NORTHWESTERN UNIVERSITY

Robot Thermodynamics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Mechanical Engineering

By

Thomas Alejandro Berrueta

EVANSTON, ILLINOIS

September 2024

ABSTRACT

Robot Thermodynamics

Thomas Alejandro Berrueta

The pursuit of precision has been a driving force in engineering since the earliest days of the steam engine. Robotics, born from industrial automation, has embraced this focus. Robots are designed for low tolerances, absolute repeatability, and predictable behavior. Any uncertainty—in the environment, perception, or movement—is seen as a problem to be eliminated. This approach stands in contrast to the messy, unpredictable inner workings of biological organisms. Yet, despite these "flaws," living beings possess a degree of autonomy no machine can match. While precise determinism has its place in engineering, its blind pursuit limits the development of true "life-like" autonomy. This thesis explores a framework that embraces noise and uncertainty as essential tools, rather than obstacles, on the path toward more adaptable, and reliable, autonomous systems.

This thesis proposes design, learning, and control principles for embodied agents with robust, nondeterministic, autonomy. It draws inspiration from (and contributes to the literature of) statistical mechanics and thermodynamics to produce results applicable to nonequilibrium systems such as robots and living organisms. Thermodynamics describes

the flow of energy through matter, and how this flow and its fluctuations can be harnessed to produce work. Analogously, this thesis—titled *Robot Thermodynamics*—investigates how actions are materialized by robot bodies, and how the fluctuations induced by these actions can affect an agent's task-capabilities. In this endeavor, our primary unit of analysis is the *path* or *trajectory distribution*, which describes all possible paths through time and space that an agent can traverse. The structure of an agent's path distribution depends on its physical or material properties, as well as its controller or policy. Exploiting the relationship between agent behavior, embodiment, and decision-making through design, learning, and control is the explicit goal of robot thermodynamics.

This thesis begins by laying the analytical foundations of robot thermodynamics. This mathematical overview serves multiple purposes: First, it introduces the principle of maximum caliber as an inference framework over path distributions. Then, it illustrates how these inferred path distributions and their properties can be used to characterize and manipulate the dynamics of complex systems. Lastly, it describes how optimal control and reinforcement learning can be framed as operations applied onto an agent's path distribution. The thesis then proceeds by demonstrating the power of this approach in several different applications across length-scales—prediction and control of nonequilibrium collectives, design of energy-harvesting colloidal microparticles, and embodied reinforcement learning—each advancing the state-of-the-art in their respective fields. Taken together, the results in this thesis highlight the promise of noise and uncertainty as versatile tools in the development of robust, life-like, real-world autonomy.

Acknowledgements

Throughout the entirety of my Ph.D., I explored intersections of robotics with a staggering (and potentially inadvisable) number of fields—statistical physics, chemical engineering, stochastic processes, materials science, physical therapy, etc.—across experimental platforms that span orders of magnitudes in scale—exoskeletons, pHRI interfaces, robot swarms, microrobots, and robotic pool noodles among others. Over the years, the only limitations I ever felt imposed on me were bans on use of flowery language. *Thus*, I would like to thank my advisor Prof. Todd Murphey for giving me the freedom and encouragement to chase my own wild ambitions—and for many more reasons than I am able to make room for in this page. Todd, your enthusiasm, empathy, and support has made the time spent working on my doctorate the best time of my life, and for that I will be forever grateful. I am certain that it will take me years to fully absorb all of the thoughtful advice and wisdom you have shared with me.

I would also like to thank my committee, Profs. Malcolm MacIver and Daniel Goldman for their inspiration and guidance. Malcolm, your intellectual curiosity and earnest embrace of interdisciplinarity has always been an inspiration to me, and for that I am most grateful. Dan, I will never forget the first time you asked me "so why does this matter?" in a meeting—this moment irreversibly shifted my approach to science and engineering for the better. Over the course of our many collaborations, you have taught me the importance of asking the right questions and of telling a good story, which I am very

grateful for. Additionally, despite not being a member of my committee, I would like to thank Prof. Ryan Truby. Ryan, I will always be thankful for your uncompromising support and willingness to help out at any moment. Your advice has been instrumental to my professional development and navigating the academic landscape, and I cannot wait for the next time our paths cross.

Thanks to my involvement in a multi-year-long a MURI project on the foundations of algorithmic matter, I also greatly benefited from the mentorship and guidance of Profs. Michael Strano, Dana Randall, Jeremy England, and Andréa Richa. Michael, I am grateful to have learned from your commitment to rigor, excellence, and disruptive innovation in science and engineering. Dana, I have always found your ability to simultaneously reason about technical details and the bigger picture inspiring. Jeremy, you taught me that the best science questions its own philosophical foundations, a lesson that I carry with me to this day. Andréa, your uncanny ability to formalize and model physical and biological phenomena is something I hope to have learned from. In addition to the professors involved with the MURI, I also interacted closely with many brilliant students. I am only able to name some of you in this page, but, if you were not mentioned, know that I am still very grateful to have met you.

The work in this dissertation would not have been possible without the help of my incredible academic collaborators. While the full list of collaborators I had the pleasure to work with looks like the author list of an experimental particle physics paper, I would like to highlight a few of my closest peers whose company had a big impact on me. Pavel, the time we hosted you at Northwestern was one of the most impactful periods of my doctorate. I admire your limitless intellectual ambitions, unabashed creativity, and

open-mindedness—no idea is too big for you, and I would be lucky if some of your courage rubbed off on me. Jingfan, there are very few people that I think of as highly professionally. Somehow, in spite of our vastly different academic backgrounds, working with you was one of the easiest and natural collaborations I have ever done. It was truly a joy to work with you, and I am thankful for your tenacious, no-nonsense, and almost supernatural ability to get things done. Allie, you embody rigor itself. So much so, that most of the time I spent working with you I did not believe my own results until you told me that you believed in them yourself. I struggle to think of an academic partnership that was as intellectually productive as ours: In one way or another, our academic disagreements always pushed us to greater heights, and for that I am very grateful. Lastly, Jamison, your empathy and wit—in technical matters as well as in comedic banter—has made it a pleasure to work closely with you in recent years.

Across the 7 years of my Ph.D., I was also very lucky to have incredible labmates as collaborators, mentors, and close friends—only a few of which I can actually fit on this page. Katie, I will never forget how after a single (very eventful) night of karaoke you invited me to your wedding even though we had only just met (shoutout to Emery too). You welcomed me to the lab with open arms, and very patiently offered me guidance and help whenever I needed it. Ola, my memories of your electrifying energy and presence are some of the best I have. I am also grateful for your never-ending willingness to question and be questioned, which made for a ton of fun and wild discussions. Ana, that stretch of three or so years when we were each other's academic partners-in-crime are some of my most cherished memories. I am very thankful for your near-infinite empathy and willingness to engage with me on wacky (or B.S.) philosophical discussions (apparently

about rocks??). Lastly, Annalisa, it is simultaneously really hard and really easy to believe that just over the course of a few years we grew close enough for you to join my wedding party as the best (and worst) woman. You are one of the few people on this page that I have to extend a special thank you to for just being willing to deal with me, and that is high praise.

