


Combining Reinforcement Learning and Tensor Networks, with an Application to Dynamical Large Deviations

Edward Gillman^{1,2}, Dominic C. Rose³, and Juan P. Garrahan^{1,2}

¹*School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, United Kingdom*

²*Centre for the Mathematics and Theoretical Physics of Quantum Non-Equilibrium Systems, University of Nottingham, Nottingham NG7 2RD, United Kingdom*

³*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, United Kingdom*

 (Received 3 November 2022; revised 28 February 2024; accepted 4 April 2024; published 7 May 2024)

We present a framework to integrate tensor network (TN) methods with reinforcement learning (RL) for solving dynamical optimization tasks. We consider the RL actor-critic method, a model-free approach for solving RL problems, and introduce TNs as the approximators for its policy and value functions. Our “actor-critic with tensor networks” (ACTeN) method is especially well suited to problems with large and factorizable state and action spaces. As an illustration of the applicability of ACTeN we solve the exponentially hard task of sampling rare trajectories in two paradigmatic stochastic models, the East model of glasses and the asymmetric simple exclusion process, the latter being particularly challenging to other methods due to the absence of detailed balance. With substantial potential for further integration with the vast array of existing RL methods, the approach introduced here is promising both for applications in physics and to multi-agent RL problems more generally.

DOI: [10.1103/PhysRevLett.132.197301](https://doi.org/10.1103/PhysRevLett.132.197301)

Introduction.—Tensor networks (TNs), routinely used in the study of quantum many-body systems [1–7], are increasingly being applied in machine learning (ML), see, e.g., Refs. [8–20]. Both domains often deal with systems with state spaces that are exponentially large in the number of degrees of freedom, say the number of qubits in a quantum system, or of pixels in images to be classified. In such situations TNs provide a powerful way to represent functions, vectors and distributions, while allowing for efficient sampling and computation of quantities such as inner products and norms.

To date, the intersection of TNs and ML has been mostly in supervised and unsupervised learning, see, e.g., Refs. [8–20]. In contrast, the combination of TNs and reinforcement learning (RL) [21] has been more limited, despite recent major advances in RL [22–26]. While some promising related directions have been explored, such as the approximation of Q functions in the context of large state spaces [27], and/or large action spaces [28], the flexible integration of TNs with RL remains an open problem, along with demonstrating useful applications.

Here, we introduce the actor-critic with tensor networks (ACTeN) method, a general framework for integrating TNs

into RL via actor-critic (AC) techniques. By combining decision-making “actors” with “critics” that judge an actor’s quality, AC methods are used in many state-of-the-art RL applications. Using TNs as the basis for modeling actors and critics within AC and RL represents a powerful combination to tackle problems with both large state and action spaces.

To demonstrate the effectiveness of our approach, we consider the problem of computing the large deviation (LD) statistics of dynamical observables in classical stochastic systems [29–35], and of optimally sampling the associated rare long-time trajectories [36–48]. Such problems are of wide interest in statistical mechanics and can be phrased straightforwardly as optimization problems that may be solved with RL [49–52] and similar techniques [53–61]. For concreteness we consider two models: (i) the East model, a kinetically constrained model used to study slow glassy dynamics, and (ii) the asymmetric exclusion process (ASEP), a paradigmatic model of nonequilibrium, in which particles hop around a lattice while blocking each others movement. In particular, and in contrast to the East model, the ASEP (with periodic boundaries) does not obey detailed balance [62], and thus evades straightforward use of TNs to compute spectral properties of the relevant dynamical generators. We demonstrate that ACTeN can be applied to both problems irrespective of the equilibrium or nonequilibrium distinction, by computing their dynamical LDs for sizes well beyond those achievable with exact methods. Given the vast array of options for improving the RL algorithm that we use, our results indicate that the

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

overall framework outlined here is highly promising for applications more generally.

Background: Reinforcement learning and actor-critic techniques.—A discrete-time Markov decision process (MDP) [21] consists at each time $t \in [0, T]$ of stochastic variables $X_t = (S_t, a_t, R_t)$, named state, action, and reward. We assume these are drawn from t -independent finite sets, \mathcal{S} , \mathcal{A} , and \mathcal{R} , where the action set may depend on the current state, $\mathcal{A}(S)$ for $S \in \mathcal{S}$. The action and state variables are associated with the policy, $\pi(a|S)$, and environment, $P(S'|S, a)$, distributions. These are sampled in a sequence of steps to generate a trajectory of the MDP, $\omega = (X_0, X_1, \dots, X_T)$, see Fig. 1(a). We assume that the reward is a deterministic function of the state and action variables.

In the typical scenario of policy optimization, the policy is controllable and known, while the environment is fixed and potentially unknown. We focus on MDPs that are “continuing” and admit a steady-state distribution independent of X_0 . We can then define the average reward per time step when following a given policy as $r(\pi) = \lim_{t \rightarrow \infty} \mathbb{E}_\pi[R_t]$, where $\mathbb{E}_\pi[\cdot]$ is the stationary state expectation over states and over transitions from those states according to policy π . The task of policy optimization is to find the policy π^* that maximizes $r(\pi)$.

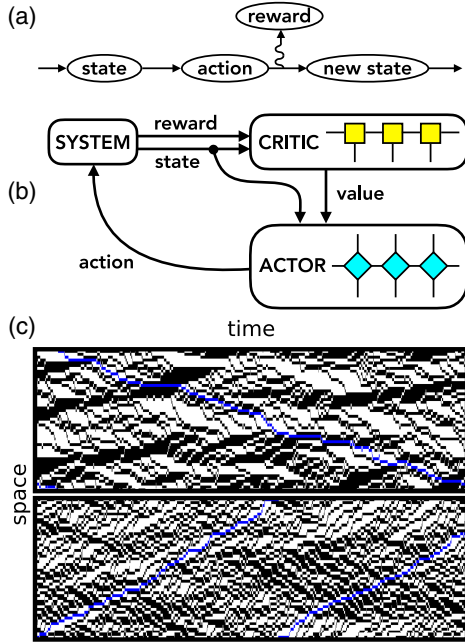


FIG. 1. Actor-critic with tensor networks. (a) Sketch of a Markov decision process. (b) In actor-critic RL, the state is passed to an actor, which chooses the action, and to a critic, which values the state given the reward. This value is used to improve the actor’s policy. In ACTeN, the function approximators for actor and critic are tensor networks. (c) Top: typical trajectory of the ASEP at half-filling and $L = 50$ sites with one particle highlighted (blue), shown for 3000 steps. Bottom: trajectory with a current large deviation, sampled from the ACTeN solution for biasing (counting) field $\lambda = -3$. See the text for details.

