf{x}_0$ using stochastic parameters $\mathbf{u} + \delta\mathbf{u}_s$ at every time step
  - **For** $k = 1...K$, compute:
    - $P\left(\boldsymbol{\tau}_{i,k}\right) = \frac{e^{-\frac{1}{\lambda}S(\boldsymbol{\tau}_{i,k})}}{\sum_{k=1}^{K}[e^{-\frac{1}{\lambda}S(\boldsymbol{\tau}_{i,k})}]}$
    - $S(\boldsymbol{\tau}_i) = \phi_{t_N} + \sum_{j=i}^{N-1}\left(q_{t_j}dt + \frac{1}{2}||\mathbf{u} + \delta\mathbf{u}_s||_{\mathbf{M}}\right)$
  - **For** $i = 1...(N-1)$, compute:
    - $\delta\mathbf{u}(\mathbf{x}_{t_i}) = \sum_{k=1}^{K} P\left(\boldsymbol{\tau}_{i,k}\right)\ \delta\mathbf{u}_s(t_i,k)$
  - **Update** $\mathbf{u} \leftarrow \max\left(\mathbf{u}_{\min}, \min\left(\mathbf{u} + \delta\mathbf{u}, \mathbf{u}_{\max}\right)\right)$

Under the optimal controls $\mathbf{u}^*$ the cost function is equal to the value function $V(\mathbf{x})$. $\boldsymbol{\mathcal{L}}$(**x**,**u**,t), the immediate cost, is expressed as:

$$\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t) = q(\mathbf{x},t) + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} \tag{3}$$

The running cost $\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t)$ has two terms: the first $q_0(\mathbf{x}_t,t)$ is an arbitrary state-dependent cost, and the second is the control cost with weight $\mathbf{R} > 0$. The optimal controls $\mathbf{u}^*(\mathbf{x},t)$ as a function of the cost to go: $V(\mathbf{x},t)$ as follows.

$$\mathbf{u}^*(\mathbf{x},t) = -\mathbf{R}^{-1}\mathbf{C}(\mathbf{x})^T\nabla_{\mathbf{x}}V(\mathbf{x},t) \tag{4}$$

The role of optimal control is to drive the system towards parts of the state space with small values of $V(\mathbf{x},t)$. The value function $V(\mathbf{x},t)$ satisfies the Hamilton Jacobi Bellman equation partial differential equation [7, 20]. Recent work on path integral work in [24, 11] and logarithmic transformations of the value function $V(\mathbf{x},t) = -\frac{1}{\lambda}\log\Psi(\mathbf{x},t)$ are exploited to transform the HJB into a linear PDEs for which their solution [8, 13] is represented via forward sampling of diffusions processes. The outcome is the expression that follows:

$$V(\mathbf{x},t_i) = -\frac{1}{\lambda}\log\int P_{path}\ e^{\frac{-\left(\phi_{t_N}+\sum_{j=i}^{N-1}q_{t_j}dt\right)}{\lambda}}d\mathbf{x}_N \tag{5}$$

where the probability $P_{path} = P(\mathbf{x}_N,t_N|\mathbf{x}_i,t_i)$ has the form of path integral starting from the initial state $\mathbf{x}_i$ and ending in state $\mathbf{x}_N$ under uncontrolled $\mathbf{u} = 0$ stochastic dynamics in (2). In [24], [12] it has been shown that under the solution of the value function in (5) the path integral optimal control takes the form:

$$\mathbf{u}_{PI}(\mathbf{x}_{t_i})dt = \lim_{dt\to 0}\int P\left(\boldsymbol{\tau}_i\right)\delta\boldsymbol{\omega}_{t_i} \tag{6}$$

where $\boldsymbol{\tau}_i$ is a trajectory in state space starting from $\mathbf{x}_{t_i}$ and ending in $\mathbf{x}_{t_N}$, so $\boldsymbol{\tau}_i = (\mathbf{x}_{t_i},...,\mathbf{x}_{t_N})$. The probability $P\left(\boldsymbol{\tau}_i\right)$ is defined as

$$P\left(\boldsymbol{\tau}_i\right) = \frac{e^{-\frac{1}{\lambda}\tilde{S}(\boldsymbol{\tau}_i)}}{\int e^{-\frac{1}{\lambda}\tilde{S}(\boldsymbol{\tau}_i)}d\boldsymbol{\tau}_i} \tag{7}$$

