Inverse articulated-body dynamics from video via variational sequential Monte Carlo

Dan Biderman[1] **Christian A. Naesseth**[1] **Luhuan Wu**[1] **Taiga Abe**[1]

Alice C. Mosberger[1] **Leslie J. Sibener**[1] **Rui Costa**[1]

James Murray[2] **John P. Cunningham**[1]

[1] Columbia University, New York, USA
{db3236, ca2733, lw2827, ta2507, acm2246
ljs2203, rc3031, jpc2181}@columbia.edu
[2] University of Oregon, Oregon, USA
jmurray9@uoregon.edu

Abstract

Convolutional neural networks for pose estimation are continuously improving in identifying joints of moving agents from video. However, state-of-the-art algorithms offer no insight into the underlying mechanics of articulated limbs. "Seeing" the mechanics of movement is of major importance for fields like neuroscience, studying how the brain controls movement, and engineering, e.g., using vision to correct for errors in the action of a robotic manipulator. In the pipeline proposed here, we use a convolutional network to track joint positions, and embed these as the joints of a linked robotic manipulator. We develop a probabilistic physical model whose states specify second-order rigid-body dynamics and the torques applied to each actuator. Observations are generated by mapping the joint angles through the forward kinematics function to Cartesian coordinates. For nonlinear state estimation and parameter learning, we build on variational Sequential Monte Carlo (SMC), a differentiable variant of the classical SMC method leveraging variational inference. We extend with a distributed nested SMC algorithm, which, at inference time, wraps multiple independent SMC samplers within an outer-level importance sampler. We extract mechanical quantities from simulated data and newly acquired videos of mice and humans, offering a novel tool for studying e.g. biological motor control.

1 Introduction

Understanding the motion of 3D bodies by sight is a fundamental task for most biological and artificial agents. While pose-estimation algorithms can now accurately estimate an articulated body's position, to robustly predict movement and understand how the brain controls it, one must still infer the underlying Newtonian mechanics. Here we propose a reliable tool for making such inferences for an articulated body from a sequence of images.

Our method, illustrated in Figure 1, starts from detecting joint positions via a convolutional network. For a multi-view setup, we reconstruct the joints' 3D positions using a Bundle Adjustment algorithm [1]. We model the (noisy) joint positions with a hidden Markov model (HMM), whose dynamics

Workshop on Differentiable Vision, Graphics, and Physics in Machine Learning at NeurIPS 2020.