Reinforcement learning (RL) refers to the group of methods that aim to discover optimal policies by using the experience gained from sampling trajectories of an MDP. In policy gradient methods, the policy is approximated by a function $\pi_w(a|S)$ with parameters w , and optimized using the gradient of $r(\pi_w)$ with respect to w . Building on this, actor-critic methods then assess $\pi_w(a|S)$ by computing the value, $v_\pi(S)$, of the states that result when following the policy. This value is defined as the difference between rewards in the future of a given state and the average reward, $v_\pi(S) = \mathbb{E}_\pi[\sum_{\tau=t}^{\infty} [R_\tau - r(\pi)] | S]$. The gradient of $r(\pi_w)$ can be written exactly in terms of these values as

$$\nabla_w r(\pi) = \mathbb{E}_\pi[\delta_t^r \nabla_w \ln \pi_w(a_t | S_t)], \quad (1)$$

where we have introduced the temporal difference (TD) error, $\delta_t^r = v_\pi(S_{t+1}) + R_{t+1} - r(\pi_w) - v_\pi(S_t)$ [21,49], which quantifies if a resultant state is better than the current one. AC methods use this information to alter the probability of taking that action in the future.

In reality, calculating the true value of every state under the current policy is impractical, and thus an auxiliary approximation for the value function, $v_\psi(S)$ with parameters ψ , is introduced, the so-called critic. To optimize the critic, we note that the value of a state is related to the value of states reachable from it, as encoded in the differential Bellman equation [21],

$$v_\pi(S_t) = \mathbb{E}_\pi[v_\pi(S_{t+1}) + R_{t+1} - r(\pi_w) | S_t]. \quad (2)$$

Minimizing the error in the Bellman equation when substituting the critic for the true values can be done by updating the weight as $\psi' = \psi + \alpha \mathbb{E}_\pi[\delta_t^r \nabla_\psi v_\psi(S)(S_t)]$, in terms of the approximate TD error, $\delta_t^r = v_\psi(S_{t+1}) + R_{t+1} - r(\pi_w) - v_\psi(S_t)$, and a learning rate α . Intuitively, the estimated expected reward before a transition occurs, $v_\psi(S_t)$, is compared to the expected reward afterwards plus the true reward for that time step, $v_\psi(S_{t+1}) + R_{t+1} - r(\pi_w)$, and $v_\psi(S)$ adjusted to make these closer. The policy is then updated by following the gradient in Eq. (1) with the exact TD error δ^r replaced by the critic’s approximate TD error δ .

Analytic example: Two-site East model.—To illustrate these ideas, consider two spins, $s_{1,2} = 0, 1$, evolving with a constrained set of transitions as in the East model studied below, such that a spin can flip only when its left neighbor (in this case simply the other spin due to periodic boundaries) is 1. The states are $\mathcal{S} = \{00, 01, 10, 11\}$. The dynamics of this model can be implemented as an MDP using a policy that (stochastically) selects which spins to flip. Denoting no-flip or flip by 0 or 1 and requiring at most one spin flip per time step, the action sets for this model are then $\mathcal{A}(10) = \{00, 01\}$, $\mathcal{A}(01) = \{00, 10\}$, and $\mathcal{A}(11) = \{10, 01\}$, with the state 00 being disconnected from the rest. An example policy that selects from all

possible actions equally would assign a probability of $1/2$ to each of these, e.g., $\pi(a = 00|S = 10) = \pi(a = 01|S = 10) = 1/2$ and similarly for the other states. The transition to a new state is then enacted by the environment. We can choose this to be deterministic, and simply apply the spin-flip operations selected by the actor to the current state, $s_i \rightarrow (1 - a_i)s_i + a_i(1 - s_i)$. Finally, an example reward can be defined via the function $R(S, a, S') = -\lambda(1 - \delta_{S',S})$, which awards $-\lambda$ every time a spin flip occurs ($a \neq 0$), encouraging activity or inactivity for negative or positive λ . For this reward, the optimal policy for negative λ will maximize activity, flipping a spin at every step, requiring going from 10 and 01 to 11 with probability 1, i.e., $\pi^*(01|10) = 1$, etc.

Method: Actor-critic with tensor networks.—We now focus on applying AC to solve problems with large state and action spaces. For example, we may wish to find the optimal dynamics of a system of L binary components, resulting in 2^L states, with individual agents and their actions associated to each component. To ensure optimal choices for a given task, actions may need to be correlated not only with other agents states, but also the actions other agents are about to take. In such problems, simple approaches such as tabular RL fail due to exponentially large memory requirements and sampling costs, and a common alternative is to use neural networks (NNs) when defining $\pi_w(a|S)$ and $v_\psi(S)$. TNs offer another approach, with polynomial memory and computational costs, and showing state-of-the-art performance in many settings; see, e.g., the review [6]. Motivated by this, we define a general framework (which we call ACTeN) that exploits TNs to efficiently represent $\pi_w(a|S)$ and $v_\psi(S)$.