The term $\tilde{S}(\boldsymbol{\tau}_i)$ is defined as:

$$\begin{aligned}\tilde{S}(\boldsymbol{\tau}_i) \propto \phi(\mathbf{x}_{t_N}) + \int_{t_i}^{t_N} q(\mathbf{x}_{t_j})\delta t \\ + \int_{t_i}^{t_N}\left\|\frac{\mathbf{x}_c(t_j+\delta t) - \mathbf{x}_c(t_j)}{\delta t} - \boldsymbol{\alpha}_c(\mathbf{x}(t_j))\right\|^2_{\boldsymbol{\Sigma}_{t_j}^{-1}}\delta t\end{aligned} \tag{8}$$

Which in discrete time is approximated by:

$$S(\boldsymbol{\tau}_i) \propto \phi(\mathbf{x}_{t_N}) + \sum_{j=i}^{N-1}\left(q(\mathbf{x}_{t_j})dt + \frac{1}{2}\delta\boldsymbol{\omega}_{t_j}^T\mathbf{M}\delta\boldsymbol{\omega}_{t_j}\right) \tag{9}$$

with $\mathbf{M} = \mathbf{B}(\mathbf{x})^T\left(\mathbf{C}(\mathbf{x})\mathbf{R}^{-1}\mathbf{C}(\mathbf{x})^T\right)^{-1}\mathbf{B}(\mathbf{x})$. Essentially the optimal control is an average of variations $\delta\omega$ weighted by their probabilities. This probability is inversely proportional to the path cost according to (7) and (9). Low-cost paths have high probability and *vice versa*. Table I illustrates the path integral control in iterative form as applied to the ACT Hand. Alternative iterative formulations based on Girsanov's theorem [13] resulting in small differences in biasing terms of the path cost $\tilde{S}$ are available in [23].

V. PI^2 Learning Variable Tendon Impedance

A. Variable Impedance

If interaction forces with the environment were neglible, a standard approach would be to apply negative feedback control with as aggressive gains as possible while still maintaining stabilty. High feedback gains, however, are dangerous for robots interacting with humans and are potentially wasteful when perturbations are not task-relevant. The idea of impedance control is to control the dynamic behavior of the manipulator in addition to commanding a reference state trajectory [10] [3] [4]. Variable stiffness impedance control specifies a schedule of gains emphasizing uncertain or unforgiving components of the task, while allowing compliance elsewhere.

For complex tasks and unknown environmental dynamics, a suitable target impedance schedule is difficult to specify *a priori*. For biological and some biomimetic systems, stiffness is achieved by coactivation of antagonist muscles, introducing futher challenge for impedance scheduling due to complex actuator dynamics [9] [16]. Here we define the basic structure of our controller, then describe how learning is applied.

Consider motor commands τ calculated via a proportional controller:

$$\tau_i = -k_i(l_i - \hat{l}_i) \tag{10}$$

where k_i is the proportional gain and $\hat{l}_i$ is the desired reference length for tendon i. By varying k in time according to a gain schedule, we may achieve variable impedance specific to each tendon. The reference trajectories may differ from the demonstration to anticipate task dynamics.

B. Learning Variable Impedance Using PI^2

We apply PI^2 to learning controls with no time-averaging of the learned parameter vector. We use a time-varying impedance controller with a differential equation on the gains and reference lengths:

$$dk(t) = (-\alpha k(t) + u_k(t))\, dt + \sigma_k \delta u_k \tag{11}$$

$$d\hat{l}(t) = \left(-\alpha \hat{l}(t) + u_{\hat{l}}(t)\right) dt + \sigma_{\hat{l}} \delta u_{\hat{l}} \tag{12}$$

where u_k and $u_{\hat{l}}$ are the change in gains and reference trajectories. δu_k and $\delta u_{\hat{l}}$ are sampled from a zero-mean Normal distribution as in the PI^2 algorithm, Table I.