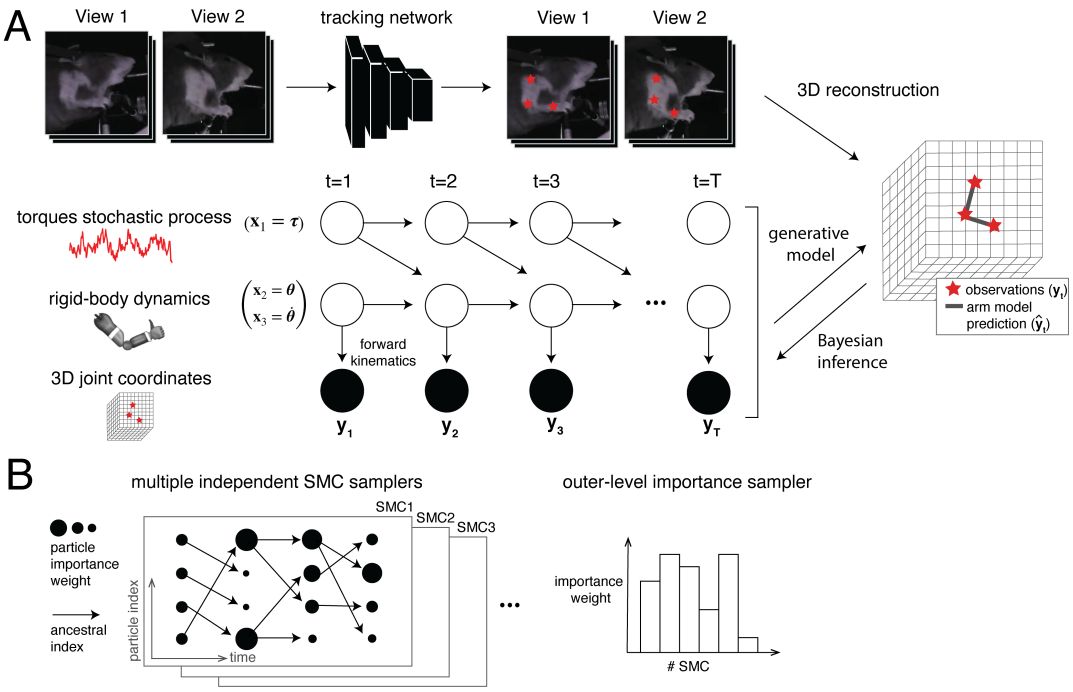

Figure 1: A) Our tracking pipeline and probabilistic model. B) Inference with multiple independent SMC samplers (left), embedded in an outer-level importance sampler (right).

include the nonlinear equations of motion and an appended stochastic torque process. The observation model nonlinearly transforms joint angles into Cartesian coordinates.

The nonlinear dynamics and observation models involved lead to an intractable Bayesian inference problem, calling for approximate solutions [3]. Furthermore, since we assume no knowledge of the manipulator's parameters such as link lengths and radii, and since we wish to allow for an arbitrarily complex torque process, we need to accommodate large-scale parameter learning. Therefore, we build on variational SMC [12, 7, 8], a differentiable variational inference method tailored to nonlinear state-estimation and parameter learning. We improve the fidelity of torque inference with a nested SMC scheme that wraps multiple independent SMC samplers within an outer-level importance sampler. Our model can be interpreted as a stochastic differentiable physics engine, and we offer a method to perform robust Bayesian inference over its full state.

One closely related line of work infers the Newtonian mechanics of non-articulated bodies from video using approximate Bayesian inference while relying on convolutional networks for tracking or segmentation [19, 21, 17]. We differ by concentrating on external torques and by using an alternative inference strategy. For articulated bodies, state-of-the-art human pose-estimation algorithms fit a geometric 3D human body model to single images [4] or use recurrent neural networks or optical flow [5] to model videos. These models neither consider rigid-body dynamics nor perform Bayesian state inference. Another relevant line of work uses deep imitation learning to reconstruct the motion of robotic manipulators and humanoids by observing their joints' state [20], which is estimated from motion capture [10] or video [15]. We differ from imitation learning by performing Bayesian inference over the state and external torques and learning the body model's parameters.

2 Probabilistic rigid-body mechanics model

We define a rigid-body including K joints with a known geometry, and learnable parameters ϕ including link shape parameters and masses. We describe the configuration of the rigid-body using joint angles as generalized coordinates [18]. According to the angular version of Newton's second law, a.k.a "forward dynamics", $\ddot{\boldsymbol{\theta}} = \boldsymbol{D}(\boldsymbol{\theta})^{-1}(\boldsymbol{\tau} - \boldsymbol{c}(\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}))$, where $\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}, \ddot{\boldsymbol{\theta}} \in \mathbb{R}^K$ are vectors representing the joint angles, angular velocities and angular acceleration, and $\boldsymbol{\tau} \in \mathbb{R}^K$ the joint torques. $\boldsymbol{D}(\boldsymbol{\theta}) \in \mathbb{R}^{K \times K}$ is the inertia tensor, $\boldsymbol{c}(\boldsymbol{\theta}) \in \mathbb{R}^K$ the gravity load, and $\mathbf{h}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) \in \mathbb{R}^K$ the

Coriolis and centripetal forces. For specific rigid bodies, we compute the entries of $\boldsymbol{D}(\cdot), \boldsymbol{c}(\cdot), \boldsymbol{h}(\cdot)$ through the Euler-Lagrange method from classical mechanics, which is a the equivalent of Newton's laws of motion for generalized coordinates [18], using SymPy [11].