A TN is a set of tensors, $\mathcal{T} = \{T_{i_1 j_1 k_1 \dots}^{[1]}, T_{i_2 j_2 k_2 \dots}^{[2]}, \dots\}$, contracted in some pattern, cf. Fig. 1(b). This results in a single tensor that can be viewed as defining a multivariate function, $\varphi(x) = T_x$, $T_x = \mathcal{C}[\mathcal{T}]$, where \mathcal{C} indicates the chosen contractions and x all remaining uncontracted indices. For a given problem, the selection of an appropriate TN depends on factors such as dimensionality and geometry. Here we consider applying ACTeN to study one-dimensional (1D) systems with periodic boundaries (PBs) and L components, such that a state is $S = (s_1, \dots, s_L)$ with s_i taking d values. To represent the value function $v_\psi(S)$ we use a translation invariant matrix product state (MPS) which mirrors the chain geometry of the system. This TN is built from a single real-valued tensor \mathcal{A}_{sij} of shape (d, χ, χ) (or equivalently d square χ -dimensional matrices $[A_s]_{ij} = A_{sij}$) whose elements encode the parameters of $v_\psi(S)$, $\psi = \{A_{sij}\}_{s=1}^d$. For a given $S = (s_1, \dots, s_L)$ this TN is defined as

$$\varphi(S) = \text{Tr}[A_{s_1} A_{s_2} \dots A_{s_L}], \quad (3)$$

i.e., for each site we select the corresponding matrix A_{s_i} , multiplying the L matrices together with a trace to produce a real scalar. To take advantage of translation invariance and apply approximations from smaller L to larger L systems, we then define the value function in terms of $\varphi(S)$ after the additional application of a square and log, which prevents the exponential growth or decay of values as L is changed for fixed A_{sij} . Hence,

$$v_\psi(S) = \log[\varphi(S)^2]. \quad (4)$$

To define $\pi_w(a|S)$ we use a matrix product operator (MPO). This TN is built from a single real-valued (d_S, d_A, χ, χ) -shaped tensor, \mathcal{A}_{asij} , equivalent to $d_S \times d_A$ χ -dimensional square matrices $[A_{as}]_{ij} = A_{asij}$, i.e., one matrix per combination of local state and action. Given a state $S = (s_1, \dots, s_L)$ and actions $a = (a_1, \dots, a_L)$, the contraction is given by the traced matrix product,

$$\varphi(a, s) = \text{Tr}[A_{a_1 s_1} A_{a_2 s_2} \dots A_{a_L s_L}]. \quad (5)$$

To use this to define a policy, we need to ensure positivity and normalization, as well as preventing the policy from producing invalid actions. To achieve this we define

$$\pi_w(a|S) = \mathcal{C}(a, S) [\mathcal{N}(S)]^{-1} \varphi(a, S)^2, \quad (6)$$

where $\mathcal{N}(S) = \sum_{a|\mathcal{C}(a,S)=1} \varphi(a, S)^2$ is the (state-dependent) normalisation factor and $\mathcal{C}(a, S)$ returns one if an action a is possible in state S or zero otherwise [63].

Application: Dynamical large deviations.—To test ACTeN we consider the problem of computing the large deviations (LDs) of trajectory observables [29,31,33] in the East model and the ASEP in 1D with PBs. Both models are many-body binary spin systems with large state-spaces for large L , $S = \{s_i\}_{i=1}^L$, with $s_i = 0, 1$ (see the Appendix for details on the models). The dynamics of these systems are subject to local constraints that lead to rich behaviors in their trajectories, $\omega_0^T = \{S_t\}_0^T$. This can be observed in the time integrals of time-local quantities, $O(\omega_0^T) = \sum_{t=1}^T o(S_t, S_{t-1})$, the moments of which are contained in derivatives of the moment generating function (MGF), $Z_T(\lambda) = \sum_{\omega_0^T} e^{-\lambda O(\omega_0^T)} P(\omega_0^T)$, where $P(\omega_0^T) = \prod_{t=1}^T P(S_t|S_{t-1})P(S_0)$ is the trajectory probability under the dynamics.

In the long-time limit, the MGF obeys a large deviation principle [29,31,33] with the scaled cumulant generating function (SCGF), $\theta(\lambda) = \lim_{T \rightarrow \infty} (1/T) \ln Z_T(\lambda)$, playing the role of a free-energy for trajectories. In principle, the SCGF can be obtained by sampling methods. However, this is exponentially hard (in time and space) using the original dynamics. An alternative is to find a more efficient sampling dynamics which may then be combined with importance sampling to obtain unbiased statistics. This can

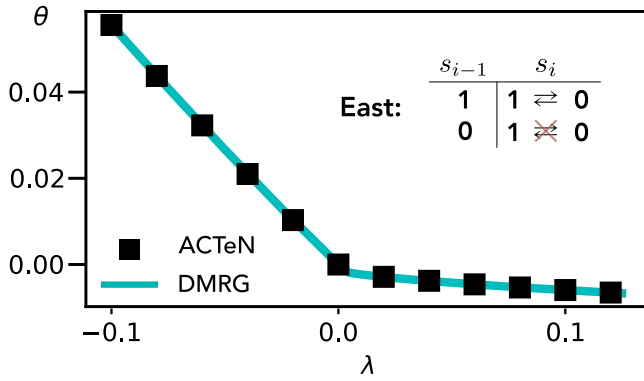


FIG. 2. Dynamical large deviations in the East model using ACTeN. Scaled-cumulant generating function for the dynamical activity of the East model as a function of biasing field λ from ACTeN (symbols), for $L = 50$ and PBC. Our RL results coincide with those obtained from the current state-of-the-art method using DMRG, cf. Ref. [64] (which is possible since the East model obeys detailed balance). Inset: Kinetic constraint of the East model; a spin, s_i , can flip, $s_i \rightarrow 1 - s_i$, only if the spin to the left is up, $s_{i-1} = 1$.

be formulated as a RL problem as follows: can we find a parametrized dynamics $P_w(S_t|S_{t-1})$ such that $P_w(\omega_0^T) = e^{-\lambda O(\omega_0^T)} P(\omega_0^T) / Z_T(\lambda)$, i.e., it reproduces a trajectory ensemble biased towards rare trajectories of the original dynamics. This dynamics is connected to an underlying policy $\pi_w(a|S)$ by a deterministic environment which returns states after receiving an associated action, i.e., $P_w[S' = f(a, S)|S] = \pi_w(a|S)$, where for each S , $f(a, S)$ returns a unique S' for each a . For example, in the East model if we take the action $a = \{a_i\}_{i=1}^L$ then sites with $a_i = 1$ are flipped and those with $a_i = 0$ are not flipped; the new state is then $S' = f(a, S) = \{(1 - a_i)s_i + a_i(1 - s_i)\}_{i=1}^L$.