The smoothing effect of these differential equations provides a means for using anything from unsmoothed ($\alpha \to \infty$) to highly smoothed ($\alpha = 0$) controls. In either case, no time averaging acts on the level of the learning algorithm [19]. The experiments presented here use $\alpha = \frac{0.9}{dt} = 180$, providing a very slight smoothing effect to the control outputs. Section VII presents learned controls which, although quite noisy, produce smooth tendon trajectories with contact-produced discontinuities, as illustrated in Figure 5.

The controls vector $\mathbf{u} = u(t) + \delta u(t)$ optimized by the learning algorithm (see Table I) corresponds to a concatenation of gain and reference length controls: $\mathbf{u} = \left(u_k(t)\, u_{\hat{l}}(t)\right)^T$. PI^2 learns the optimal gain and reference length for each tendon at each timestep, a considerable number of parameters. To combat this high-dimensionality without adversely effecting the learning, we subsample the injected noise δu to change each 50 timesteps during a rollout. Such windowing is especially important for systems like ours whose high-order dynamics filter out much of the injected noise.

Learning proceeds through iterative revision of the policy parameters. Each of these revisions we refer to as a *trial*. A sample trajectory is queried from the system by sampling δu, and actually performing a switch-slide using the resulting $u + \delta u$. We refer to one of these exploratory executions of the task as a *rollout*. To revise u at the end of a trial, each sampled control strategy is weighted according to the cost encountered by the corresponding rollout (Table I). The results reported here use $\sigma_k = 50$ for sampling gains and $\sigma_{\hat{l}} = .05$ for sampling reference trajectories. The smaller this exploration variance is, the more similar rollouts are, so the magnitute of σ should depend on the natural stochasticity of the plant, though here it is set by hand. Convergence is qualitatively insensitive to the exact value of σ, and high variance results in low control cost (Section IV) and therefore increased control authority. Each trial consists of ten rollouts, and after every third trial performance is evaluated by executing three *exploration-free rollouts* ($\sigma = 0$).

The controls learned for each tendon are linked to the others only through their shared effect on the system reflected in the rollout costs. In this way, each tendon-specific PI^2 can be considered a parallel agent operating on its own.

VI. Cost Function and Experiments

A. Cost Function

We conduct three experiments using the same setup and demonstration. For all experiments, the cost-to-go function for a rollout having duration T may be expressed as :

$$C_t = q_{terminal}(x_T) + \sum_t^T q(x_t) + \mathbf{u_t}^T \mathbf{R} \mathbf{u_t} \tag{13}$$

Here x_t corresponds to the position of the switch at time t. $q(x_t)$ is the cost weighting on the switch state, with $q_{terminal}(x_T)$ the terminal cost at the end of the rollout. $\mathbf{R}$ is the cost weighting for controls. The experiments presented here use $q(x_t) = 20000x_t, q_{terminal}(x_t) = 300q(x_t)$, meaning the terminal state cost is as costly as 300 time steps.

B. Gain Scheduling: Experiment One

For the first experiment PI^2 learns only the gains K. Therefore the sampling noise $\sigma_{\hat{l}} = 0$. This means that the learned strategy will always kinematically resemble the demonstration, but will optimize compliant control of the dynamics.

C. Gains and Reference Lengths: Experiment Two

The second experiment is to learn both gains and reference lengths simultaneously. The learned strategy may take advantage of kinematic poses not experienced in the demonstration.

D. Stabilizing Spring Dynamics: Experiment Three

In the third experiment PI^2 learns both gains and reference lengths, and we also introduce spring behavior to the switch dynamics, requiring a qualitatively different learned optimal strategy. The intrinsic switch dynamics are retained, but the motor coupled to the switch also resists the finger with a stiffness force $f_{spring} = k(x_{max} - x)$. This force springs the switch back up toward its home position if the finger pushes too far and loses contact. For the first two experiments, the optimal strategy could include pushing past the point of losing contact with the switch, but for the third the finger must proceed quickly to the edge of losing contact but go no further.