Most importantly, I want to acknowledge the support of my friends and family. Sam, the relief and comfort that you have been able to offer by always being there to hear me out is hard to fully express. Cory, I know that I can always count on you to make home feel like home. To my parents Alejandro and Isabel, your love and support has been the bedrock of my day-to-day. I owe everything to you, and every day I wake up thinking of how lucky I am to have you. The other person I must extend a special thank you to for being willing to deal with me is my wonderful wife Sami, without whom I would be unanchored, unmoored, and incomplete. I will admit that getting married to you might have also played a role in making this chapter of my life my favorite one yet, and I am eternally grateful that you had the courage to follow me to Chicago for this crazy adventure.

Lastly, I want to acknowledge the funding sources that supported the work, without which this work would not have been possible: US Army Research Office MURI grant W911NF-19-1-0233, US Office of Naval Research grant N00014-21-1-2706, National Science Foundation grant CBET-1637764, and the Northwestern University Presidential Fellowship. I also want to acknowledge hardware loans and technical support from Intel Corporation, whose equipment facilitated the reinforcement learning elements of this thesis.

Table of Contents

List of Figures

a random walk on a large random graph (1000 states), where P_{ss}—the probability at a state—is approximately given by $\mathcal{E}$—that state's exit rate. (C) shows an active matter system of shape-changing agents: an enclosed ensemble of 15 "smarticles" in simulation. (D) realizes similar agents experimentally with an enclosed three-robot smarticle ensemble. The middle row shows that relaxation to the steady-state of a uniform initial distribution is accompanied by monotonic decay in the average rattling value in all cases—analogous to free energy in equilibrium systems. The bottom row shows the validity of the nonequilibrium Boltzmann-like principle in Eq. 3.3, where the black lines in (A, B, and C) illustrate the theoretical correlation slope for a sufficiently large and complex system (see supplementary materials). The mesoscopic regime in (D) provides the most stringent test of rattling theory (where we observe deviations in γ from 1), while also exhibiting global self-organization. In (A and B, middle) time units are arbitrary, and for (C and D, middle) time is in seconds, where the drive period is 2 s. 114

3.2 **Self-organization in a smarticle robotic ensemble**. (A) Front, back and top view of a single smarticle. Of its five degrees of freedom, we consider the time-varying arm angles (α_1, α_2) as "external" driving, since these are controlled by a pre-programmed microcontroller, while the robot coordinates (x, y, θ) are seen as "internal" system configuration, since these respond interdependently to the arms. (B) An example periodic arm motion pattern. (C) Top view of three smarticles confined

serves as an upper bound for probabilities of other configurations, which are predicted by dotted curves (with fitting parameter $\gamma = 3.1$). (B) illustrates the robustness of the relation between probability and rattling as friction in the system is changed. This, along with panel A, shows that measuring the overdamped dynamics $\tau = 0$ is sufficient to predict system behaviors for all lower friction values. 126

4.1 **Emergence of chemomechanical microparticle self-oscillation.** (A), Schematic of a self-limited system of a single particle resting still at the air-liquid interface of a H_2O_2 drop. The particle is composed of a catalytic patch of Pt (yellow) underneath a polymeric disc (blue). The O_2 formation slows down asymptotically over time as the gas bubble restricts the available catalytic surface area. (B), A 2-particle system, in contrast, exhibits an emergent and self-sustained beating behaviour as the bubble merger restores the previously hindered reactivity, thus disrupting the equilibrium state. (C),(D), Micrograph sequence (in (C)) and tracked particle coordinates (in (D)) of a 1-particle system that remains still for an extended period of time. (E),(F), Micrograph sequence (in (E)) and tracked coordinates (in (F)) of a 2-particle system with emergent beating. The breathing radius, $r(t)$, is the distance from the collective's centroid to each particle, averaged over all particles. (G), The long-term breathing radius trajectory of the same system as in (E) and (F) demonstrates the robustness of the beating behaviour. The shaded portion is magnified in the right panel, where mechanistic model simulations (black) are shown

observe that the breathing radius of a homogeneous $N = 8$ system is not periodic. (D), Asymmetry-induced order across N predicted by Rattling Theory. A quantification of collective disorder, the system's Rattling $\mathcal{R}$ is predicted to be lower (i.e. more orderly) if the relative burst intensity of one particle is increased beyond or decreased below 1x, which signifies homogeneity. This is experimentally realized by modulating the Pt patch size on a "designated leader" (DL) particle relative to the others. The curves are offset to make all $\mathcal{R} = 0$ at 1x intensity to highlight the effect of system heterogeneity on Rattling. (E), Same as (B), but for heterogeneous systems of equal particle numbers, where the DL broke the permutation symmetry. In contrast to the homogeneous systems (B), they remain robustly periodic across N. It is important to recognize that the polymeric disc size of a DL is unchanged. Scale bar, 500μm. (F), Breathing radius for an 8-particle DL system (i.e., $N = 7 + 1$DL), which reliably beats periodically. The period of 14.2s extracted from $r(t)$ coincides with the most probable interarrival time in ((E), bottom). 139

4.3 **Rattling as a function of patch size in diffusive model.** Here, we study the effect of a given particle's U parameter (in analogy to Pt patch size) on the rattling of collectives of varying sizes. Note that we subtract the constant offset in rattling due to system size so that $\mathcal{R} = 0$ at $U = 0$ for all N. We find that any variability in the size of the particle's patch produces a drop in rattling, leading to asymmetry-induced order. When a particle becomes inert as U increases, it stops contributing to system-level

dynamics simple enough to explicitly write down and whose policy admits a globally optimal analytical solution. The system's 4-dimensional state space is comprised of its planar positions and velocities. We parametrize its controllability through $\beta \in [0, 1]$, where $\beta = 0$ produces a formally uncontrollable system. The task is to translate the point mass from p_0 to p_g within a fixed number of steps at different values of β, and the reward is specified by the negative squared Euclidean distance between the agent's state and the goal. We compare state-of-the-art model-based and model-free algorithms, NN-MPPI and SAC respectively, to our proposed maximum diffusion (MaxDiff) RL framework. (B),(D), Representative snapshots of MaxDiff RL, NN-MPPI, and SAC agents (top to bottom) in well-conditioned ($\beta = 1$) and poorly-conditioned ($\beta = 0.001$) controllability settings. (C), Even in this simple system, poor controllability can break the performance of RL agents. As $\beta \to 0$ the system's ability to move in the x-direction diminishes, hindering the performance of NN-MPPI and SAC, while MaxDiff RL remains task-capable. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test.