Optimizing the KL divergence between the two trajectory ensembles gives a regularized form of RL with a reward depending on the policy [49]

$$R_t = -\lambda o(S_t, S_{t-1}) - \ln \left(\frac{P_w(S_t|S_{t-1})}{P_{\text{orig}}(S_t|S_{t-1})} \right), \quad (7)$$

with its expected value becoming the SCGF at optimality [49]. Intuitively, choosing actions (e.g., flips) to maximize the first term increases the likelihood of rare events with extreme values of the observable, while maximizing the second term minimizes the difference between the parametrized and original dynamics, thus making the event more probable. Maximizing this reward is a balancing act between these two aims, resulting in dynamics biased towards rare events in a way representative of their occurrence in the original dynamics. In the appendices, we illustrate ACTeN by solving explicitly the two-site East model and showing how this can be exactly represented by the TN ansatz.

(i) East model and dynamical activity: Figure 2 shows the SCGF of the dynamical activity [total number of spin flips in a trajectory, defined by $o(S_t, S_{t-1}) = 1 - \delta_{S_t, S_{t-1}}$], calculated using ACTeN (symbols). Since the East model obeys detailed balance, the SCGF is the log of the largest eigenvalue of a Hermitian operator and can be estimated via density matrix renormalization group (DMRG) methods, cf. Refs. [64–70] (here we tensors.jl [71]). Figure 2 shows that the DMRG results (blue curve) coincide with ACTeN (black squares) for size $L = 50$, which is well beyond what is accessible to exact diagonalization (ED). Note that DMRG with PBs tends to be much less numerically stable than for open boundaries. Nonetheless, ACTeN can reach $L \gtrsim 50$ without the need for any special stabilization techniques.

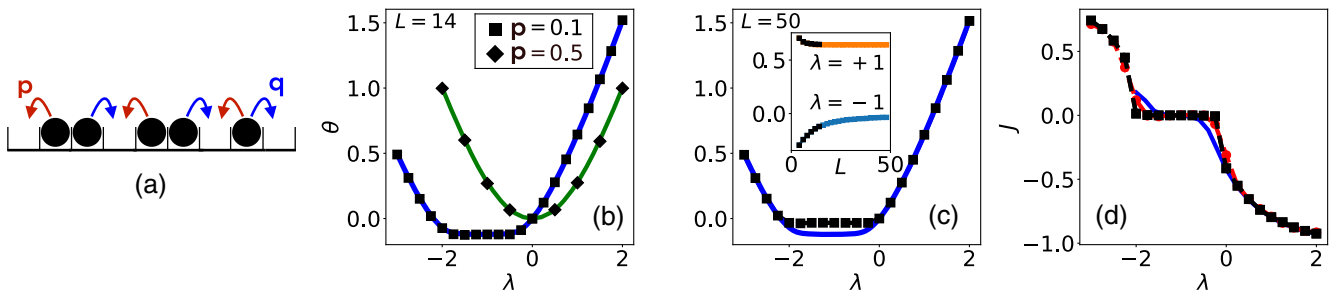


FIG. 3. Dynamical large deviations in the ASEP using ACTeN. (a) In the ASEP particles can only move to an unoccupied neighboring site, with probability p to the left and $q = 1 - p$ to the right. (b) SCGF for the time-integrated particle current as a function of biasing field. We show results from ACTeN for $p = 0.1$ (squares) and $p = 1/2$ (diamonds). The lack of detailed balance for PBC and $p \neq 1/2$ prevents straightforward application of DMRG, but for small sizes (here $L = 14$) we can compare to exact diagonalization (blue curve for $p = 0.1$, green for $p = 1/2$). (c) SCGF for $p = 0.1$ from ACTeN for size $L = 50$ which is beyond the scope of ED. Compared to $L = 14$ (blue curve from ED), we see that ACTeN captures the flattening of the SCGF for larger sizes indicative of a LD phase transition, cf. Ref. [32]. The inset shows the smooth convergence of our ACTeN numerics with L for two values of λ . (d) Since ACTeN provides direct access to the optimal dynamics, observables such as the time-integrated current can be evaluated directly (black squares for $L = 50$). We show for comparison the numerical differentiation of the ACTeN SCGF (red circles) and of the ED SCGF at $L = 14$ (blue line).

(ii) ASEP and particle current: Figure 3 presents the LDs of the time-integrated particle current, defined by $o(s_t, s_{t-1}) = \frac{1}{2} \sum_{i=1}^L s_{t-1}^i s_t^{i+1} - s_t^i s_{t-1}^{i+1}$. Figure 3(b) shows the SCGF obtained via ACTeN (black squares or diamonds). Unlike the East model, for asymmetric hops ($p \neq 1/2$) Hermitian DMRG cannot be applied directly to the ASEP, so for comparison we show results from exact diagonalization for both $p = 0.1$ (blue line) and $p = 1/2$ for $L = 14$. Beyond $L = 14$ ED becomes prohibitive, while ACTeN remains feasible. Figures 3(c) and 3(d), show the expected phase transition behavior [32] and convergence with L up to $L = 50$. The optimal dynamics itself, i.e., the learned policy, can be used to generate trajectories representative of $\lambda \neq 0$, see Fig. 1(c), and directly sample rare values of the integrated current, see Fig. 3(d).

Outlook.—ACTeN compares very favorably with state-of-the-art methods for computing rare events without some of the limitations, such as boundary conditions or detailed balance. From the corpus of research in both TNs and RL, our approach has considerable potential for further improvement and exploration. These include: numerical improvements to precision via hyper-parameter searches; stabilization strategies for large systems; integration with trajectory methods such as transition path sampling or cloning; integration with advanced RL methods such as those offered by the DeepMind ecosystem [72]; generalization to continuous-time dynamics; and applications to other multi-agent RL problems, such as PistonBall [73], via integration with additional processing layers particularly those for image recognition.