VII. Results

The task is successfully learned for each experiment. Figure 6 depicts the sum cost-to-go results as learning proceeds for 100 trials (revisions of control) for two separate executions of each experiment. Every third trial, three unexploratory rollouts are performed, and their means and standard deviation are denoted by the solid line and shaded width in the figure. Cost

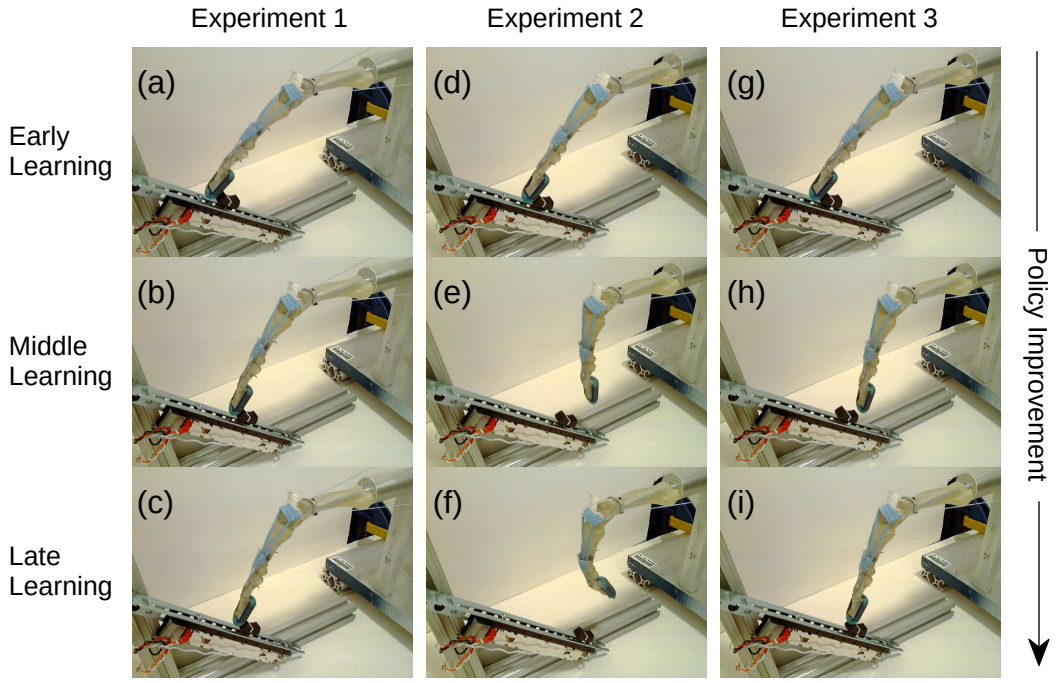

Fig. 4: Final posture attained for early, middle, and late learning (rows) for each experiment (columns). In pane (f) we observe the switch being thrown past the last point of finger contact to its physical limit. Pane (h) depicts a strategy which has pushed the finger too far, and the switch has bounced upward due to the added spring force (experiment three, Section V). In pane (i) we see the very tip of the finger just preventing the switch from returning.