We introduce a HMM with a transition distribution defined by rigid-body dynamics and an emission distribution mapping joint angles to Cartesian coordinates. We represent the dynamics with a blocked state vector $\mathbf{x}$, such that the current state is a function of just the previous state

$$\mathbf{x} = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{pmatrix}, \quad \dot{\mathbf{x}} = \begin{pmatrix} f(\boldsymbol{\tau}) \\ \dot{\boldsymbol{\theta}} \\ \boldsymbol{D}(\boldsymbol{\theta})^{-1}(\boldsymbol{\tau} - \boldsymbol{c}(\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})) \end{pmatrix} \tag{1}$$

where the forward dynamics is emphasized in red, and the torque dynamics model $f(\boldsymbol{\tau})$ is a user choice. Here, we chose to model the torques from an Ornstein-Uhlenbeck process to ensure that they remain bounded:

$$\dot{\boldsymbol{\tau}} = -\lambda\boldsymbol{\tau} + \sigma_\tau \boldsymbol{\eta}, \quad \lambda > 0, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2}$$

An interesting generalization that we leave for future work is to let $f(\boldsymbol{\tau})$ be a Gaussian Process or a recurrent neural network. We evolve the system using Euler-Maruyama integration ($\mathbf{x}_t = \mathbf{x}_{t-1} + \delta_t \dot{\mathbf{x}}_{t-1}$) and can write a distribution over the current state given the previous state

$$\mathbf{x}_t \sim f(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t \middle| \begin{pmatrix} \boldsymbol{\tau}_{t-1} + \delta_t(-\lambda\boldsymbol{\tau}_{t-1}) \\ \boldsymbol{\theta}_{t-1} + \delta_t \dot{\boldsymbol{\theta}}_{t-1} \\ \dot{\boldsymbol{\theta}}_{t-1} + \delta_t \ddot{\boldsymbol{\theta}}_{t-1} \end{pmatrix}, \begin{pmatrix} \delta_t^2 \boldsymbol{\sigma}_\tau^2 \boldsymbol{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \epsilon \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \epsilon \boldsymbol{I} \end{pmatrix}\right), \tag{3}$$

where we introduced diagonal covariance terms ϵ to avoid a degenerate distribution, implicitly admitting that our physical model might not be a perfect description of the system. The emission model from states to observations implements the forward kinematics function $\boldsymbol{\mu}_y(\boldsymbol{\theta}, \boldsymbol{\phi})$ that computes the Cartesian coordinates of each joint given the angles and link lengths (specified in Danevit-Hartenberg notation [18])

$$\mathbf{y}_t \sim g(\mathbf{y}_t|\mathbf{x}_t) := \mathcal{N}(\mathbf{y}_t|\boldsymbol{\mu}_y(\boldsymbol{\theta}, \boldsymbol{\phi}), \sigma_y \cdot \mathbf{I}), \tag{4}$$

In summary, our HMM is defined by $p(\mathbf{x}_{0:T}, \mathbf{y}_{0:T}) = \underbrace{f(\mathbf{x}_0)}_{\text{initial}} \prod_{t=1}^{T} \underbrace{f(\mathbf{x}_t|\mathbf{x}_{t-1})}_{\text{transition}} \prod_{t=0}^{T} \underbrace{g(\mathbf{y}_t|\mathbf{x}_t)}_{\text{emission}}$, where $\mathbf{x}_{0:T} := \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T\}$ and $\mathbf{y}_{0:T} := \{\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_T\}$.

