INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Mishra S, van Rees WM, Mahadevan L. 2020 Coordinated crawling via reinforcement learning. *J. R. Soc. Interface* **17**: 20200198.
http://dx.doi.org/10.1098/rsif.2020.0198

Received: 23 March 2020
Accepted: 27 July 2020

Subject Category:
Life Sciences–Physics interface

Subject Areas:
computational biology, biomechanics

Keywords:
crawling, locomotion, reinforcement learning, neuromechanics

Author for correspondence:
L. Mahadevan
e-mail: lmahadev@g.harvard.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5100823.

THE ROYAL SOCIETY PUBLISHING

Coordinated crawling via reinforcement learning

Shruti Mishra[1], Wim M. van Rees[4] and L. Mahadevan[1,2,3]

[1]Paulson School of Engineering and Applied Sciences, [2]Department of Physics, and [3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[4]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

WMvR, 0000-0001-6485-4804; LM, 0000-0002-5114-0519

Rectilinear crawling locomotion is a primitive and common mode of locomotion in slender soft-bodied animals. It requires coordinated contractions that propagate along a body that interacts frictionally with its environment. We propose a simple approach to understand how this coordination arises in a neuromechanical model of a segmented, soft-bodied crawler via an iterative process that might have both biological antecedents and technological relevance. Using a simple reinforcement learning algorithm, we show that an initial all-to-all neural coupling converges to a simple nearest-neighbour neural wiring that allows the crawler to move forward using a localized wave of contraction that is qualitatively similar to what is observed in *Drosophila melanogaster* larvae and used in many biomimetic solutions. The resulting solution is a function of how we weight gait regularization in the reward, with a trade-off between speed and robustness to proprioceptive noise. Overall, our results, which embed the brain–body–environment triad in a learning scheme, have relevance for soft robotics while shedding light on the evolution and development of locomotion.

1. Introduction

The locomotion of an animal is a result of coordination of its nervous system with its body and environment. Understanding coordinated motions that involve sensory feedback and proprioception requires a theoretical framework integrating the brain, body and environment [1–3]. But how do these smooth rhythmic motions arise in the first place?

Experiments on locomotory dynamics in model systems, such as the fly larva of *Drosophila melanogaster* [4], suggest that, early in larval morphogenesis, neurons are part of a well-connected network. During development, the pruning of neuronal connections reduces the connectivity of neurons via both biochemical and biomechanical feedback modulated by behaviour and function embodied in twitching that gradually gives way to coordinated locomotion [5,6]. In the larva and more generally in many soft-bodied organisms, motion arises via rectilinear crawling [7,8], wherein rhythmic contraction and relaxation of muscles create waves that propagate either forward (prograde) or backward (retrograde) along the length of the body. This induces forward locomotion when the interaction with the substrate is asymmetric, e.g. when friction in the forward direction and that in the backward direction are very different. The asymmetry in friction has both a passive and an active component: the presence of anisotropic denticles allows the body to slide more easily in one direction than another passively, while dorso-ventral muscles can partially lift the body to modulate friction actively [4]. In either case, the result is the conversion of waves of contraction to net motion of the body, which has been studied for over a century [9].

Substantial previous experimental work characterizing *D. melanogaster* crawling has highlighted the role of sensory feedback in initiating and maintaining the gait [10] and has inspired recent theoretical work on the dynamics of a segmented, soft-bodied crawler moving on a frictional surface [11,12]. These studies have shown that minimal representations of the musculature and neural dynamics

suffice to explain a number of these experimental observations that include the onset and propagation of contractile waves that lead to locomotion, and further suggest that the rhythmic gait can arise without a central pattern generator. Here, neural impulses drive the activation of muscle forces, resulting in deformation of the body, producing biomechanical strain. Proprioceptive sensing of this strain in turn drives neural impulses, thereby closing the feedback loop. The result is that the crawler moves forward by simultaneously lifting and contracting its body segments, starting from the posterior segments and moving towards the anterior end. Critically, in these and most other studies, the neural system is assumed to have a fixed, predetermined connectivity.

Since the muscles, body wall and connective tissue in the body of a *D. melanogaster* larva develop asynchronously [10], a natural question is how these subsystems are wired together for robust performance. Indeed, could the crawler use proprioceptive feedback to learn a coordinated gait for forward crawling, i.e. rewire the neuronal connections using experience-driven sensory feedback to achieve a coordinated gait, as observed experimentally [10]? To explore this, we use the framework of reinforcement learning (RL) [13]. Originally inspired by observations of how animals learn to perform certain functions, the approach has gained significant traction recently in the context of training computers in games [14], strategies for moving through a fluid [15,16] and other domains. We frame our question in terms of the coupled dynamics of a neurophysical system for the crawler and an RL algorithm for neuronal wiring, using sensory feedback to maximize a reward associated with crawling forward.

2. Mathematical model of a crawler

Our mathematical model is chosen to roughly mimic a soft-bodied crawler, the larva of the fruitfly *D. melanogaster*, given that it has become a focus to understand the link between molecules, circuits, physiology and behaviour using a variety of approaches [17,18]. The soft-bodied cylindrical larva consists of 10 segments and is about 1 mm in length and 200 µm in diameter in the first instar stage and 4 mm in length and about 800 µm in diameter in the third instar stage. Previous work introduced a coupled neurodynamical model, with a given neural and mechanical connectivity to explain how coordinated crawling can arise even without a central pattern generator [11], and later included a more detailed neuromechanical network model in a similar framework to quantify the experimentally observed importance of proprioception [12]. This sets the stage to ask how the larva might learn to coordinate crawling, given the importance of properly wiring the neuromuscular system for robust functioning of its locomotory system.