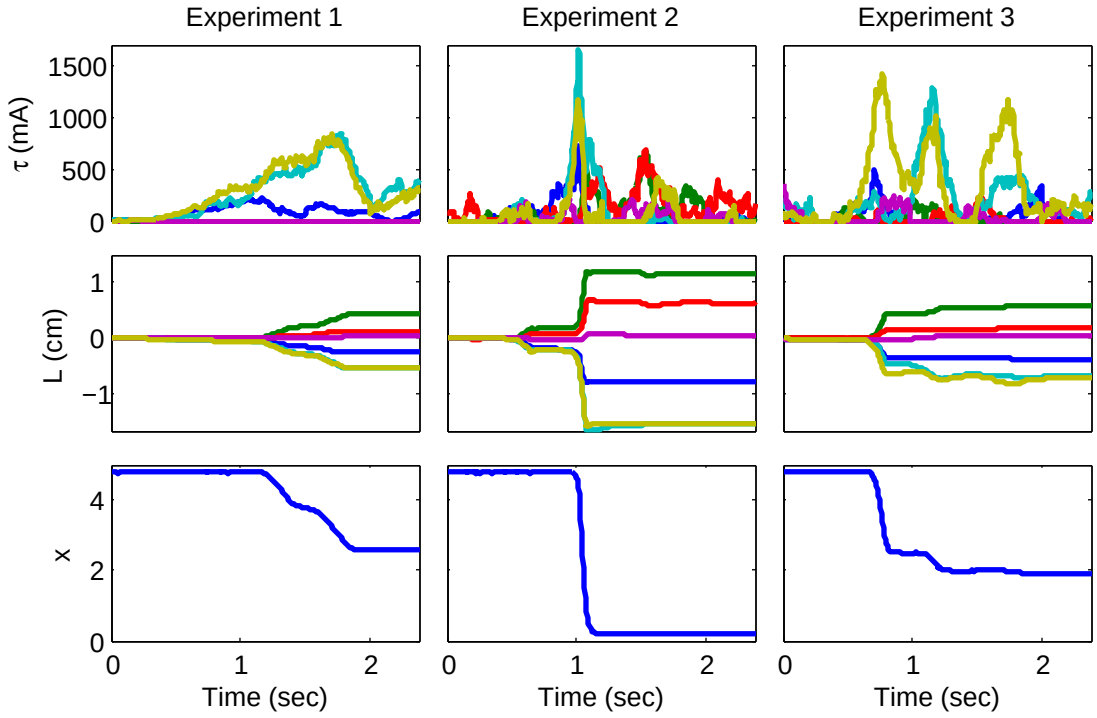

Fig. 5: Results after learning (trial 100) for all three experiments. The first row depicts the torque commands learned for all tendons. Dynamic requirements of the tasks are apparent in the learned strategies. For instance, Experiment 2 appears to require two bursts of torque to achieve vigourous flexion, but Experiment 3 requires another in order to arrest finger motion before losing contact. The middle row depicts the resulting tendon trajectories, and the third row depicts the position of the switch.

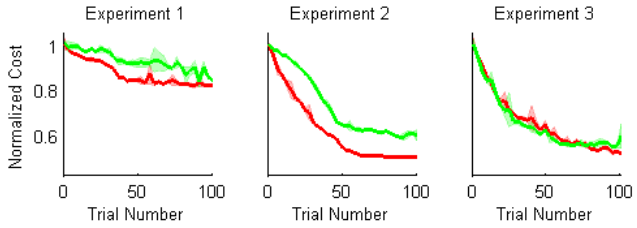

Fig. 6: Learning as trials progress. In experiment one (left), only the gains are learned and the reference lengths recorded during the demonstration are used. In experiment two (middle), both the gains and reference lengths of the tendons are learned. In experiment three (right), additional spring dynamics resist switch displacement from initial position. Solid line indicates mean cost for three unexploratory rollouts every third trial; the shaded area is standard deviation. Each of the two curves corresponds to a repetition of the experiment beginning with the same initial control strategy.

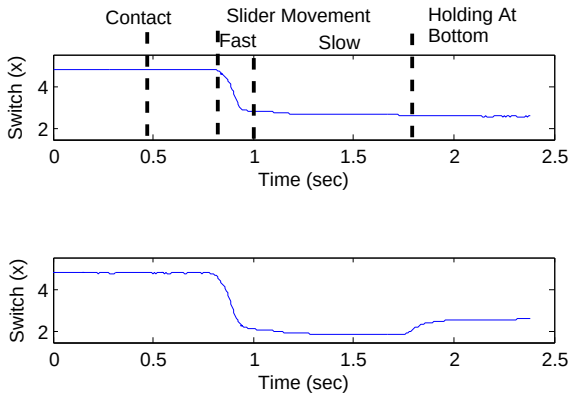

Fig. 7: Above: Switch position learned for Experiment 3, in which additional spring dynamics resist switch displacement from initial position. Below: Switch position for a middle-learning failure to stop before leaving contact with the switch.

magnitude is different for each experiment due to different control exploration (Equation 13), so sum cost-to-go for the first noise-free rollout is normalized for comparison.

Figure 5 shows the commanded torques, resulting tendon trajectories, and switch position trajectories for the final learned control strategies for each experiment.

A. Gain Scheduling: Experiment One Results

In Experiment 1, the optimized gain schedule produces a movement which smoothly pushes the switch to near the position attained in the demonstration. As can be seen in Figure 5, top left pane, the learned strategy uses the two primary flexors very similarly. After 100 trials, the total cost is reduced by about 15% from the initial policy.