We would like to thank Christopher J. Turner for useful discussions. We acknowledge funding from The Leverhulme Trust Grant No. RPG-2018-181, EPSRC Grant No. EP/V031201/1, and University of Nottingham Grant No. FiF1/3. We are grateful for access to the University of Nottingham’s Augusta HPC service. D. C. R. was supported by funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 853368). We thank the creators and community of the Python programming language [74], and acknowledge use of the packages JAX [75], NumPy [76], PANDAS [77], Matplotlib [78] and H5PY [79]. We also thank the creators and community of the JULIA programming language [80], and acknowledge use of the packages ITensors.jl [71] and HDF5.jl [81].

Appendix A: On kinetically constrained models.—

(i) East model: Flipping of a spin is constrained on the spin to its left being in state 1 [31,82]. Dynamics then amounts to two steps: first, select a random site i with probability $1/L$; second, if spin $s_{i-1} = 1$ then flip s_i . Given N spins in state 1 and periodic boundary conditions, the transition probability is $P(s'|s) = (1/L)$ for each possible new state $s' \neq s$, and probability $P(s|s) = 1 - (N/L)$ for no flip occurring.

(ii) Asymmetric simple exclusion process: The constraint is particle exclusion: a particle at site i ($s_i = 1$) can move left or right only if the destination is unoccupied [83]. The movement of a particle to, say, the right thus corresponds to $10 \rightarrow 01$, i.e., a flip of both spin variables. The dynamics again amounts to two steps: first, select a particle with probability $1/N$, with $N = \sum_i s_i$ the particle number; second, choose whether this particle hops right or left with probabilities p or $1 - p$, respectively, with the hop occurring if the new site is unoccupied. The transition probabilities are then $P(s'|s) = p/N$ for a right hop; $P(s'|s) = (1 - p)/N$ for a left hop; and, given the number of neighboring particles $N_{\text{nn}} = \sum_{i=1}^L s_i s_{i+1}$, with $s_{L+1} := s_1$, the probability of no change is $P(s|s) = 1 - N_{\text{nn}}/N$. In the main text, we consider the case of half-filling, $N = L/2$.

Appendix B: On analytical solution of the two-site East model with tensor network ansatz.—An exact solution for the optimal policy in the two-site East model can be found by analytically constructing the so-called “Doob dynamics” [34,84] as follows. First we define a “tilted” evolution operator [29,31,33] $P_\lambda(S_t|S_{t-1}) = e^{-\lambda o(S_t, S_{t-1})} \times P(S_t|S_{t-1})$, such that $Z_T(\lambda) = \langle -|P_\lambda^T|P_{\text{ss}} \rangle$ where $\langle -| = \sum_S \langle S|$ and $|P_{\text{ss}} \rangle$ is the stationary state vector for P . In the large T limit the SCGF is the log of the largest eigenvalue of P_λ [29,31,33]. The corresponding left eigenvector l_λ is related to the dynamics that maximizes the expected value of Eq. (7), given by the so-called Doob (or optimal) dynamics $P_\lambda^D(S_t|S_{t-1}) = [l_\lambda(S_{t-1}) \times P_\lambda(S_t|S_{t-1})/e^{\theta(\lambda)} l_\lambda(S_{t-1})]$. Applied to the two-site East model, defining the function $a(\lambda) = [4/(1 + \sqrt{1 + 8e^{-2\lambda}})]$, we find $\theta(\lambda) = -\ln(a(\lambda))$, with optimal dynamics

$$\begin{aligned} P_\lambda^D(10|10) &= P_\lambda^D(01|01) = \frac{a(\lambda)}{2}, \\ P_\lambda^D(11|10) &= P_\lambda^D(11|01) = 1 - \frac{a(\lambda)}{2}, \\ P_\lambda^D(10|11) &= P_\lambda^D(01|11) = \frac{1}{2}. \end{aligned}$$

We may then find the corresponding value function by solving the differential Bellman equation (2). To do this, note first that by symmetry the values of states 10 and 01 are identical, and second that Eq. (2) is invariant under an overall shift of the value function by a constant. Therefore, we may choose the value of states 10 and 01 to be 0, and thus find $V_\lambda(11) = -\lambda - \theta(\lambda)$.

These results are what is expected intuitively. Trajectories biased towards enhanced activity ($\lambda < 0$) have $a(\lambda) < 1$, making $P_\lambda^D(11|10) = P_\lambda^D(11|01) > 1/2$, i.e., the system is more likely to transition to the state that is guaranteed to flip at the next step rather than remain in 01 or 10. Furthermore, $V_\lambda(11) > 0$, i.e., the state guaranteed to flip is more valuable. In contrast, trajectories biased towards reduced activity ($\lambda > 0$) show the opposite behavior.

To connect P_λ^D to our policy ansatz Eq. (6), we first rewrite it as an operator. Using projection operators $P_1 = |1\rangle\langle 1|$, $P_0 = |0\rangle\langle 0|$, and flip operators $\sigma_+ = |1\rangle\langle 0|$, $\sigma_- = |0\rangle\langle 1|$, we define $A(\lambda) = \sigma_- + a(\lambda)\sigma_+ + [2 - a(\lambda)]P_0$. We may thus write $P_\lambda^D = 0.5(P_1 \otimes A(\lambda) + A(\lambda) \otimes P_1)$.