Anatomically, the segments are connected at their boundaries (nodes) as shown in figure 1*a*. Each segment is assumed to behave like a linear viscoelastic solid, which can be actively contracted by muscles that respond to neuronal inputs as schematized in figure 1*a*. The firing of a segmental neuron, minimally modelled as a leaky integrator, causes muscular activation to deform the segment, which then moves if the forces overcome friction; simultaneously, the segment transmits forces to neighbouring segments, where neurons can be activated if the strain crosses a threshold. This leads to a propagating wave even in the absence of a central pattern

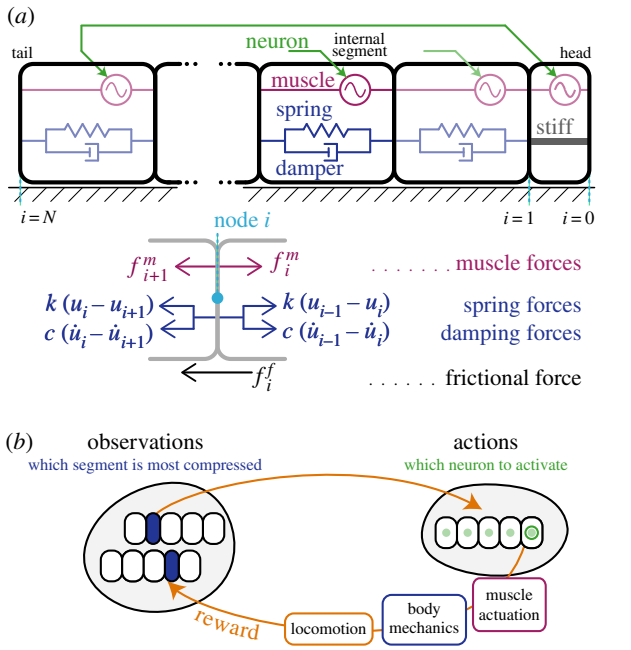

Figure 1. Schematic of the crawler. (*a*) Each segment of the soft-bodied crawler is represented by a spring-damper system and a muscle. Each muscle acts to stretch the segment and is driven by a single neuron. (*b*) Interactions between the different components of the crawler as it learns using the feedback from its environment.

generator. We now turn to quantifying the three sub-systems corresponding to the body, the brain and the environment.

2.1. Body mechanics

We assume that the passive viscoelastic properties of each segment can be described in terms of an elastic modulus E and viscosity μ. To construct a simple one-dimensional model for its mechanical behaviour, we integrate across the cross-section of the body, so that the segmented crawler resembles a set of active viscoelastic springs with stiffness $k = EA/L$ and damping constant $c = \mu A/L$, where A is the area of cross-section of a segment of length L; see figure 1*a*. The segment boundaries or nodes $i \in [0, 10]$ are mechanically characterized by their displacements $u_i(t)$, which change due to a contractile force f_i^m exerted by muscles on either side of node i and due to a frictional force f_i^f from the external environment at node i. Ignoring the role of inertia, since the animals move slowly, force balance at node $i \in [1, 9]$ in figure 1*a* implies that

$$k(u_{i+1} - 2u_i + u_{i-1}) + c(\dot{u}_{i+1} - 2\dot{u}_i + \dot{u}_{i-1}) + f_i^m - f_{i+1}^m = f_i^f. \tag{2.1}$$

The force-balance equations at the head and the tail are different from those at the internal nodes as the head and tail do not have a segment ahead of and behind them, respectively. At the head ($i = 0$)

$$k(u_1 - u_0) + c(\dot{u}_1 - \dot{u}_0) + f_0^m - f_1^m = f_0^f, \tag{2.2}$$

while at the tail ($i = N = 10$)

$$k(u_{N-1} - u_N) + c(\dot{u}_{N-1} - \dot{u}_N) + f_N^m = f_N^f. \tag{2.3}$$

To complete the formulation of the model for how the body responds, i.e. solve for $u_i(t)$, $i \in [0, 10]$, we need to specify dynamical laws for the evolution of the neural dynamics that drive the internal muscular forces f_i^m and the environmental frictional forces f_i^f.

2.2. Neuromuscular dynamics

In each segment, we assume that the neural dynamics follow a minimal first-order dynamical law known as the θ-model [19] to drive the activation of neuron i. Assuming the time scale of neuronal relaxation to be τ_θ, we may then write

$$\tau_\theta \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = 1 - \cos\theta_i + (1 + \cos\theta_i)\min[1, I_i(t)]. \tag{2.4}$$

Here $I_i(t)$ is the time-dependent input to the neuron i, and the neuron fires every time $\theta 2\pi > 0$ so that the set of spike times t_i^s for neuron i is given as

$$\{t_i^s\} = \{t |\ \mathrm{mod}(\theta_i(t) - \pi, 2\pi) = 0\}, \tag{2.5}$$

resulting in muscle actuation. The input $I_i(t)$ is then restricted to be a binary variable, with only one neuron active at a given time

$$I_i(t) = \begin{cases} 1 & \text{if } i = a, \\ 0 & \text{if } i \neq a, \end{cases} \quad a \in \{0, \ldots, N-1\}. \tag{2.6}$$

Noting that experimental observations of larval crawling show that the head and the tail move together [4], we activate the tail neuron I_N every time the head neuron is activated, i.e. when $I_0 = 1$, we set $I_N = 1$.