we examine the terminal returns of swimmer agents (10 seeds each) across values of α with $m_s = 1$. Diffusive exploration leads to greater returns until a critical point (inset dotted line), after which the agent starts valuing diffusing more than accomplishing the task. (C), Using $\alpha = 100$, we compared MaxDiff RL against SAC and NN-MPPI with $m_s = 0.1$. We observe that MaxDiff RL outperforms comparisons on average with near-zero variability across random seeds, which is a formal property of MaxDiff RL agents. For all reward curves, the shaded regions correspond to the standard deviation from the mean across 10 seeds. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test. 179

5.4 **Trained system embodiment determines deployed system performance**. (A), Two variants of the MuJoCo swimmer environment: One with $m_s = 1$ and one with $m_s = 0.1$. As a baseline, we deploy learned representations on the same swimmer variant trained on. Then, we carry out a transfer experiment where the trained and deployed swimmer variants are swapped. (B), Baseline experiments confirm previous results: All algorithms benefit from a more controllable swimmer. (C), Both NN-MPPI and SAC performance degrades when deployed on

a more controllable system than was trained on, which is undesirable. In contrast, MaxDiff RL benefits from the "Heavy-to-Light" transfer and we also observe that MaxDiff RL performance further increases in the "Light-to-Heavy" transfer experiment. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test. 186

5.5 **Maximally diffusive RL agents are capable of single-shot learning.** (A), Illustration of MuJoCo ant environment. (B), Typical algorithms learn across many different initializations and deployments of an agent, which is known as multi-shot learning. In contrast, single-shot learning insists on a single task attempt, which requires learning through continuous deployments. Here, we prove that MaxDiff RL agents are equivalently capable of single-shot and multi-shot learning in a broad variety of settings. (C), Single-shot learning depends on the ability to generate data samples ergodically, which MaxDiff RL guarantees when there are no irreversible state transitions in the environment. (D), Single-shot learning in the swimmer MuJoCo environment. We find that MaxDiff RL achieves robust performance comparable to its multi-shot counterpart. (E), In contrast to the swimmer, the MuJoCo ant environment contains irreversible state transitions (e.g., flipping

upside down) preventing ergodic trajectories. Nonetheless, MaxDiff RL remains state-of-the-art in single-shot learning. Note that we report returns over a window of 1000 steps in analogy to our multi-shot results, where episodes consist of 1000 environment interactions. For all reward curves, the shaded regions correspond to the standard deviation from the mean across 10 seeds. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean and the data distribution is plotted directly ($n = 10$ seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test. 189

CHAPTER 1

Introduction

The field of thermodynamics has been concerned with technological innovation from its very inception [1]. Unlike most natural sciences, early progress in thermodynamics was propelled by inventions like the steam engine, which served as a sort of vessel or "engineering substrate" for scientific discovery. This is evident in the fact that the same research aimed at understanding and improving the efficiency of heat engines [2] also established the foundation for such fundamental concepts as entropy and the second law of thermodynamics [3]. Since then, thermodynamics has maintained a symbiotic relationship with engineering, providing a theoretical framework for harnessing the inherent randomness and *nondeterminism* of molecular motion to achieve useful work and accomplish tasks. This powerful synergy fueled the Industrial Revolution, which, ironically, created a socioeconomic demand for increasingly precise, predictable, and efficient machinery [4]. As a result, the pursuit of low tolerances, repeatability, and *determinism*—often referred to as "precision engineering"—became the prevailing philosophy across engineering disciplines.

Robotics, itself emerging from industrial automation, has embraced this doctrine from its very beginnings [5]. However, this pursuit of determinism has led to a fundamental tension in robotics. While robots are often designed with the goal of precise and repeatable behavior, the real world is ever-changing, uncertain, and unpredictable. Robots must constantly interact with complex environments, cope with unexpected disturbances, and

make decisions in the face of incomplete information. This mismatch between the deterministic ideals of precision engineering and the stochastic nature of real-world operation has created a significant challenge for the development of truly autonomous and adaptable robotic systems.

In this work, we argue that embracing uncertainty, rather than fighting against it, is essential for advancing the field of robotics. By drawing inspiration from the principles of thermodynamics, we can develop new approaches to robot design, learning, and control that explicitly account for—and exploit—inherent randomness and uncertainty in the real-world. This shift in perspective opens up exciting possibilities for creating robots that are not only more robust and adaptable but also capable of exhibiting emergent behaviors and self-organization, much like the complex systems found in nature.

1.1. Main Contributions

This thesis embraces uncertainty and nondeterminism in the search for novel design, learning, and control principles for embodied autonomy. To this end, it presents a framework grounded in the statistical physics of the principle of maximum caliber, which we refer to as *robot thermodynamics*. Throughout the chapters of this thesis, we develop new predictive principles, propose alternative design strategies, and derive novel learning and control methods, each advancing the state-of-the-art in their respective areas of robotics. We explore these questions through extensive theory-crafting, substantial simulation-development, and exhaustive experimental validation. *By framing design, learning, and control problems in terms of distributions of stochastic trajectories, this thesis develops*

methods for parsimoniously reasoning about agent embodiment and decision-making as two sides of the same coin—and, in doing so, it paves the way towards more life-like autonomy.

1.2. Thesis Outline

The main contents of the thesis are subdivided into the following four chapters, the first of which presents the theoretical backbone for most results discussed in the latter chapters. We conclude this thesis with an additional chapter that discusses future potential directions for this body of work.

1.2.1. Foundations of Robot Thermodynamics

In this chapter, we lay down the mathematical foundations of the robot thermodynamics framework. Many concepts are defined and given an introductory treatment prior to their use in future chapters. In particular, we introduce the primary mathematical object of study throughout this thesis—the path (or trajectory) distribution. We proceed by outlining procedures to *infer* the path distributions of complex systems through the principle of maximum caliber. Then, we show how these distributions can be used to make nontrivial predictions about complex system behavior, such as steady-state occupancy statistics. Lastly, we outline procedures for *reshaping* path distributions via control and policy optimization, which allow us to recover canonical results in optimal control and reinforcement learning from the perspective of robot thermodynamics.

The contributions of this chapter are the following:

(1) We present a methodology for inference and synthesis of agent behavior based on trajectory distributions and their properties, which we refer to as *robot thermodynamics*.

(2) We provide an original, unpublished derivation of the steady-state occupancy statistics of a broad class of stochastic processes that recovers the low-rattling selection principle.

(3) We present an original, unpublished derivation of Pontryagin's maximum principle from the principle of maximum caliber, and establish connections between KL-control and stochastic optimal control.