The policy ansatz Eq. (6) involves the elementwise square of an operator $\varphi(S', S)$: we thus seek the elementwise square root of P_λ^D . We define $[\tilde{A}(\lambda)]_{ij} = \sqrt{[A(\lambda)]_{ij}}$. Because of their sparsity structures, we have $\tilde{A}(\lambda) \odot P_1 = 0$, where \odot is element-wise (Hadamard) multiplication. Since $(A \otimes B) \odot (C \otimes D) = (A \odot C) \otimes (B \odot D)$, we find $\varphi = \sqrt{0.5}(P_1 \otimes \tilde{A}(\lambda) + \tilde{A}(\lambda) \otimes P_1)$ is such that $\varphi \odot \varphi = P_\lambda^D$. It remains to factor φ into an MPO of the form of Eq. (5). We may rewrite this as $\varphi = \sum_{d_1, d_2=1}^\chi T_{d_1 d_2} \otimes T_{d_2 d_1}$ where each $T_{d'd}$ is a 2×2 matrix acting on the single site state space. This T can be constructed by taking $\chi = 2$ with $T_{11} = T_{22} = 0$, $T_{12} = 0.5^{1/4}P_1$, and $T_{21} = 0.5^{1/4}\tilde{A}(\lambda)$. We thus find this order-4 tensor T reproduces the exact Doob dynamics from our translation invariant MPO-based ansatz.

For the value function, taking an exponential and square-root element-wise of the value function to invert Eq. (4) leads to the vector $\tilde{V}^\lambda = (e^{V_\lambda(11)/2} - 1)|1\rangle \otimes |1\rangle + |-\rangle \otimes |-\rangle$. Since this is already a sum of symmetric products, it is easy to rewrite it as a translation invariant MPS with $\chi = 2$, i.e., $\tilde{V}_{s_1 s_2}^\lambda = \sum_{d_1, d_2=1}^2 v_{s_1}^{d_1 d_2} v_{s_2}^{d_2 d_1}$, with order-3 tensor v such that $v_1^{11} = \sqrt{e^{V_\lambda(11)/2} - 1}$, $v_1^{22} = v_0^{22} = 1$, and $v_s^{d'd} = 0$ otherwise.

Appendix C: On training procedure.—We now provide details on the procedure used to obtain the policies of the

main text. First, we outline the update step used to improve the policy and value function approximations. Second, we outline size annealing, where we apply transfer learning using systems of increasing size. Finally, we discuss policy evaluation and selection, whereby the best policy is chosen from a set of candidates. Further implementation details are provided in [85,86].

(i) Basic outline of training: We start by initializing the parameters w_0 , ψ_0 , and \bar{r}_0 , where \bar{r}_t is an estimate of the average reward per-time step, $r(\pi_w)$, after t training steps, along with the environment and initial state s_0 . Choosing the three learning rates α_π , α_v , and α_r , for each step $t \in [0, T]$ we (1) Sample an action $a_t \sim \pi_w(\cdot|s_t)$ [where $x \sim Y(\cdot)$ stands for x sampled from Y], and from it get its log probability and eligibility, $\ln \pi_w(a_t|s_t)$, $\nabla_w \ln \pi_w(a_t|s_t)$. (2) Get the next state and reward given the current state and action, $(s_{t+1}, r_{t+1}) \sim P(\cdot, \cdot|a_t, s_t)$. (3) Get the temporal difference error with the current value function, $\delta_{t+1} = v_\psi(s_{t+1}) + r_{t+1} - \bar{r}_t - v_\psi(s_t)$. (4) Update the parameters of the value function, $\psi_{t+1} = \psi_t + \alpha_v \delta_{t+1} \nabla_\psi v_\psi(s_t)$. (5) Update the parameters of the policy, $w_{t+1} = w_t + \alpha_\pi \delta_{t+1} \nabla_w \ln \pi_w(a_t|s_t)$. (6) Update the estimate of the average reward per time step, $\bar{r}_{t+1} = \bar{r}_t + \alpha_r \delta_{t+1}$.

(ii) Annealing and transfer learning: In the context of machine learning, “annealing” (sequentially solving an optimization problem reusing solutions to improve an initial guess) can be considered a form of *transfer learning*. In our case, we anneal the size of the system: the optimal policies for two system sizes will be similar as long as $L' \gtrsim L$, and in the settings considered the optimal dynamics should converge as $L \rightarrow \infty$.

We first approximate the optimal policy for a small system, $L = 4$, starting from random initial weights. The

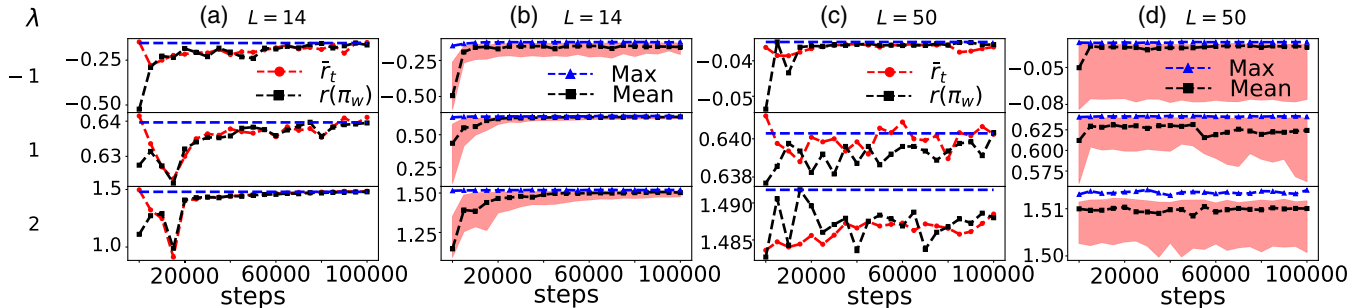


FIG. 4. Training procedure and learning curves. (a) For each bias [we show $\lambda = -1$ (top row), $\lambda = 1$ (middle row), $\lambda = 2$ (bottom row)] TN-based policies and value functions are produced via actor-critic optimization. These are initiated at random for $L = 4$ with $\chi = 16$ and trained for 10^6 steps. Every 5000 training steps the average reward of the policy is evaluated over 10^4 steps (black squares) and the weights of the policy (which we call a “snapshot” for that time) are stored. The evaluated values can be compared to the training estimate of $r(\pi)$ (red circles), which tends to overestimate $r(\pi)$ initially. The policy snapshot with the highest evaluated r (blue dashed line) is used to initiate the policy for higher values of L . This is repeated every $\Delta L = 2$ up to $L = 50$, with $L = 14$ shown here. (b) For each bias, several policies (here six) are independently trained via the same procedure from different random initial conditions. This produces a distribution of evaluated average rewards, here represented by the median (black squares) and interquartile range (red-shaded region). The policy with the maximum average reward at each L is selected as the optimal dynamics (blue triangles). (c) Same as (a) for $L = 50$. The learning curves appear noisier than in (a) but note that the vertical scale is much smaller. The learning rate is kept fixed throughout. (d) The distribution of r across parallel agents for $L = 50$ is again much tighter than for $L = 14$.

weights after optimization at this size are then used as the initial weights for $L = 6$. This is repeated in steps of $\Delta L = 2$, up to the maximum desired L . This process ensures that effectively much longer training times are used for larger systems, and produces smooth convergence curves in L , which can be used both as diagnostic tools and for extrapolation [cf. Fig. 3(c) inset].