We note that a more sophisticated neural model with a population of excitatory and inhibitory neurons that are coupled is probably more realistic as a model for the fruitfly larva [12] but yields the same qualitative results as the simple integrate and fire model above, and so we will limit ourselves to the current simple model.

In each segment, we assume that the muscle forces vary between zero and a maximum $F^m_{\max}$. The contracting muscles respond asymmetrically to the timing of neuronal spikes with a characteristic rise time that is about half that of the exponential decay [4]. For simplicity, we use a symmetric rise and decay with a built-in temporal decay constant τ_m and a limiter to set the maximum force amplitude so that

$$\tau_m \frac{\mathrm{d}f_i^m}{\mathrm{d}t} = -f_i^m + F^m_{\max}\min[1, F_i^m(t)], \tag{2.7}$$

where

$$F_i^m(t) = \Sigma_{t^s \in \{t_i^s\}}\, \mathrm{e}^{-(t-t^s)/\tau_\theta} \tag{2.8}$$

is the sum of all the forces due to the spiking neurons. This dynamical law is consistent with our previous models [11,12]. Here, we note that it is possible for several segments to have active muscles even though only one neuron can be active at a particular time, because the muscle forces can decay much more slowly than neural activity.

2.3. Body–substrate frictional interaction

To complete the formulation of our model, we need to prescribe a frictional law for the interaction of the crawler with the environment. Experimental observations [4] show that the larva actively reduces friction in a contracting segment by lifting it off the substrate, and the presence of asymmetrically shaped denticles on the ventral surface make the passive friction asymmetric, so that forward motion experiences less friction than backward motion. While previous work [12] has included both these effects, in our one-dimensional model we do not distinguish between the passive and active components of the friction for simplicity. We further impose the condition that the friction force vanishes whenever $\dot{u}_i = 0$, leading to a smooth transition between the positive and negative values for forward and backward velocity. Then the friction force on the ith node is given by

$$f_i^f = \frac{f^f_{\max}}{2}\left[(1 + \eta_f)\tanh\left(\frac{\dot{u}_i - \dot{u}^0}{\varepsilon^f}\right) + (1 - \eta_f)\right], \tag{2.9}$$

where η_f is the ratio of the maximum frictional force in the backward to the forward directions, ϵ_f is a smoothing parameter and $\dot{u}^0$ is a constant chosen such that $f^f(0) = 0$. All together, our mathematical model equations (2.1)–(2.9) determine the gait and locomotion of the crawler: given the neural connectivity weights and an initial neural impulse leads to an input that drives equation (2.4); this drives equation (2.8) and (2.9) and thence equations (2.1)–(2.3).

2.4. Scaling and parameter choices

We scale the relevant variables in our model using the time scale of neuronal activity τ_θ, the equilibrium length of a segment L and the stiffness of a segment k. Then the dimensionless parameters corresponding to the variables presented in the mechanical model are: τ_m/τ_θ, which is the ratio of time scales for muscular and neuronal relaxation; $c\tau_\theta/k$, which is the dimensionless damping; $f^f_{\max}/kL$, which is the scaled maximum frictional force; and $F^m_{\max}/kL$, which is the scaled maximum muscular force. The specific values for these non-dimensional parameters used throughout this work, given in table 1 in appendix A, are consistent with experimental estimates for a *D. melanogaster* larva [12].

To compare our results of converged coordinated crawling shown in figure 2*a* (see also electronic supplementary material, videos S1 and S2) to the experimental observations of *D. melanogaster* larva we use our simulations to extract the scaled maximum segment deformation $\Delta u/L$, the characteristic wave speed $v\ \tau_\theta/L$, and the speed of the larva $v_{\mathrm{crawler}}\ \tau_\theta/L$. For the parameter values given in table 1 in appendix A, we find that the peak contraction of a segment is 33%, yielding a contraction speed of 0.026 waves$/\tau_\theta$ and a crawler forward speed of $0.0056L/\tau_\theta$. For a third instar larva [20], using the value of 1.5 waves per second and a length of 4 mm implies that $\tau_\theta = 17$ ms and $L = 4/10 = 0.4$ mm, respectively, so that the forward speed of the crawler is 0.13 mm s^{-1}. For a first instar larva [4], using a wave speed of 0.5–1.5 waves per second and a length of 1 mm, we get a range of τ_θ of 17–51 ms, which translates to a forward speed of 11–33 m.s^{-1}, compared with the observed range of 45–120 m.s^{-1}.

3. Reinforcement learning strategy

With the established physical model and parameter choices for the crawler, we turn to RL to determine the neural weights for efficient crawling. The framework of RL consists of an agent interacting with its environment, with the aim of achieving a goal. An agent moves through different environmental observable states by taking actions. While so, it accumulates rewards from the environment, with the goal of taking actions that maximize its long-term rewards, calculated as a discounted sum of successive rewards. This goal is achieved by learning a mapping that links an action to its current environmental observable state; this mapping is known as the agent's policy. The RL description is summarized in figure 1*b*.