3 Inference and learning via variational sequential Monte Carlo

The time series of observations $\mathbf{y}_{0:T}$ includes the joint coordinates, obtained from a pose-estimation algorithm taking video as its input. The goal is to infer the posterior distribution over the latent (unobserved) states $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$ and maximize the log marginal likelihood $\log p(\mathbf{y}_{0:T})$ with respect to the articulated-body's constant parameters such as link lengths or masses. Due to the nonlinearities in the transition and emission distributions, computing the posterior analytically is intractable. We propose to leverage variational sequential Monte Carlo (VSMC) [13, 7, 8] for inference and learning. VSMC combines sequential Monte Carlo (SMC) [14] with variational inference [2] to approximate the posterior. SMC approximates the sequence of target distribution $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$ for $t = 0, ..., T$ using N weighted particles

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t}) \approx \widehat{p}(\mathbf{x}_{0:t}|\mathbf{y}_{0:t}) := \sum_{i=1}^{N} \frac{w_t^i}{\sum_l w_t^l} \delta_{x_{0:t}^i}. \tag{5}$$

The weighted set of particles are obtained iteratively for $t = 0, \ldots, T$ by *resampling* (stochastically choosing samples from $x_{1:t-1}^i$ according to their weights), *propagating* (generating new samples x_t^i) and *weighting* (computing w_t^i). The full SMC algorithm is summarized in the Appendix. SMC provides an unbiased estimate of the marginal likelihood $\widehat{p}(\boldsymbol{y}_{0:T}) = \prod_{t=0}^{T} \frac{1}{N} \sum_{i=1}^{N} w_t^i$.

Variational SMC [13] interprets a draw from Eq. (5) as a sample from the variational approximation to the posterior $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$. This interpretation allows learning parameters both in the probabilistic mechanics model and in the proposal distribution, thereby improving inference efficiency. Variational SMC optimizes $\mathbb{E}[\log \widehat{p}(\boldsymbol{y}_{0:T})]$, a *differentiable* lower bound on the true log marginal likelihood $\log p(\boldsymbol{y}_{0:T})$, using stochastic gradient descent. For a thorough introduction to VSMC and SMC we refer to [14, 13]. To improve torque inference, we leverage GPUs to run multiple independent SMC samplers with an outer-level importance sampler (see the Appendix for details).

4 Results

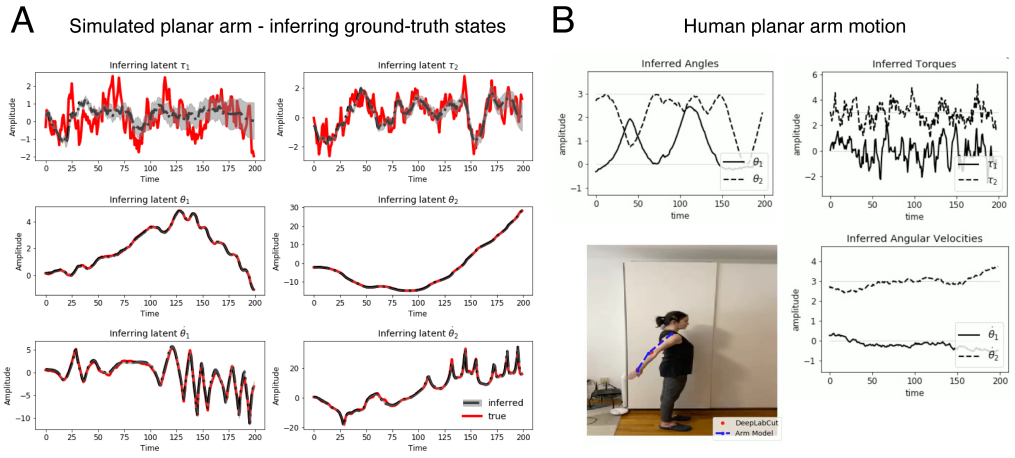

Figure 2: A) Recovering ground-truth angular dynamics and torques from noisy joint coordinates. B) Reconstructing human planar arm motion and extracting angular dynamics and torques.

Planar arm (simulated data). We simulate noisy joint coordinates of a probabilistic planar arm model, and show that we can reconstruct the arm's trajectory in this Video, while recovering the ground-truth mechanical states including the external torques, as seen in Fig.2. We initialized the unknown link lengths by the median empirical norm between joints, and further refined them in learning (masses were held fixed). Moreover, we initialized the joint angles near their maximum-likelihood value for the first frame, and found that it increased the effective sample size both within each SMC sampler and between independent SMC samplers at the outer-level importance sampling.