The work in this chapter is a combination of original, unpublished results, as well as results previously published in [6] and [7].

1.2.2. Predicting Self-Organization in Active and Robotic Matter

In this chapter, we explore the first application of our framework: We reinterpret our steady-state occupancy predictions made in the previous chapter from the perspective of statistical mechanics in order to develop an understanding of self-organization in far-from-equilibrium systems. As such, this chapter motivates *rattling theory* and proposes it as a candidate explanation for many emergent phenomena in broad classes of systems. We experimentally validate the theory's predictions in a robotic active matter system of "smarticles"—originally introduced by [8]—whose re-design from the ground up into a platform capable of distributed control is a contribution of this thesis. Lastly, we provide proof-of-concept demonstrations of control strategies based on the principles of rattling theory by manipulating drive entropy and frictional interactions.

The contributions of this chapter are the following:

(1) We present a novel theory of self-organization in nonequilibrium statistical mechanics, which we refer to as *rattling theory*.
(2) We introduce and experimentally validate a Boltzmann-like principle for predicting the steady-state behavior of complex systems.
(3) We outline multiple methodologies for engineering and designing desired nonequilibrium steady-state behaviors in complex physical systems.

The work in this chapter consists of results largely published in [6]. As a result, we acknowledge the contributions of Pavel Chvykov to this work and to the theory in particular, whose initial developments precede my involvement in this intellectual project. This work was done in collaboration with the groups of Drs. Jeremy England, Daniel Goldman, and Kurt Wiesenfeld.

1.2.3. Designing for Emergence in Robotic Microsystems

This chapter explores the first application of robot thermodynamics beyond inference. We explore how the design parameters of complex systems can be used to reshape their path distributions, resulting in novel behaviors. To this end, we use our framework to analyze the complex dynamics of an active collective of colloidal microparticles, developing a theoretical model that captures their most salient features. Using this model, we optimize system design parameters in hopes of inducing self-organized states in the system dynamics. Then, we implement these parameters in an experimental instantiation of the system, confirming the emergence of self-organized self-oscillations. Lastly, as a proof-of-concept

demonstration, we illustrate how these self-organized states can be exploited towards microrobotic task-capabilities by using them to power microrobot arms.

The contributions of this chapter are the following:

(1) We analyze and model from first-principles the complex dynamics of an active collective of colloidal microparticles.

(2) We present a novel thermodynamic mechanism for *asymmetry-induced order* in complex systems grounded in rattling theory.

(3) We exploit self-organization as a means of generating self-oscillating electrical currents aboard a microparticle collective, which we demonstrate by powering a state-of-the-art microrobot arm.

The work in this chapter consists of results largely published in [9]. As a result, we acknowledge the contributions of Jingfan Yang, whose experimental design and microfabrication expertise made this project feasible at all. This work was done in collaboration with the groups of Drs. Michael Strano and Marc Miskin.

1.2.4. Overcoming Temporal Correlations in Robot Learning

In this chapter, we fully realize the promise of robot thermodynamics as an inference and *synthesis* framework for embodied agents. We discuss the challenges that embodied reinforcement learning agents face, such as violations of the *i.i.d.* property during data acquisition. Then, we introduce maximum diffusion reinforcement learning (MaxDiff RL) in order to overcome these limitations. The chapter proceeds by demonstrating that MaxDiff RL agents are capable of reliable performance across benchmarks with provable performance guarantees. Lastly, we also prove that MaxDiff RL agents are capable of

learning in single-shot deployments, which is essential to real-world deployment of deep reinforcement learning solutions.

The contributions of this chapter are the following:

(1) We present a novel reinforcement learning framework built from the ground up with embodied agents in mind, which we refer to as *MaxDiff RL*.

(2) We prove that MaxDiff RL is a generalization of the state-of-the-art MaxEnt RL framework, and prove that MaxDiff RL agents are robust to initializations and are capable of learning in single-shot deployments.

(3) We perform extensive empirical evaluations and achieve state-of-the-art performance in spatial navigation robotics benchmarks in RL.

The work in this chapter consists of results largely published in [7]. As a result, we acknowledge the contributions of Allison Pinosky, whose algorithmic genius provided a rock solid foundation for our exploration of interesting questions in robot learning.

CHAPTER 2

Foundations of Robot Thermodynamics

Few fields have had as foundational an impact on the progress of science as statistical mechanics, without which the quantum revolution could not have taken place. However, even within the seemingly deterministic realm of classical physics, statistical mechanics has revealed that true certainty is elusive [10]. Instead of individual particles following predictable paths, statistical mechanics showed that their behavior is often best understood through probabilities and distributions, introducing a fundamental uncertainty into our understanding of the classical world. Much of what drives its predictive power is that it describes deeper truths than those immanent in the laws of physics. In a sense, statistical mechanics is simply concerned with deriving mathematical truths about statistical populations and their constituents. This is the groundbreaking realization first made by E. T. Jaynes in 1957 [11]—that statistical mechanics can be largely understood as a general probabilistic inference framework applied to the analysis of physical systems.

Here, I aim to extend this powerful framework beyond inference and into *synthesis*, forming the core of what I term *Robot Thermodynamics*. This shift represents a fundamental transformation; by considering synthesis, we unlock the potential to both model and specify goal-oriented behavior. At its heart, robot thermodynamics poses two key questions: "What is the structure of an optimal agent's dynamics?" and "How can such dynamics be realized?" To address these questions, this chapter will cover the representation of agent dynamics through path distributions, the inference of goal-directed behavior under

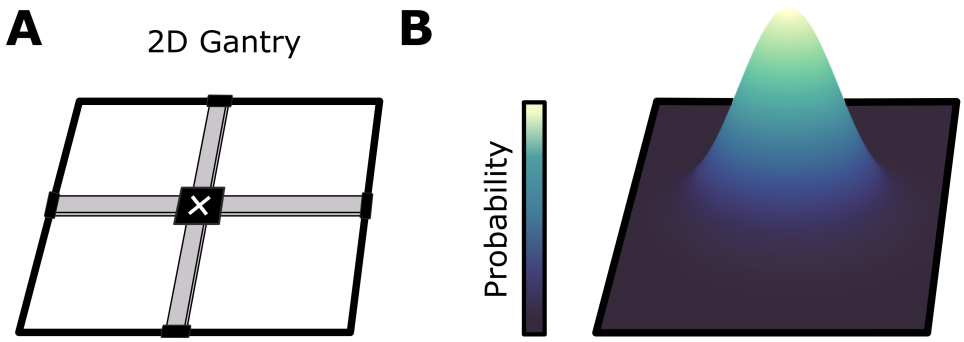


Figure 2.1. **Example system: 2D gantry.** (A) The state-space of the system consists of all possible coordinates in the plane within the rectangle described by the gantry frame. The gantry motors control the horizontal and vertical position of the end-effector via either position or velocity commands. (B) When gantry positions are sampled from a normal distribution at the center of the frame, its states are normally distributed as well.

dynamical constraints, and the synthesis of policies and controllers capable of achieving such behavior after covering some preliminaries. This chapter carefully builds up the mathematical vocabulary that will span the entirety of this dissertation, as such its contributions are many and diverse. These contributions will lead to design, learning, and control principles that will form the mathematical basis for the applications explored in future chapters.