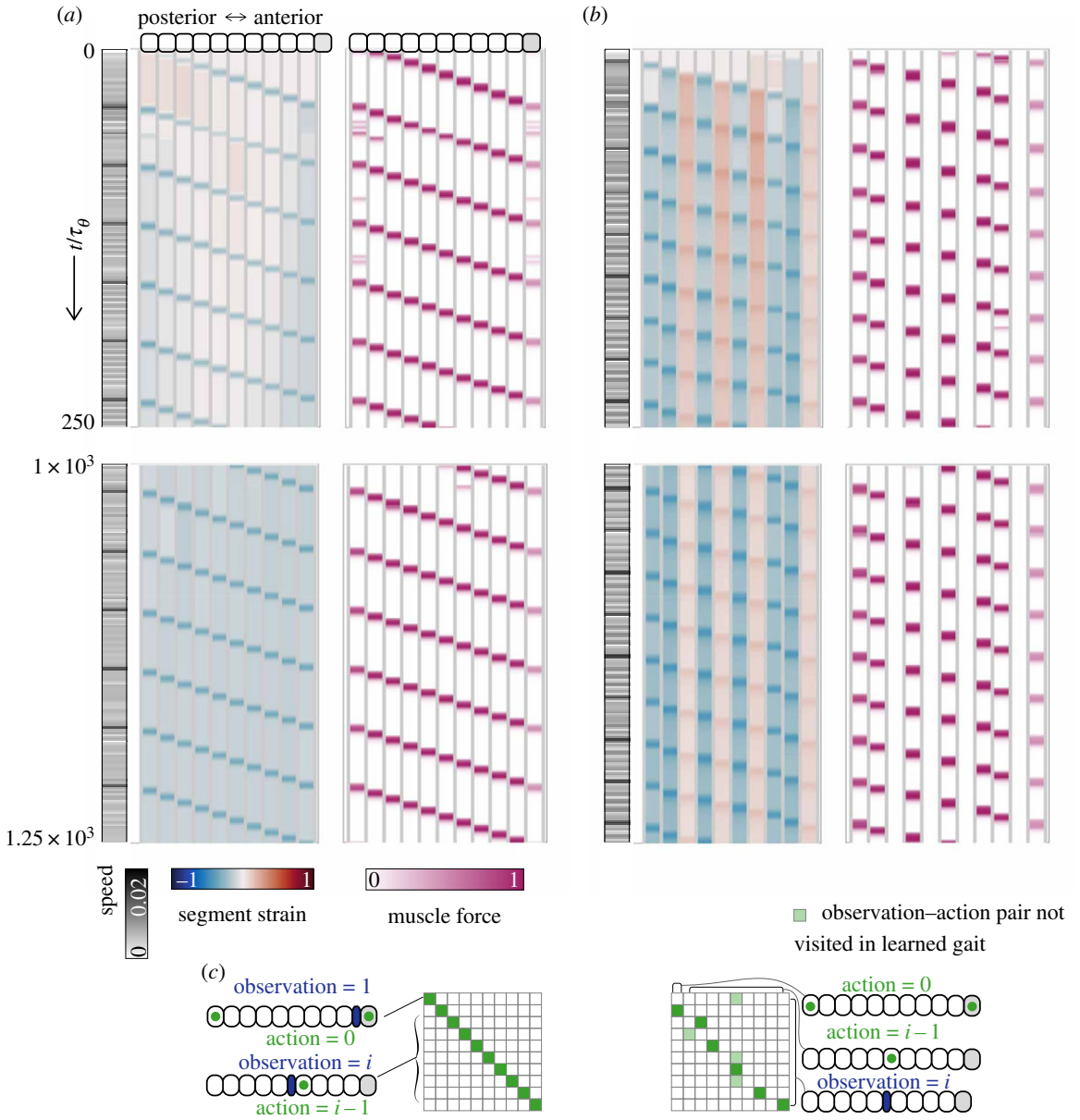

Figure 2. Learning of coordinated gaits in a neurophysical model determined using equations (2.1)–(3.3). (*a*) A regularized gait ($\epsilon = 0.01$), with the speed of the centre of mass of the crawler shown in grey, the segment strains shown in blue-red and the muscle forces in each segment shown in white-pink, and (*b*) an unregularized gait ($\epsilon = 0$). The parameter values are summarized in table 1 in appendix A. (*c*) Converged policy corresponding to the gaits in (*a*,*b*), with the dark green and light green squares corresponding to $\pi(a|o) = 1$ in the final policy and light green squares corresponding to observations which are never reached in the converged gait.

3.1. Formulation of observable state, action and reward

In our formulation, the observation of the agent is an incomplete knowledge of itself and its frictional environment. Given the established importance of proprioception [11] in locomotion, it is likely to be important in the learning process as well. A minimal approach accounting for this is via the observation o associated with the index of the segment that is most strongly contracted, since that requires knowledge of a single variable that can be easily computed via a series of pair-wise comparisons. Then

$$o = \text{argmin}_{i\in(1,\ldots,N)}(u_i - u_{i-1}). \tag{3.1}$$

The action a is the input to the θ-model that drives neuronal activity, resulting in muscle actuation, i.e. $I_i(t)$ in equation (2.4). We further restrict this by allowing the input $I_i(t)$ to have values of 0 (OFF) or 1 (ON), with only one neuron active at a given time, as described in (2.6).