B. Gains and Reference Lengths: Experiment Two Results

The strategy found for Experiment 2, learning both the gains and references, differed between the two executions of the experiment. Both learned strategies quickly move the switch down, past the kinematic pose observed in the demonstration. During the second execution, however, PI^2 learned that the switch could be thrown down, using the inertia of the switch to carry it past the last point of finger contact, x_{reach}. The result of this highly dynamic behavior can be seen in Figure 4, pane (f). The switch is thrown to its physical limit x_{min} in a single ballistic motion, as evidenced in the switch position trajectory pictured in Figure 5, middle-bottom.

C. Stabilizing Spring Dynamics: Experiment Three Results

Experiment 3 presents the most dynamically challenging task. In this experiment, the motor resisting switch displacement implicitly imposes a task constraint: the finger must move quickly to push the switch as before, but now it must stop quickly to prevent overshoot. As in the other experiments, Figure 5, bottom-right presents the switch trajectory for the learned strategy. The learned strategy makes a large

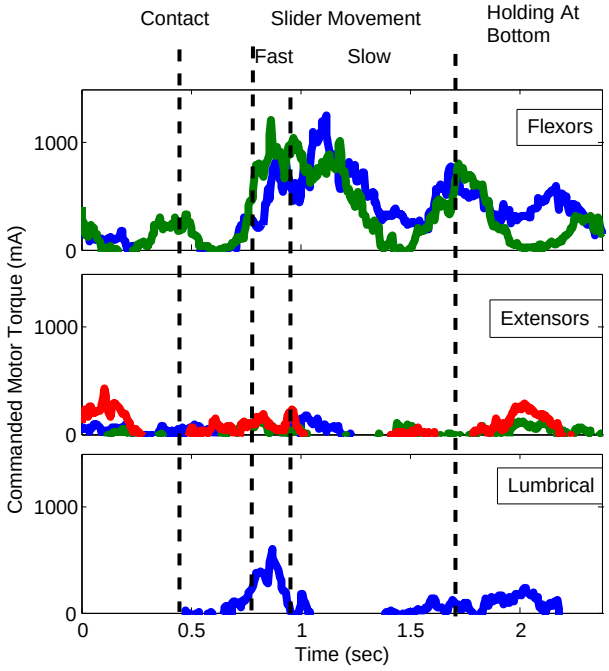

Fig. 8: Torques resulting from the learned control strategy for experiment three, in which additional spring dynamics resist switch displacement from initial position.

initial switch movement followed by a burst of flexion and then co-contraction of flexors and extensors. In order to better illustrate the effect of the additional spring dynamics, Figure 7 presents a suitably stopping (top) and failed stop (bottom) switch trajectory. Recall that the control policy incorporates feedback of tendon lengths only; measurement of the switch position is used in evaluating a rollout but not within the control loop.

D. The Role of the Lumbrical Tendon

Experiment 3 yields insight into the role of the Lumbrical tendon. In that experiment the finger must move quickly but then stabilize the switch at the extent of the finger's reach, corresponding to switch position x_{reach} in Figure 3. This requires more elaborate use of the extensor tendons than in the first two experiments, in which extensors primarily stabilize the abduction-adduction motion of the finger.

Examining the strategy learned by the controller for this tendon serves as a novel way of illuminating the complex role it plays. The tendon extensor mechanism matches the human tendon network in important ways [30], including finger lumbricals. Lumbricals are smaller tendons which attach to other tendons, both flexors and extensors, instead of bone. This produces a moment arm relationship to joint torques which is complex and highly pose-dependent. Roughly, it is a flexor of the metacarpophalangeal (MCP) joint, the "knuckle", while being an extensor of the two distal joints. With the hand outstretched, thumb up, the lumbrical would be used to push the fingertip as far as possible from the palm, straightening the finger while flexing the knucle. This is exactly the motion necessary for the finger to contact the switch for the longest amount of time during downstroke.

By examining the control strategy learned for the lumbrical tendon in concert with the more straightforward tendons, we observe its protean nature. As illustrated in Figure 8, it acts as a flexor at the beginning of the motion, with a clear peak of torque being delivered in concert with the FDP/FDS flexors at the moment of greatest switch movement. However, after motion has ceased we observe cocontraction of all tendons, but the pattern of activation of Lumbrical more closely matches the torque profile of the extensors.