(iii) Policy evaluation and selection: To determine the quality of a policy, we use it to generate trajectories without any change to the policy weights [cf. Fig. 1(c) of the main text]. The set of rewards along these trajectories can then be averaged to estimate $r(\pi_w)$ for the policy, allowing for different policies to be compared.

To ensure that we obtain the best policies possible, we then employ policy selection in two ways. First, throughout training a given policy we store its weights periodically. After some number of periods, these weight snapshots are then evaluated and the best one is selected, ensuring that the policy can only improve with more training. Second, we run parallel policy optimisations and evaluations, starting from different random initial weights, with the best one selected.

The specific processes of policy evaluation and selection used to produce the results in the main text are illustrated for the ASEP in Fig. 4 (details in caption).

-
- [1] S. R. White, *Phys. Rev. Lett.* **69**, 2863 (1992).
 [2] S. R. White, *Phys. Rev. B* **48**, 10345 (1993).
 [3] U. Schollwöck, *Ann. Phys. (Amsterdam)* **326**, 96 (2011).
 [4] J. Eisert, M. Cramer, and M. B. Plenio, *Rev. Mod. Phys.* **82**, 277 (2010).
 [5] J. Eisert, *Model. Simul.* **3**, 520 (2013).
 [6] S. Montangero, *Introduction to Tensor Network Methods* (Springer International Publishing, New York, 2018).
 [7] R. Orús, *Nat. Rev. Phys.* **1**, 538 (2019).
 [8] E. Stoudenmire and D. J. Schwab, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4799–4807, [arXiv:1605.05775](#).
 [9] Y. Liu, X. Zhang, M. Lewenstein, and S.-J. Ran, [arXiv:1803.09111](#).
 [10] Z.-Z. Sun, C. Peng, D. Liu, S.-J. Ran, and G. Su, *Phys. Rev. B* **101**, 075135 (2020).
 [11] C. Roberts, A. Milsted, M. Ganahl, A. Zalcman, B. Fontaine, Y. Zou, J. Hidary, G. Vidal, and S. Leichenauer, [arXiv:1905.01330](#).
 [12] Z. Ghahramani and M. I. Jordan, *Mach. Learn.* **29**, 245 (1997).
 [13] S. Efthymiou, J. Hidary, and S. Leichenauer, [arXiv:1906.06329](#).
 [14] E. M. Stoudenmire, *Quantum Sci. Technol.* **3**, 034003 (2018).
 [15] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, *Phys. Rev. X* **8**, 031012 (2018).
 [16] Z.-F. Gao, S. Cheng, R.-Q. He, Z. Y. Xie, H.-H. Zhao, Z.-Y. Lu, and T. Xiang, *Phys. Rev. Res.* **2**, 023300 (2020).
 [17] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, *Phys. Rev. Lett.* **122**, 065301 (2019).
 [18] C. Guo, Z. Jie, W. Lu, and D. Poletti, *Phys. Rev. E* **98**, 042114 (2018).
 [19] I. Glasser, N. Pancotti, and J. I. Cirac, [arXiv:1806.05964](#).
 [20] I. Glasser, R. Sweke, N. Pancotti, J. Eisert, and I. Cirac, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. Alche-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Vancouver, Canada, 2019), pp. 1496–1508.
 [21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).
 [22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, *Nature (London)* **518**, 529 (2015).
 [23] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, [arXiv:1812.05905](#).
 [24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, *Nature (London)* **529**, 484 (2016).
 [25] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, in *International Conference on Machine Learning* (Proceedings of Machine Learning Research, 2015).
 [26] T. Wang, X. Bao, I. Clavera, J. Hoang, Y. Wen, E. Langlois, S. Zhang, G. Zhang, P. Abbeel, and J. Ba, [arXiv:1907.02057](#).
 [27] F. Metz and M. Bukov, *Nat. Mach. Intell.* **5**, 780 (2023).
 [28] A. Mahajan, M. Samvelyan, L. Mao, V. Makoviychuk, A. Garg, J. Kossaifi, S. Whiteson, Y. Zhu, and A. Anandkumar, [arXiv:2106.00136](#).
 [29] H. Touchette, *Phys. Rep.* **478**, 1 (2009).
 [30] J. P. Garrahan, *J. Stat. Mech.* (2016) 073208.
 [31] J. P. Garrahan, *Physica (Amsterdam)* **504A**, 130 (2018).
 [32] R. L. Jack and P. Sollich, *Eur. Phys. J. Special Topics* **224**, 2351 (2015).
 [33] R. L. Jack, *Eur. Phys. J. B* **93**, 74 (2020).
 [34] R. Chetrite and H. Touchette, *Ann. Henri Poincaré* **16**, 2005 (2015).
 [35] C. Casert, T. Vieijra, S. Whitelam, and I. Tamblyn, *Phys. Rev. Lett.* **127**, 120602 (2021).
 [36] S. N. Majumdar and H. Orland, *J. Stat. Mech.* (2015) P06039.
 [37] R. Chetrite and H. Touchette, *J. Stat. Mech.* (2015) P12001.
 [38] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Annu. Rev. Phys. Chem.* **53**, 291 (2002).
 [39] C. Giardinà, J. Kurchan, and L. Peliti, *Phys. Rev. Lett.* **96**, 120603 (2006).
 [40] C. Giardinà, J. Kurchan, V. Lecomte, and J. Tailleur, *J. Stat. Phys.* **145**, 787 (2011).
 [41] F. Cérou and A. Guyader, *Stoch. Anal. Appl.* **25**, 417 (2007).
 [42] V. Lecomte and J. Tailleur, *J. Stat. Mech.* (2007) P03004.
 [43] S. Whitelam, D. Jacobson, and I. Tamblyn, *J. Chem. Phys.* **153**, 044113 (2020).
 [44] B. De Bruyne, S. N. Majumdar, and G. Schehr, *Phys. Rev. E* **104**, 024117 (2021).