Since the goal is to move forward, we set the reward r accordingly,

$$r = (\bar{u}_{t+\Delta t} - \bar{u}_t) - \epsilon r_2, \tag{3.2}$$

where $\bar{u}$ is the position of the centroid of the crawler, t denotes time, Δt is the size of the discrete time step (see table 1 in appendix A), and $r_2 = \max_i(|u_{i+1} - 2u_i + u_{i-1}|)$ is a penalty on large variations in strain along the length of the crawler, with ϵ determining the relative contributions from this strain gradient to the reward r.

We use a form of RL known as Q-learning [13], with a discrete representation for the observation and action spaces. The entries in the Q-matrix, $Q(o, a)$, represent how much cumulative reward the crawler expects to get after taking an action a after an observation o, i.e. $\sum_{k=0}^{\infty} \gamma^k r_{t+(k+1)\Delta t}$, where r_t is the reward at time t and $\gamma \in [0, 1)$ is the discount factor (see table 1 in appendix A) that weighs the long-term rewards relative to the short-term rewards. To maximize the expected discounted cumulative sum of rewards, the entries $Q(o, a)$ are

updated each time the agent takes an action after an observation, according to the update rule

$$Q(o, a) = (1 - \alpha)Q(o, a) + \alpha(r_t + \gamma \max_a (Q(o', a))), \quad (3.3)$$

where α is the learning rate and o' is the subsequent observation made by the agent. The policy is a greedy policy, meaning that, after each observation, the agent takes the action that corresponds to the highest value. The learning is done in episodes; each episode corresponds to the crawler moving a fixed distance forward, after which it is reset to its original undeformed configuration. The crawler goes through a number of episodes in this manner, gaining experience in the interactions between neurons, body mechanics and environment, updating its Q-matrix as it goes through the episodes. It is worth emphasizing that our learning algorithm has just two parameters, a learning rate α and a discount factor γ, in contrast to many recent variants of RL that have many hyper-parameters; thus most reasonable choices for these will converge and yield similar policies. We choose $\alpha = 0.05$ to allow for stochastic effects and $\gamma = 0.95$ to strive towards the case of high long-time rewards [13].

3.2. Experimental results: regularized and unregularized gaits

We initialize the crawler in an undeformed state, with a Q-matrix of values that are uniform and high. Then the crawler is equally likely to take any action independent of the observable state of the crawler, and since the values are high, i.e. the reward is lower than the expected reward, the crawler explores other actions. This leads to uncoordinated gaits; an example is shown in appendix A, figure 4. As the Q-matrix converges towards its steady-state value, the rewards become closer to the expectation of the crawler and the policy converges to a coordinated gait.

Figure 2 shows two coordinated gaits corresponding to two values of the regularization parameter $\epsilon = 0.01$ (figure 2*a*) and $\epsilon = 0$ (figure 2*b*), as defined in equation (3.2). In both of the gaits, the crawler moves by means of a travelling wave of contraction from tail to head. The regularized gait corresponds to observations of a larva consistent with experiments, wherein a localized wave causing sequential segmental contraction moves from tail to head as shown in figure 2*a*. By contrast, the unregularized gait, corresponding to $\epsilon = 0$, is characterized by a 10% higher speed, and larger variations in segment strain, and is due to the fact that some muscles are never activated (figure 2*b*, right), leading to pairs of segments moving together (see electronic supplementary material, video S2). The policies for both gaits are shown in figure 2*c*. These results justify our use of a regularization penalty in the reward to recover gaits that are biologically plausible and are also consistent with the diagonal neuronal weights that result.

To further compare the gaits, we show the power expenditure, cycle duration and robustness to noise in figure 3. The power exerted at each node,

$$p_i = |f_i - f_{i-1}||u_i|, \quad (3.4)$$

is a periodic function for both cases. For the regularized gait, the maximum power and the duration for which power is non-zero are both more uniform across the interior nodes, while for the unregularized gait there is a larger variation in power across nodes (figure 3*a*). Figure 3*b* shows the distribution of cycle duration for the two gaits, and shows that the higher speed of the unregularized gait is achieved via a faster propagation of waves along the length of the crawler.

To test whether these policies are robust, we explored the response of the two gaits to uncertainty in the crawler's ability to sense proprioceptive strain. We implement this by replacing the deterministic observation of the most compressed segment, given by (3.1), by a noisy version with $o = \text{argmin}_{i\in(1,\ldots,N)}(u_i - u_{i-1} + U)$, where $U \in [-n_0, n_0]$ is a uniformly distributed random variable and n_0 is the maximum amplitude of the noise. We find that while the regularized gait has a lower speed than the unregularized gait at low levels of noise n_0, the regularized gait maintains its speed as the noise level increases, while the unregularized gait does not. This trade-off between speed and robustness to noise is demonstrated by the crossover in figure 3*c*. Comparing the segment strain over the course of a cycle, we observe that the unregularized gait varies over a smaller range than the regularized gait (denoted by a smaller contrast in colours for a particular segment in figure 2*b* versus figure 2*a*). This suggests that the unregularized gait should be more susceptible to proprioceptive noise, consistent with what is observed in figure 3*c*.