Real human planar arm motion. We trained a ResNet-50 in DeepLabCut [9] to track the shoulder, elbow and wrist in videos of humans, given ≈ 350 labeled frames. For a test video including a planar motion, we initialize the link lengths near the empirical median stick length, and initialize the angles with their maximum-likelihood value for the first frame. We find a match between the observed 2D joints and the our model's predictions, as seen in the Video. For each frame, we extract previously inaccessible quantities including torques, inertia tensor, and gravity load. We expect an improved inference performance once we improve our initial tracking performance.

3D arm model (simulated data). We developed a new probabilistic model for a robotic arm in 3D. In our model, each joint has two axes of rotation - yaw and pitch, so the state-vector is twice larger than in the 2D case. We successfully recover the 3D arm's state from noisy observations (see Figure 1 in the Appendix). Note that our torque inference is slightly less accurate than in the 2D case, potentially stemming from poor scaling of SMC with the state's dimension [14].

3D mouse reaching. In this experiment, a mouse is moving a planar joystick to obtain a sucrose reward, while neural activity in the motor cortex is recorded through calcium imaging. We film the mouse from two views and shave its forelimb to better expose the joints for tracking using ResNet-50. As a first analysis, we fit a 3D forward-kinematics model with random-walk angular dynamics using a forward-filtering backward-sampling approach [3], learning stick lengths. This simple dynamics model fits the observations well, as seen in this Video.

5 Conclusions

To understand biological movement, we need to explain it from first physical principles like torques applied to joints. In this work we proposed to combine pose-estimation networks with articulated-body models and approximate Bayesian inference. We plan to extend our approach to mechanically model stroke patients' arm reaching impairments [6], and to dissect the role of different neural populations within the motor cortex in mice [16].

Acknowledgements

We thank Liam Paninski for helpful comments and suggestions.

References

[1] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *European conference on computer vision*, pages 29–42. Springer, 2010.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[3] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later.

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[5] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[6] J. W. Krakauer and S. T. Carmichael. *Broken movement: the neurobiology of motor recovery after stroke*. MIT Press, 2017.

[7] T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood. Auto-encoding sequential monte carlo. *arXiv preprint arXiv:1705.10306*, 2017.

[8] C. J. Maddison, J. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6573–6583, 2017.

[9] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[10] J. Merel, Y. Tassa, D. TB, S. Srinivasan, J. Lemmon, Z. Wang, G. Wayne, and N. Heess. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*, 2017.

[11] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

[12] C. Naesseth, F. Lindsten, and T. Schön. Nested sequential Monte Carlo methods. volume 37 of *Proceedings of Machine Learning Research*, pages 1292–1301, Lille, France, 07–09 Jul 2015. PMLR.

[13] C. Naesseth, S. Linderman, R. Ranganath, and D. Blei. Variational sequential monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 968–977. PMLR, 2018.

[14] C. A. Naesseth, F. Lindsten, and T. B. Schön. Elements of sequential monte carlo. *arXiv preprint arXiv:1903.04797*, 2019.

[15] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018.

[16] G. M. Shepherd. Corticostriatal connectivity and its role in disease. *Nature Reviews Neuroscience*, 14(4):278–291, 2013.

[17] K. Smith, L. Mei, S. Yao, J. Wu, E. Spelke, J. Tenenbaum, and T. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, pages 8985–8995, 2019.

[18] M. W. Spong and M. Vidyasagar. *Robot dynamics and control*. John Wiley & Sons, 2008.

[19] T. D. Ullman, A. Stuhlmüller, N. D. Goodman, and J. B. Tenenbaum. Learning physical parameters from dynamic scenes. *Cognitive psychology*, 104:57–82, 2018.

[20] Z. Wang, J. S. Merel, S. E. Reed, N. de Freitas, G. Wayne, and N. Heess. Robust imitation of diverse behaviors. In *Advances in Neural Information Processing Systems*, pages 5320–5329, 2017.

[21] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.