2.1. States, Paths, and their Probabilities

In this section, we will cover some mathematical preliminaries leading to our discussion of path distributions and how we can use them to describe the behavior of robots and complex dynamical systems. The crux of this approach is to model agent experiences as random events constrained by their tasks and physical embodiment. To motivate this perspective in a robotics context, we will consider a 2D gantry as an example system that will guide us throughout the introduction of preliminary concepts (see Fig. 2.1).

2.1.1. Random Variables

To model dynamical systems and their behavior in time, we first need to define their states and the properties of its state-space. For the system in Fig. 2.1(A), we can capture its behavior by tracking the the planar coordinates of its end-effector, $p_x(t), p_y(t)$, at each point in time, t, which we refer to as its state, $x(t) = [p_x(t), p_y(t)]^T$. At all points in time over some time interval, the system's states take value in a state-space determined by the dimensions of the gantry frame. Expressed more formally, $\forall t \in \mathcal{T} \subset \mathbb{R}^+$, we have that $x(t) \in \mathcal{X}$, where $\mathcal{X} = \{(p_x, p_y) \in [-L, L] \times [-L, L]\}$ and $L \in \mathbb{R}^+$. This state-space has some properties of interest: $\mathcal{X}$ is *compact* and simply *connected*, which implies that it is of finite volume and that every state is accessible from every state. It is important to note, however, that the state-spaces of more general systems do not often have these properties.

Since the focus of this work is on embodiment rather than perception, agent "experiences" refer to the states a system can find itself in over time. Equipped with a state-space, we may now begin to model agent experiences as collections of *random variables*. Random variables are functions that map from a sample-space, Ω, onto a measurable state-space, $\mathcal{X}$. To satisfy the axioms of probability, the sample-space must be a part of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sample-space Ω is the space of possible variable outcomes, whereas the state-space $\mathcal{X}$ represents the values one can measure for each of these outcomes. The event-space $\mathcal{F}$ is a Borel σ-algebra, which is a collection of all possible subsets of random variable outcomes. Lastly, $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure that describes the likelihood of any event [12]. To assess the likelihood of individual samples in the state-space, the concept of probability density functions can be useful. Probability density functions $p : \mathcal{X} \to [0, \infty)$ are functions that satisfy the following relationship for any given

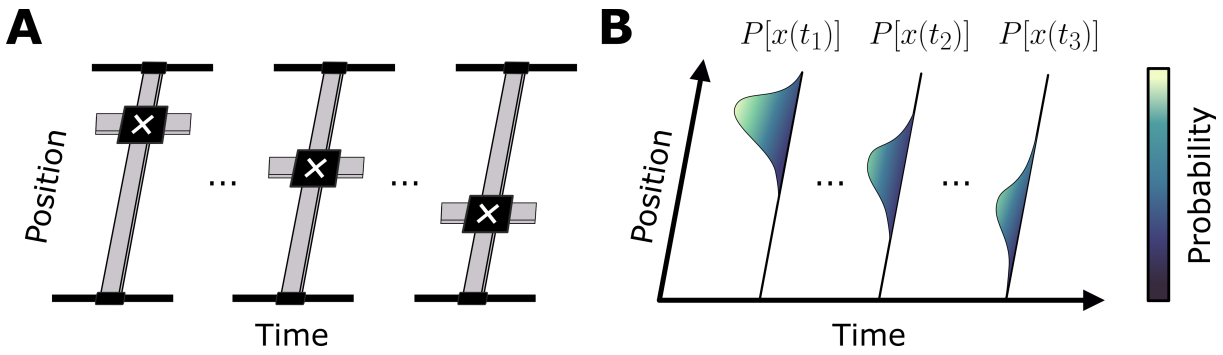


Figure 2.2. **Sample path distribution for the gantry system.** (A) Snapshots of sampled gantry states. The horizontal position of the gantry end-effector is held constant while the vertical position is stochastically sampled from the system's path distribution at three points in time. (B) Temporal cross-sections of the 2D gantry system's path distribution, $P[x(t)]$. At each point in time, the spatial distribution of gantry experiences varies.

$A \in \mathcal{X}$:

$$\mathbb{P}(X^{-1}(A)) = \int_A p(x)dx, \tag{2.1}$$

where $X^{-1} : \mathcal{X} \to \Omega$ is assumed to exist. In the context of the 2D gantry example, consider providing random position commands to the system motors. If we model our gantry dynamics as $x(t) = u(t)$ with $u(t) \sim \mathcal{N}(0, \mathrm{Id})$, then the robot's experiences can be modelled by a normally distributed random variable, as in Fig. 2.1(B). Alternatively, consider the possibility there is actuation noise or uncertainty. Then, we may model the gantry as $x(t) = u(t) + \xi$, where $\xi \sim \mathcal{N}(0, \mathrm{Id})$, realizing the same statistics as before when $u(t) = 0$. As long as $u(t)$ is constant, the gantry's experiences may be modelled by a random variable.

2.1.2. Stochastic Processes and Stochastic Control Processes

As soon as the system's experiences become time-varying, we can no longer model them effectively as a single random variable. In the previous illustration, uncertainty entered the system directly through its states. As a result, the effect of noise on the system was static and time-invariant, which allowed us to model agent experiences with a random variable. However, this changes the moment we consider using velocity commands to control the gantry. For example, consider the gantry dynamics in the presence of noisy velocity commands, $\dot{x}(t) = u(t) + \xi$, where we use Langevin notation for simplicity with $\xi \sim \mathcal{N}(0, \Sigma)$. Now, let $u(t)$ be given by some constant velocity, $c = [0, v_y]^T$, taken in the negative y-direction. If we ignore boundary conditions and let $x(t_i) = [0, L]^T$, then the gantry's experiences are distributed in time according to $x(t) \sim \mathcal{N}(x(t_i) - ct, t\Sigma)$ for all $t \in [t_i, t_f] \subset \mathbb{R}^+$, as illustrated in Fig. 2.2. Crucially, the agent's experiences can no longer be described by a single random variable, but rather by a collection of random variables—each distributed according to a different law at each moment in time. Thus, in this more general setting we must now model gantry experiences as *stochastic processes* due to their time-dependent nature. Adapting the definition in [13], stochastic processes are formally defined in the following way.