4. Discussion

There is now much interest in using a range of machine learning techniques for modelling movement control in the context of bipedal walking, running and jumping [21–23] and more complex modes of movement such as swimming, gliding, etc. [15,16]. Our minimal approach to learning a coordinated gait for rectilinear crawling embeds the question of determining the neural weights via RL in a framework linking the brain, the body and the environment. Our observations and actions are all couched in terms of biophysically motivated quantities such as relative displacements, forces and neural spike patterns, along with rewards that are characterized by speed and internal strains and strain rates. The solutions that we converge to recover the wiring patterns and propagating contractile waves similar to experimental observations [4] and theoretical studies [11,12]. Regularizing the reward to penalize strain gradients provides smooth gaits that expend power more uniformly in space and time, as well as gaits that are robust to uncertainty in the crawler's ability for proprioception, but at the cost of speed. Indeed there is a trade-off between speed and robustness when these gaits are challenged by proprioceptive noise. This qualitative result of our study is suggested in the developmental neurobiology literature [4,10] and has potential applications in the design of robotic analogues of crawlers [24].

Quantifying the functional form of this speed–robustness trade-off is a natural next question which will require specific choices for the form of the uncertainty in the environment (e.g. friction), in the body (e.g. mechanical properties that change due to developmental defects or injury) and in the brain (e.g. wiring that changes due to molecular mechanisms of external-induced injury). Looking ahead, our one-dimensional model needs to be augmented to account for bending deformations to reproduce larval motions that include both axial and transverse motions. Recent theoretical work [25] has suggested that a minimal reflexive strategy embedded in a neuromechanical model can give rise to a basis of exploratory locomotory gaits that include rectilinear crawling and turning. By modifying our model and changing the reward structure in

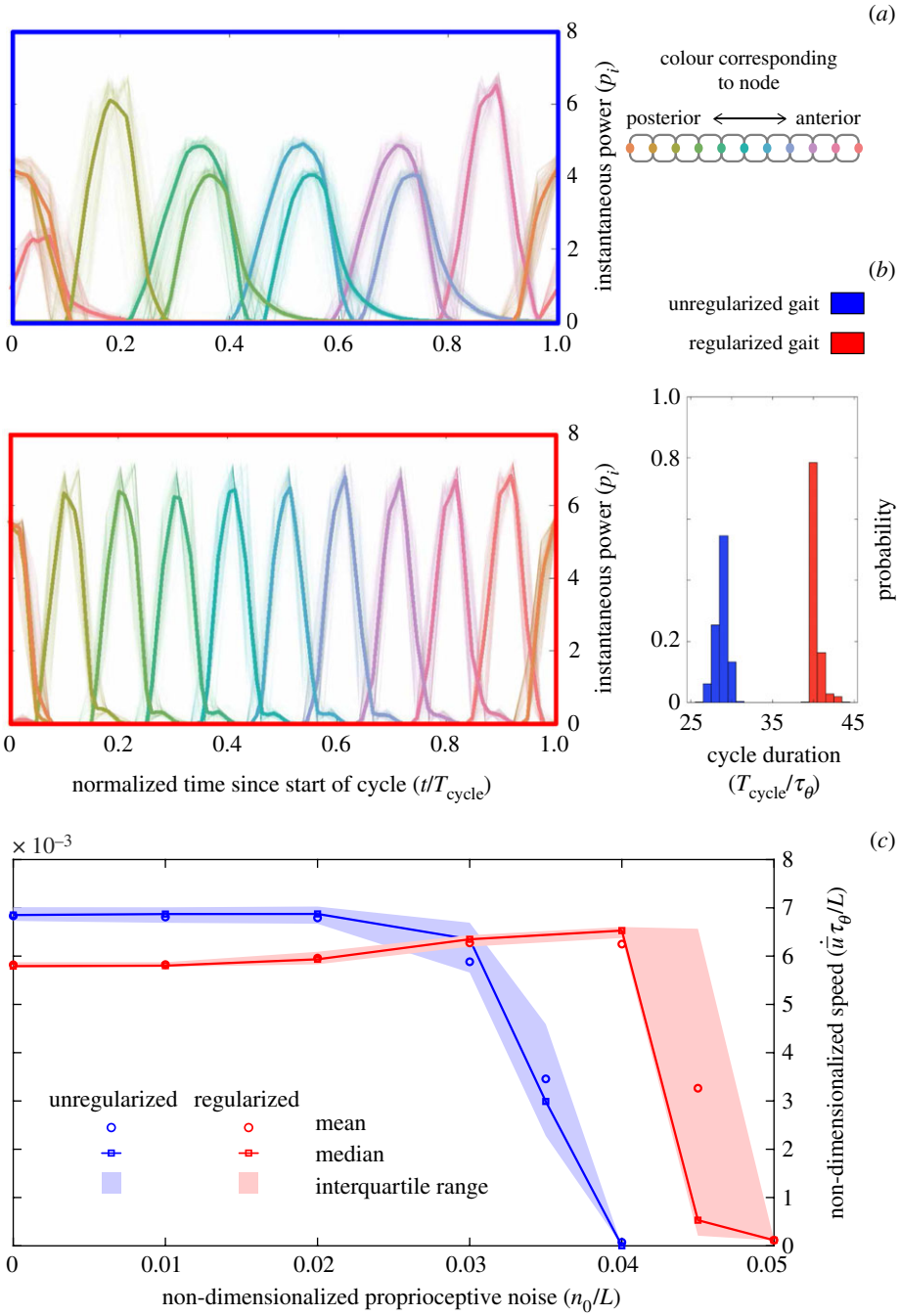

Figure 3. Comparison of the unregularized (blue) and regularized (red) gaits. (*a*) Power p_i computed using equation (3.4) for each node $i \in [0, 10]$ as a function of phase in a cycle, for a number of cycles. The different colours correspond to node numbers as labelled. (*b*) Histogram for the duration of a cycle for the unregularized (blue) and regularized (red) gaits. (*c*) Speed versus proprioception noise for the unregularized (blue) and regularized (red) gaits.