VIII. Related Work and Discussion

A. Optimal Control for Tendon-Driven Systems

Optimal control provides a principled approach for robots to expertly perform tasks without relying on pre-programmed or stereotyped motions. Recent successes in applying algorithms like iLQG [17, 15, 18] suggest that difficult control tasks, like dexterous manipulation, may be better achieved by formulating models and cost functions than by laboriously hand-tuning reference trajectories or feedback gains directly. Notably, optimal control can solve for time-varying torque and stiffness profiles to achieve dynamic tasks compliantly[2].

In contrast however with [17, 15, 2] we do not use state space dynamical models. PI^2 is a derivative-free method in the sense that it does not require linearization of dynamics and quadratic approximations of cost functions on state space trajectories. More importantly the optimal control in our approach is an average over sampled controls based on how well they performed on the real system and thus no differentiation of the value functions is performed. With respect to previous work on variable stiffness control with reinforcement learning [4, 3], our approach does not use function approximators to represent gains or desired trajectories. This lack of policy parameterization allows us to learn non-smooth trajectories and control gains required for tasks that involve contact, which is particularly important when controlling tendon-actuated manipulators. Tendon-driven systems are particularly advantageous in applications like dexterous hands where it is desirable to have compact and low inertia end effectors [14]. The associated increase in control dimensionality and nonlinearity, as well as the relatively greater prominence of task dynamics make system identification of the task and the robot less tractable. Model-based control methods for systems reflecting the biomechanical complexities of hands are very much an open research topic due mostly to the interaction of tendons in the network.

This paper shows that by using a model-free reinforcement learning algorithm like PI^2, we can bridge the gap between optimal control and tendon-driven systems: we simultaneously enjoy the benefit of choosing cost functions instead of reference trajectories or gains while also circumventing the need for an accurate model of the robot and environment. Future work will investigate the generalizability of learned policies and the possibility of using this algorithm also as a data collection mechanism for system identification.

The absence of models for the underlying hand and task dynamics as well as the lack of policy parameterization does not come without cost. As demonstrated in this paper, our learning approach convergences to an optimal solution for all experiments, nevertheless this may require the execution of many rollouts on the real system. This is not a surprising observation since it is related to the exploitation-exploration trade-off in reinforcement learning literature [21] in correspondence to how much information of the underlying dynamics is initially provided to the learning algorithm. Clearly, once a model is known, its use could speed up learning. Future work will explore the use of Stochastic Differential Dynamic Programming (SDDP) [25] or iLQG on contact-less models and integrate the resulting control policies with PI^2 to learn tasks with contact and motion to force control transitions.

B. Discovery of System Phenomena

The second execution of Experiment 2 discovered that the switch could be thrown beyond x_{reach}, and the learned behavior for Experiment 3 involved cocontraction of antagonists, as seen in biological tendon actuation, without explicit instruction. By interacting directly with the environment, the robotic system learned to perform dynamic behavior never encountered in the demonstration, exploiting subtle phenomena without system ID. The advantages of this kind of embodied learning have been explored in a variety of experiments, for example learning circuit configurations [26] or robot control and morphology [1]. Manipulators incorporating complex actuation and eventually sensory capabilities could benefit from learning from task-relevant experience as opposed to expert knowledge.

Similarly, analysis of the effects of varying aspects of complex systems like tendon networks can be difficult due to dimensionality, model mismatch, and nonlinear phenomena.

As described in Section VII-D, the learned policy produces torques which validate the biomimetic anatomical properties of the tendon hood extensor mechanism. The approach outlined in these experiments is valid for a variety of complex systems which might otherwise be difficult to measure or simulate. By observing that the learned usage the Lumbrical tendon produces dynamic consequences similar to those known for humans, we now can more confidently assert the biomimicry of that aspect of the tendon network.

REFERENCES

[1] J. Bongard. The utility of evolving simulated robot morphology increases with task complexity for object manipulation. *Artificial Life*, 16(3):201–223, 2010.

[2] D.J. Braun, M. Howard, and S. Vijayakumar. Exploiting variable stiffness in explosive movement tasks. In *Robotics: Science and Systems Conference*, 2011.

[3] J. Buchli, E. Theodorou, F. Stulp, and S. Schaal. Variable impedance control - a reinforcement learning approach. In *Robotics: Science and Systems Conference*, 2010.

[4] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal. Learning variable impedance control. *International Journal of Robotics Research*, 30(7):820, 2011.