Definition 2.1. *(Stochastic Process) A stochastic process is a collection of random variables parametrized by a totally ordered indexing set* $\mathcal{T}$*,*

$$\{X_t\}_{t\in\mathcal{T}} \text{ when } \mathcal{T} \text{ is discrete, or } \{X(t)\}_{t\in\mathcal{T}} \text{ when } \mathcal{T} \text{ is continuous,}$$

defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$*. The sample space* Ω *is measurable and equipped with a Borel* σ*-algebra* $\mathcal{F}$*, as well as a probability measure* $\mathbb{P}$*. Each random variable takes value in a measurable state-space* $\mathcal{X}$*, and each sample path of the stochastic process takes value in a measurable space* $\mathcal{X}^{\mathcal{T}}$ *with Borel* σ*-algebra* $\mathcal{B}(\mathcal{X}^{\mathcal{T}})$*.*

In other words, stochastic processes are families of random variables indexed according to some "time-like" set, $\mathcal{T}$. Throughout this thesis, we take the state-space to be some subset of the reals, $\mathcal{X} \subset \mathbb{R}^d$. Each individual realization of the stochastic process, $\omega \in \Omega$, results in a different sample path, which we typically notate as $x_{\mathcal{T}}(\omega) = \{X(t, \omega)\}_{t \in \mathcal{T}}$. When considering processes that are continuous in time, we often take $\mathcal{T}$ to be the halfline or an interval of the reals given by some initial and final time, $\mathcal{T} = [t_i, t_f] \subset \mathbb{R}^+$. When $\mathcal{T}$ is discrete, which throughout this thesis we take to mean $\mathcal{T} = \{1, \cdots, T\}$, we write $x_{\mathcal{T}}(\omega) = \{X_t(\omega)\}_{t \in \mathcal{T}}$ instead. We note that sometimes we omit the ω from our notation for defining stochastic processes for simplicity, but the dependence on ω is always implicit.

Since we are interested in describing agent experiences directly, and not just the probabilities of events $\omega \in \Omega$, we must be able to describe the measures of sets of states and state trajectories. To this end, we let $\mathbb{P}(x_{\mathcal{T}} \in A) = \mathbb{P}(x_{\mathcal{T}}^{-1}(A))$ for any given $A \subset \mathcal{X}^{\mathcal{T}}$, where we note that $x_{\mathcal{T}}^{-1} : \mathcal{X}^{\mathcal{T}} \to \Omega$. Then, to describe the likelihoods of individual sample paths, we must define—and assume the existence of—a probability density function. To this end, for each ω we use $x(t) = x_{\mathcal{T}}(\omega) \in \mathcal{X}^{\mathcal{T}}$ to denote individual realizations of the stochastic process, and refer to $x(t)$ as its experiences, *state trajectories*, or *paths*. When $\mathcal{T}$ is discrete and finite, e.g., $\{1, \cdots, T\}$, we use $x_{1:T} = x_{\mathcal{T}}(\omega) \in \mathcal{X}^{\mathcal{T}}$ instead. Then, the probability density function associated with the measure is given by $P : \mathcal{X}^{\mathcal{T}} \to [0, \infty)$,

such that

$$\mathbb{P}(x_{\mathcal{T}} \in A) = \mathbb{P}(x_{\mathcal{T}}^{-1}(A)) = \int_{x_{\mathcal{T}}^{-1}(A)} d\mathbb{P}(\omega) = \int_A P[x(t)]\mathcal{D}x(t) \tag{2.2}$$

$$= \int_A P[x_{1:T}]\mathcal{D}x_{1:T}, \tag{2.3}$$

where $\mathcal{D}x(t)$ and $\mathcal{D}x_{1:T}$ denotes integration over sample paths depending on whether $\mathcal{T}$ is continuous or discrete, in line with the Feynman path integral formalism [14]. Thus, we will refer to this density over paths as the *path* or *trajectory distribution*, and use $P[x(t)]$ or $P[x_{1:T}]$ to express the probability density of a given state trajectory. In this formalism, we also have a natural way to express expected values. Consider some real-valued function $f(\cdot)$ of $x_{\mathcal{T}}$, then we define its expectation over sample paths as

$$E[f(x_{\mathcal{T}})] = \int_\Omega f(x_{\mathcal{T}}(\omega))d\mathbb{P}(\omega) = \int_{\mathcal{X}^{\mathcal{T}}} P[x(t)]f(x(t))\mathcal{D}x(t), \tag{2.4}$$

which is consistent with our above definition of probability densities over state trajectories.

An important note is that Definition 2.1 merely states that stochastic processes are parametrized collections of random variables. As such, our definition does not provide or rely on any information regarding the dynamics or laws describing the time-evolution of the random variables forming a part of said process. As we saw in our examples in Figs. 2.1 and 2.2, the dynamics of the underlying process and its properties are essential to modelling an agent's experiences. Moreover, in the context of control systems, we saw in Fig. 2.2 that the choice of controller also plays an important role in the structure of the agent's path distribution and their experiences. In general, we model controllers as functions, $u(t) : \mathcal{T} \to \mathcal{U}$, that produce control inputs to the system at each point in time,

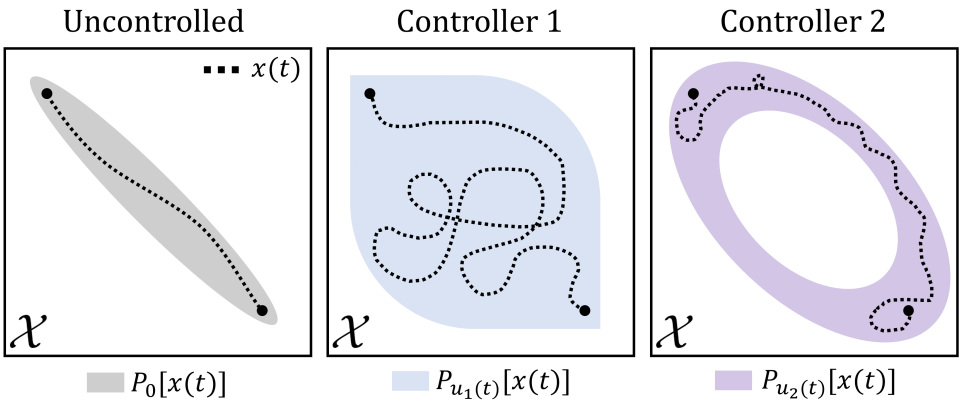


Figure 2.3. **Effect of controllers on the sample path distribution of stochastic control processes**. (left) Sample path and support of the probability density over the paths of an autonomous stochastic process (i.e., with null controller "0"). (middle and right) Sample paths and distributions induced by two distinct controllers $u_1(t)$ and $u_2(t)$. Here, we illustrate that depending on the nature of the controller the distribution over sample paths can be nontrivial. Note that we do not illustrate the values of the probability densities, only their support.

where $\mathcal{U} \subset \mathbb{R}^m$. Clearly, depending on the nature of the underlying controller and the system dynamics, the underlying density over system trajectories can vary strongly (see Fig. 2.3). To reflect this, we extend our definition of stochastic processes as follows.