our RL framework dynamically as a function of additional internal observations of the crawler, it might be possible to bias the gaits towards coordinated crawling that is exploitative, or towards random turning that is exploratory. More generally, our study is but the first step in determining neural actuation patterns for a range of complex tasks by combining models that couple the mechanics of the brain, body and environment with machine learning.

Data accessibility. All codes for the RL algorithm are available from the corresponding author upon request.

Authors' contributions. L.M. and W.M.v.R. conceived of the study; S.M., W.M.v.R. and L.M. designed the study; W.M.v.R. and S.M. wrote the code; S.M. ran experiments and made the figures; S.M. and L.M. wrote the paper; all authors edited the paper.

Competing interests. We declare we have no competing interest.

Funding. We acknowledge the following sources for partial financial support: Swiss National Science Foundation, US ARO W911NF-15-1-0166, NSF DMR-2011754, NSF EFRI-1830901 and NSF DMR-1922321.

Acknowledgements. We thank Daniel Fortunato, Jordan Hoffmann and Vamsi Spandan for discussions and feedback on the paper.

Appendix A

A.1. Parameter values

See table 1 summarized.

A.2. Videos

We include two files that show the converged gait of the crawler with and without regularization, corresponding to figure 2*a* and *b*, respectively.

Table 1. Parameters and their values used in the simulation, where all lengths are scaled by the segment length L and all times are scaled by the neuronal relaxation time τ_θ.

symbol	quantity	value
L	segment length	1
τ_θ	neuronal timescale	1
$c\tau_\theta/k$	scaled damping	3.5
$f^m_{\max}/kL$	scaled muscular force	1
localized τ_m/τ_θ	scaled muscular timescale	1
$f^f_{\max}/kL$	scaled backward frictional force	9
ε^f	frictional smoothing	10^{-6}
η_f	friction anisotropy	30
Δt	scaled discrete timestep	0.01

Video S1—Regularized gait: coordinated gait that arises from an initial uncoordinated gait with a regularization parameter $\epsilon = 0.01$.

Video S2—Unregularized gait: coordinated gait that arises from an initial uncoordinated gait with no regularization parameter, i.e. $\epsilon = 0$, which leads to motion where multiple segments that move concurrently. This gait is not robust to proprioceptive noise and is easily disrupted (see figure 3*c* and corresponding text for details).

A.3. Unlearned gait

See figure 4.

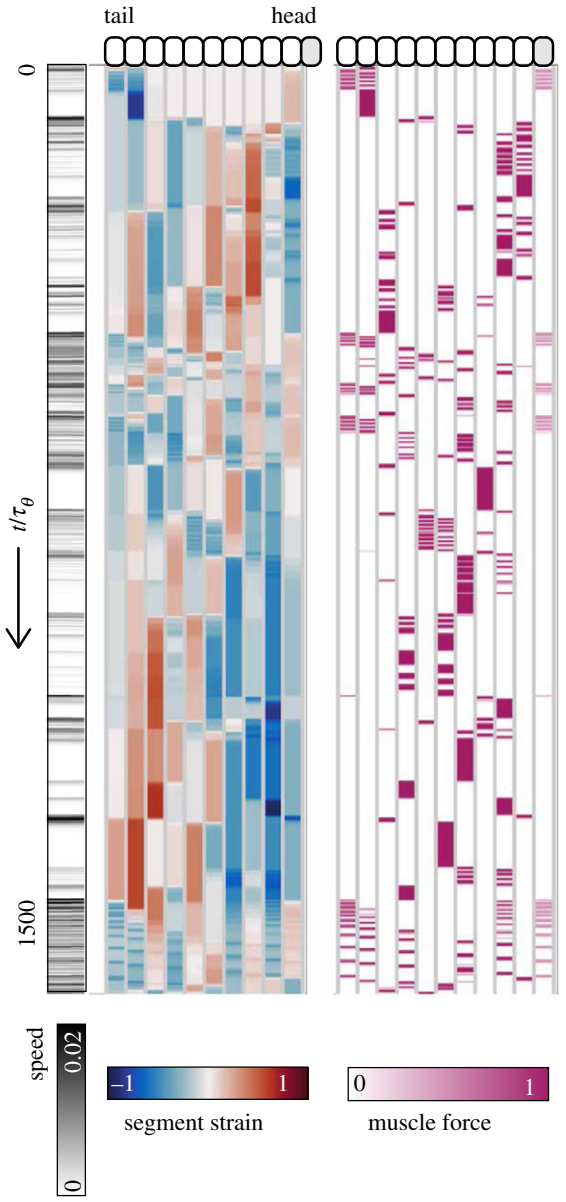

Figure 4. Uncoordinated gait resulting from a fully connected neuronal network, as summarized by equations (2.1)–(2.8) of the main text. The speed of the centre of mass is in grey (left), corresponding segment strains in blue-red (middle) and muscle force in pink (right). The parameter values are summarized in table 1.