[5] A.D. Deshpande, J. Ko, D. Fox, and Y. Matsuoka. Anatomically correct testbed hand control: muscle and joint control strategies. In *IEEE International Conference on Robotics and Automation*, pages 4416–4422, 2009.

[6] A.D. Deshpande, Z. Xu, M. J. V. Weghe, L. Y. Chang, B. H. Brown, D. D. Wilkinson, S. M. Bidic, and Y. Matsuoka. Mechanisms of anatomically correct testbed (ACT) hand. *IEEE Trans. Mechatronics*, to appear, 2012.

[7] P. Dorato, V. Cerone, and C. Abdallah. *Linear Quadratic Control: An Introduction*. Krieger Publishing Co., Inc., 2000.

[8] A. Friedman. *Stochastic Differential Equations And Applications*. Academic Press, 1975.

[9] N. Hogan. Adaptive control of mechanical impedance by coactivation of antagonist muscles. *IEEE Transactions on Automatic Control*, 29(8):681–690, 1984.

[10] N. Hogan. Impedance control: An approach to manipulation. *Journal of Dynamic Systems, Measurement, and Control*, 107(2):17, 1985.

[11] H. J. Kappen. Linear theory for control of nonlinear stochastic systems. *Phys Rev Lett*, 95:200–201, 2005.

[12] H. J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 11:P11011, 2005.

[13] I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, August 1991.

[14] H. Kobayashi, K. Hyodo, and D. Ogane. On tendon-driven robotic mechanisms with redundant tendons. *The International Journal of Robotics Research*, 17(5):561–571, 1998.

[15] P. Kulchenko and E. Todorov. First-exit model predictive control of fast discontinuous dynamics: Application to ball bouncing. In *IEEE International Conference on Robotics and Automation*, pages 2144–2151, 2011.

[16] S.A. Migliore, E.A. Brown, and S.P. DeWeerth. Biologically inspired joint stiffness control. In *IEEE International Conference on Robotics and Automation*, pages 4508–4513. IEEE, 2005.

[17] D. Mitrovic, S. Nagashima, S. Klanke, T. Matsubara, and S. Vijayakumar. Optimal feedback control for anthropomorphic manipulators. In *IEEE International Conference on Robotics and Automation*, pages 4143–4150.

[18] D. Mitrovic, S. Klanke, and S. Vijayakumar. Learning impedance control of antagonistic systems based on stochastic optimization principles. *International Journal of Robotics Research*, 30(5):556–573, 2010.

[19] E. Rombokas, E. Theodorou, M. Malhotra, E. Todorov, and Y. Matsuoka. Tendon-driven control of biomechanical and robotic systems: A path integral reinforcement learning approach. 2012.

[20] R.F. Stengel. *Optimal Control and Estimation*. Dover Publications, New York, 1994.

[21] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning : An introduction*. Adaptive computation and machine learning. MIT Press, Cambridge, 1998.

[22] E. Theodorou. *Iterative Path Integral Stochastic Optimal Control: Theory and Applications to Motor Control*. PhD thesis, University of Southern California, 2011.

[23] E. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connection to path integral and kl control. *Submitted*, 2012.

[24] E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral approach to reinforcement learning. *Journal of Machine Learning Research*, (11):3137–3181, 2010.

[25] E. Theodorou, Y. Tassa, and E. Todorov. Stochastic differential dynamic programming. In *American Control Conference*, 2010.

[26] A. Thompson. An evolved circuit, intrinsic in silicon, entwined with physics. *Evolvable Systems: From Biology to Hardware*, pages 390–405, 1997.

[27] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control.

[28] F.J. Valero-Cuevas, F.E. Zajac, C.G. Burgar, et al. Large index-fingertip forces are produced by subject-independent patterns of muscle excitation. *Journal of Biomechanics*, 31(8):693–704, 1998.

[29] M. Venkadesan and F.J. Valero-Cuevas. Neural control of motion-to-force transitions with the fingertip. *Journal of Neuroscience*, 28(6):1366–1373, 2008.

[30] D.D. Wilkinson, M.V. Weghe, and Y. Matsuoka. An extensor mechanism for an anatomical robotic hand. In *IEEE International Conference on Robotics and Automation*, 2003.