Definition 2.2. *(Stochastic Control Process) A stochastic control process is a stochastic process (Definition 2.1) on a probability space* $(\Omega, \mathcal{F}, \mathbb{P}_{u(t)})$*, with indexing set* $\mathcal{T}$*, where sample paths take value in a measurable space* $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}(\mathcal{X}^{\mathcal{T}}))$*, and the measure and its resulting density* $P_{u(t)} : \mathcal{X}^{\mathcal{T}} \to [0, \infty)$ *are parametrized by a controller* $u(t) : \mathcal{T} \to \mathcal{U}$*.*

Thus, we think of control systems as stochastic processes that are parametrized by their controllers, or equivalently as a collection of distinct stochastic processes for each choice of controller. As before, this definition does not capture *how* each random variable is affected by a choice of controller, but it does acknowledge their dependence on a controller.

2.1.3. Markov Processes and Markov Decision Processes

Neither of the definitions of stochastic processes provided thus far explicitly describe the dynamics underlying the time-evolution of their random variables. This is because doing so requires characterizing the transition structure and dynamical laws governing the stochastic process, which can be challenging to do in general. One particular class of stochastic processes where this is doable is discrete-time Markov processes, which are prevalent across applications of algorithmic decision-making.

Discrete-time Markov processes are stochastic processes (in the sense of Definition 2.1) with the additional requirement that the probability of any given event is only dependent on the state attained in the previous event. As a result, the probability density function associated with a discrete-time Markov process' probability measure is conditionally factorable across time increments. To see this concretely, consider a stochastic process $\{X_t\}_{t\in\mathcal{T}}$ with $\mathcal{T} = \{1, \cdots, T\}$ and a given initial condition x_1 distributed according to $x_1 \sim \rho$. If the stochastic process is Markovian the following relationship holds:

$$P[x_{1:T}] = \rho(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t), \tag{2.5}$$

where $p(x_{t+1}|x_t)$ is the conditional state transition density. In other words, for a Markov process its dynamics only depend on the current state. The dynamics of the process themselves are fully determined by $p(x_{t+1}|x_t)$, which encodes the probability of transitioning from a given state to any other state at any point in time. When the transition dynamics do not explicitly depend on time, i.e., when $p(x_{t+1}|x_t) = p(x_{t+1+l}|x_{t+l})$ for any valid l, we refer to the Markov process as time-homogeneous, which will be our focus.

In many application areas, the dynamics of the underlying Markov process do not evolve autonomously in time. Instead, they are subject to inputs such as control actions, $u_t \in \mathcal{U}$, that steer their dynamics in particular state trajectories. In this setting, we often model the action-dependence of the system's state transitions through state transition models of the form $p(x_{t+1}|x_t, u_t)$, where $p : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \to [0, \infty)$. When these actions are deterministically produced by a controller, we may think of them as parameters that shape the underlying path distribution of the stochastic process, as in Definition 2.2. However, we may also think of control actions as random variables drawn from a given distribution at each point in time. This distribution is referred to as a *policy* and it is typically notated as $\pi(u_t|x_t)$, where $\pi : \mathcal{U} \times \mathcal{X} \to [0, \infty)$. An important distinction between this setting and the standard Markov process setting is that the stochastic nature of control actions induces a path distribution over the actions themselves, leading to joint path distributions of the following form:

$$P[x_{1:T}, u_{1:T}] = \rho(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t). \tag{2.6}$$

Nonetheless, we can take a look at the *expected* paths that the system takes by defining $p_\pi(x_{t+1}|x_t) = E_\pi[p(x_{t+1}|x_t, u_t)$, which results in a distribution

$$P_\pi[x_{1:T}] = \rho(x_1) \prod_{t=1}^{T-1} p_\pi(x_{t+1}|x_t) \tag{2.7}$$

whose structure resembles Eq. 2.5. These expressions formalize the sense in which the role of a controller or policy is to steer a Markov process.

The process of optimizing an objective function by selecting policies that steer a Markov process towards desirable states is referred to as a *Markov decision process (MDP)*. MDPs

are defined as 4-tuples $(\mathcal{X}, \mathcal{U}, p, r)$ in terms of the state and action spaces $\mathcal{X}$ and $\mathcal{U}$, the state transition model p, and lastly a bounded *reward* function $r : \mathcal{X} \times \mathcal{U} \to [r_{\min}, r_{\max}]$ (sometimes we use a cost function instead). Note that MDPs can be defined as 5-tuples when a reward-discounting factor $\gamma = [0, 1)$ is included to ensure the convergence of decision-making processes in infinite-horizon tasks. We will return to the MDP problem setting when we discuss optimal control and reinforcement learning in future sections.

2.2. The Principle of Maximum Caliber

In proposing that statistical mechanics could be understood as a statistical inference procedure applied to the states and dynamics of physical systems, E. T. Jaynes presented a general mathematical framework from which one can perform such inferences: The principle of maximum entropy (MaxEnt) [11], and later on the principle of maximum caliber (MaxCal) [15]. In this section, we will briefly review the basics of the MaxEnt and MaxCal frameworks, as they are essential to the results in this thesis.

Much like other inference frameworks, the underlying goal of MaxEnt and MaxCal is to find a probability distribution that best describes observations from some underlying process. What distinguishes both MaxEnt and MaxCal from other inference frameworks is their ability to incorporate hard constraints into the underlying optimization, which is essential when dealing with embodied systems subject to the rigid constraints imposed upon them by the laws of physics. In a sense, MaxEnt and MaxCal attempt to codify the "common wisdom" of Occam's razor: If you had to infer the distribution of random variable from partial information, then the best guess is the one that introduces the least

additional assumptions—or, equivalently, the one that is least-biased—while remaining consistent with prior knowledge we might have about the random variable.