References

1. Chiel HJ, Beer RD. 1997 The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.* **20**, 553–557. (doi:10.1016/S0166-2236(97)01149-1)
2. Rossignol S, Dubuc R, Gossard J-P. 2006 Dynamic sensorimotor interactions in locomotion. *Physiol. Rev.* **86**, 89–1554. (doi:10.1152/physrev.00028.2005)
3. Calhoun AJ, Murthy M. 2017 Quantifying behavior to solve sensorimotor transformations: advances from worms and flies. *Curr. Opin. Neurobiol.* **46**, 90–98. (doi:10.1016/j.conb.2017.08.006)
4. Heckscher ES, Lockery SR, Doe CQ. 2012 Characterization of *Drosophila* larval crawling at the level of organism, segment, and somatic body wall musculature. *J. Neurosci.* **32**, 12 460–12 471. (doi:10.1523/JNEUROSCI.0222-12.2012)
5. Narayanan DZ, Ghazanfar AA. 2014 Developmental neuroscience: how twitches make sense. *Curr. Biol.* **24**, R971–R972. (doi:10.1016/j.cub.2014.08.052)
6. Berni J, Pulver SR, Griffith LC, Bate M. 2012 Autonomous circuitry for substrate exploration in freely moving *Drosophila* larvae. *Curr. Biol.* **22**, 1861–1870. (doi:10.1016/j.cub.2012.07.048)
7. Eltringham SK. 1971 *Life in mud and sand*. New York, NY: Crane, Russak.
8. Trueman E. 1975 *The locomotion of soft bodied animals*. London, UK: Edward Arnold.
9. Garrey WE, Moore AR. 1915 Peristalsis and coordination in the earthworm. *Am. J. Physiol. Legacy Content* **39-2**, 139–148. (doi:10.1152/ajplegacy.1915.39.2.139)
10. Suster ML, Bate M. 2002 Embryonic assembly of a central pattern generator without sensory input. *Nature* **416**, 174–178. (doi:10.1038/416174a)
11. Paoletti P, Mahadevan L. 2014 A proprioceptive neuromechanical theory of crawling. *Proc. R. Soc. B* **281**, 20141092. (doi:10.1098/rspb.2014.1092)
12. Pehlevan C, Paoletti P, Mahadevan L. 2016 Integrative neuromechanics of crawling in *D. melanogaster* larvae. *Elife* **5**, e11031. (doi:10.7554/eLife.11031)
13. Sutton RS, Barto AG. 1998 *Reinforcement learning: an introduction*. New York, NY: MIT Press.
14. Silver D *et al.* 2017 Mastering the game of go without human knowledge. *Nature* **550**, 354–359. (doi:10.1038/nature24270)
15. Colabrese S, Gustavsson K, Celani A, Biferale L. 2017 Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.*

118, 158004. (doi:10.1103/PhysRevLett.118.158004)
16. Reddy G, Celani A, Sejnowski TJ, Vergassola M. 2016 Learning to soar in turbulent environments. *Proc. Natl Acad. Sci. USA* **113**, E4877–E4884. (doi:10.1073/pnas.1606075113)
17. Kohsaka H, Guertin P, Nose A. 2017 Neural circuits underlying fly larval locomotion. *Curr. Pharm. Design* **23**, 1722–1733. (doi:10.2174/1381612822666161208120835)
18. Vaadia RD, Li W, Voleti V, Singhania A, Hillman E, Grueber WB. 2019 Characterization of proprioceptive system dynamics in behaving *Drosophila* larvae using high-speed volumetric microscopy. *Curr. Biol.* **29**, 935–944. (doi:10.1016/j.cub.2019.01.060)
19. Ermentrout GB, Kopell N. 1986 Parabolic bursting in an excitable system coupled with a slow oscillation. *SIAM J. Appl. Math.* **46**, 233–253. (doi:10.1137/0146017)
20. Hughes CL, Thomas JB. 2007 A sensory feedback circuit coordinates muscle activity in *Drosophila*. *Mol. Cell. Neurosci.* **35**, 383–396. (doi:10.1016/j.mcn.2007.04.001)
21. Heess N *et al.* 2017 Emergence of locomotion behaviours in rich environments. (http://arxiv.org/abs/1707.02286)
22. Peng XB, Berseth G, Yin K, Van De Panne M. 2017 DeepLoco: dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Trans. Graphics* **36**, 41.
23. Hwangbo J, Lee J, Dosovitskiy A, Bellicoso D, Tsounis V, Koltun V, Hutter M. 2019 Learning agile and dynamic motor skills for legged robots. *Sci. Rob.* **4**, eaau5872.
24. Chen S, Cao M, Sarparast Y, Yuan H, Dong X, Tan L, Cao C. 2020 Soft crawling robots: design, actuation, and locomotion. *Adv. Mater. Technol.* **5**, 1900837. (doi:10.1002/admt.201900837)
25. Loveless J, Lagogiannis K, Webb B. 2019 Modelling the mechanics of exploration in larval *Drosophila*. *PLoS Comput. Biol.* **15**, e1006635. (doi:10.1371/journal.pcbi.1006635)