- [45] B. D. Bruyne, S. N. Majumdar, H. Orland, and G. Schehr, *J. Stat. Mech.* (2021) 123204.
- [46] B. D. Bruyne, S. N. Majumdar, and G. Schehr, *J. Phys. A* **54**, 385004 (2021).
- [47] B. De Bruyne, S. N. Majumdar, and G. Schehr, *Phys. Rev. Lett.* **128**, 200603 (2022).
- [48] G. Pozzoli and B. D. Bruyne, *J. Stat. Mech.* (2022) 113205.
- [49] D. C. Rose, J. F. Mair, and J. P. Garrahan, *New J. Phys.* **23**, 013013 (2021).
- [50] A. Das and D. T. Limmer, *J. Chem. Phys.* **151**, 244123 (2019).
- [51] A. Das, D. C. Rose, J. P. Garrahan, and D. T. Limmer, *J. Chem. Phys.* **155**, 134105 (2021).
- [52] A. Das, B. Kuznets-Speck, and D. T. Limmer, *Phys. Rev. Lett.* **128**, 028005 (2022).
- [53] V. S. Borkar, S. Juneja, and A. A. Kherani, *Commun. Inf. Syst.* **3**, 259 (2003).
- [54] G. Ferré and H. Touchette, *J. Stat. Phys.* **172**, 1525 (2018).
- [55] J. Yan, H. Touchette, and G. M. Rotskoff, *Phys. Rev. E* **105**, 024115 (2022).
- [56] J. Yan and G. M. Rotskoff, *J. Chem. Phys.* **157**, 074101 (2022).
- [57] T. Nemoto and S. I. Sasa, *Phys. Rev. Lett.* **112**, 090602 (2014).
- [58] T. Nemoto, R. L. Jack, and V. Lecomte, *Phys. Rev. Lett.* **118**, 115702 (2017).
- [59] H. J. Kappen and H. C. Ruiz, *J. Stat. Phys.* **162**, 1244 (2016).
- [60] P. G. Bolhuis, Z. F. Brotzakis, and B. G. Keller, *arXiv:2207.04558*.
- [61] L. Holdijk, Y. Du, F. Hooft, P. Jaini, B. Ensing, and M. Welling, *arXiv:2207.02149*.
- [62] By obeying detailed balance we specifically mean that a system has a dynamical generator that can be made Hermitian via a similarity transformation defined through the stationary state (i.e., equilibrium) distribution.
- [63] There are two situations where one might simply set $C(a, S) = 1$ for all state-action pairs: (i) when all actions are possible from any state; (ii) the form of the constraint is unknown and must be learned, for example, by penalizing the policy when disallowed actions are selected.
- [64] M. C. Bañuls and J. P. Garrahan, *Phys. Rev. Lett.* **123**, 200601 (2019).
- [65] P. Helms, U. Ray, and Garnet Kin -Lic Chan, *Phys. Rev. E* **100**, 022101 (2019).
- [66] P. Helms and Garnet Kin -Lic Chan, *Phys. Rev. Lett.* **125**, 140601 (2020).
- [67] L. Causer, I. Lesanovsky, M. C. Bañuls, and J. P. Garrahan, *Phys. Rev. E* **102**, 052132 (2020).
- [68] L. Causer, M. C. Bañuls, and J. P. Garrahan, *Phys. Rev. E* **103**, 062144 (2021).
- [69] L. Causer, M. C. Bañuls, and J. P. Garrahan, *Phys. Rev. Lett.* **128**, 090605 (2022).
- [70] L. Causer, J. P. Garrahan, and A. Lamacraft, *Phys. Rev. E* **106**, 014128 (2022).
- [71] M. Fishman, S. White, and E. Stoudenmire, *SciPost Phys. Codebases* **4** (2022), 10.21468/scipostphyscodeb.4.
- [72] DeepMind, Using JAX to accelerate our research (2022), <https://www.deepmind.com/blog/using-jax-to-accelerate-our-research>.
- [73] J. K. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. Santos, R. Perez, C. Horsch, C. Dieffendahl *et al.*, *arXiv:2009.14471*.
- [74] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009), ISBN 1441412697.
- [75] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne *et al.*, JAX: Composable transformations of Python + NumPy programs (2018), <http://github.com/google/jax>.
- [76] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, *Nature (London)* **585**, 357 (2020).
- [77] The Pandas Development Team (2022), 10.5281/zenodo.7093122.
- [78] J. D. Hunter, *Comput. Sci. Eng.* **9**, 90 (2007).
- [79] A. Collette, *Python and HDF5* (O'Reilly Media, Inc., 2013).
- [80] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *SIAM Rev. Soc. Ind. Appl. Math.* **59**, 65 (2017).
- [81] HDF5.jl (2022), <https://github.com/JuliaIO/HDF5.jl>.
- [82] J. Jäckle and S. Eisinger, *Z. Phys. B Condens. Matter* **84**, 115 (1991).
- [83] F. Spitzer, *Adv. Math.* **5**, 246 (1970).
- [84] R. L. Jack and P. Sollich, *Prog. Theor. Phys. Suppl.* **184**, 304 (2010).
- [85] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.132.197301> for details on the implementation of the models (the policy and value function) used in the main text.
- [86] E. Gillman and D. C. Rose, Acten, actor critic with tensor networks, https://github.com/RL-with-TNs/acten_code (2022).