To formalize this common wisdom, MaxEnt makes use of information theory and the concept of *entropy*, arguing that Shannon entropy and Boltzmann entropy are interchangeable at some level. Now, let X be an random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with some state-space $\mathcal{X}$, and let $p : \mathcal{X} \to [0, \infty)$ be its corresponding probability density function whose form we are interested in inferring. Then, we can define an entropy operator S that acts directly on arbitrary probability density functions in the following way:

$$S[p(x)] = -\int_{\mathcal{X}} p(x) \log p(x) dx. \tag{2.8}$$

Using this operator, we can formalize the MaxEnt inference procedure as finding the distribution $p^{ME}(x)$ that is the result of the following optimization:

$$p^{ME}(x) = \underset{p(x)}{\operatorname{argmax}}\ S[p(x)] \tag{2.9}$$

$$= \underset{p(x)}{\operatorname{argmin}}\ D_{KL}(p(x)||p_{\text{uniform}}) \tag{2.10}$$

where we note the equivalence between entropy maximization and minimization of a KL-divergence with respect to a uniform distribution p_{uniform}. More generally, prior beliefs over the candidate distribution, $p_0(x)$, can be introduced by minimizing $D_{KL}(p(x)||p_0(x))$ instead. However, as written these optimizations do not produce valid probability densities, requiring us to introduce constraints into the optimization. To this end, let $\hat{S}$ be

$$\hat{S}[p(x)] = S[p(x)] - \lambda_0\Big(\int_{\mathcal{X}} p(x)dx - 1\Big), \tag{2.11}$$

where λ_0 is a Lagrange multiplier enforcing the constraint that the probability density integrates to 1 over the domain of the random variable. By optimizing $\hat{S}$ instead of S, we can ensure that the resulting density is valid. The resulting optimization is then

$$p^{ME}(x) = \underset{p(x)}{\operatorname{argmax}}\ \hat{S}[p(x)]. \tag{2.12}$$

By incorporating different constraints and priors in this manner, we can use MaxEnt to infer the statistical properties of different random variables. For example, optimizing the MaxEnt objective with Eq. 2.11 as written results in the uniform distribution over the compact domain $\mathcal{X}$. Thus, the least-biased guess we can make about the distribution of a random variable over a compact domain—when given no additional information—is that it is uniform, which motivates the sense in which MaxEnt captures the common wisdom of Occam's razor. Importantly, distributions that maximize entropy functionals are not only the "least-biased" distributions subject to constraints—they are also the distributions with the greatest support over the domain of the distribution. In other words, they are the ones with the most spread probability mass over the state-space.

Given an understanding of path distributions, we can understand MaxCal as a straightforward generalization of MaxEnt from random variables to stochastic processes. In other words, MaxCal is to densities over state trajectories what MaxEnt is to densities over states [16]. More formally, let $\{X(t)\}_{t\in\mathcal{T}}$ be a stochastic process over a time interval $\mathcal{T}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a path distribution $P[x(t)]$. In order to infer the path distribution of some dynamical process, $P[x(t)]$, the MaxCal principle proposes to optimize the "caliber" of process, which is equivalent to the entropy of the process'

path distribution,

$$S[P[x(t)]] = -\int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \log P[x(t)] \mathcal{D}x(t). \tag{2.13}$$

Just as before, we require a normalization constraint to make the MaxCal framework produce valid path distributions, which we can enforce with Lagrange multipliers in the following way,

$$\hat{S}[P[x(t)]] = S[P[x(t)]] - \lambda_0 \Big(\int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \mathcal{D}x(t) - 1 \Big), \tag{2.14}$$

and then optimize

$$P^{MC}[x(t)] = \underset{P[x(t)]}{\operatorname{argmax}} \, \hat{S}[P[x(t)]]. \tag{2.15}$$

In the MaxEnt equivalent of this example, we saw that optimizing Eq. 2.11 results in the uniform distribution. Here, solving Eq. 2.14 results in a uniform path distribution, which describes an *i.i.d* uniformly-random sampling process. In other words, the maximum entropy sampling process describes the dynamics of an agent capable of teleporting around the state-space at-will.

Solving MaxEnt and MaxCal problems with different constraints is procedurally identical. Given an objective function $\hat{S}$ that incorporates all relevant constraints, one proceeds by taking the variational derivative of the entropy functional and setting it to zero:

$$\frac{\delta \hat{S}}{\delta p} = 0, \text{ or } \frac{\delta \hat{S}}{\delta P} = 0. \tag{2.16}$$

Then, one would solve for p^{ME} or P^{MC} as well as for the different Lagrange multipliers in the expression of $\hat{S}$, which may or may not have closed-form analytical expressions. While the procedure for solving MaxEnt and MaxCal problems is relatively simple, the real challenge is to formalize system constraints in ways that are simultaneously analytically or computationally tractable, as well as true to the underlying system.

Equipped with an understanding of the basics of the MaxCal framework, in the following sections we will formulate several MaxCal optimizations with different constraints and priors, and explore their relationship to fields such as physics, design, learning, and control. In the first few of these sections we will focus on inference, asking first *how can we infer the behavior of complex embodied agents?* Then, *how can we infer the behavior of goal-directed agents?* In the last few sections we will focus on synthesis techniques grounded in these principles, asking *how can we synthesize optimal goal-directed behavior based these principles?* Our goal will be to illustrate how inferring and manipulating an agent's path distributions can provide solutions to many problems of interest, such as embodied exploration, predicting steady-state behavior, nonlinear optimal control, and reinforcement learning.

2.3. Continuity, Exploration, and Diffusion

Throughout this section, our goal will be to formalize path continuity as a constraint on stochastic control processes, and to motivate its potential impact on embodied exploration and learning. To this end, we will derive solutions to maximum caliber optimizations with path continuity constraints and discuss their properties from information-theoretic,

control-theoretic, physics-based perspectives. We note that much of the work in this section was originally published in the supplement of [7].

2.3.1. Path Continuity and the *i.i.d.* Assumption

As we saw in the previous section, the unconstrained (but normalized) maximum entropy path distribution describes a uniformly random sampling process. This is an idealized sampling process whose samples satisfy the *independent and identically distributed (i.i.d.)* assumption. The *i.i.d.* assumption is ubiquitous in all of machine learning, optimization, and statistics. Effectively, data samples are *i.i.d.* when they are all drawn from the same underlying distribution, and when the probability of drawing a given sample does not depend on the probability of drawing any other sample. We may envision an *i.i.d.* sampling process by imagining an agent capable of teleporting around its environment—one whose dynamics are discontinuous, such as the idealized 2D gantry system in Fig. 2.1. In general, robots tend to follow *continuous* paths through their state-space, as in the 2D gantry example in Fig. 2.2. Spatiotemporal continuity of experience is a defining feature of physical embodiment. Thus, as long as data is being sampled by an embodied system, the *i.i.d.* property will likely be violated, which has a profound impact on any downstream learning or data-driven optimization.

Another important consequence of violating the *i.i.d.* property is sample redundancy. Consider drawing identically distributed samples from a stochastic process at two particular points in time, $x_t \sim X_t$ and $x_{t+1} \sim X_{t+1}$. How much information do we stand to gain from observing these samples in expectation? This is precisely what Shannon entropy

quantifies [17]. Generally, Shannon entropy is *subadditive*:

$$S[p(x_t, x_{t+1})] \leq S[p(x_t)] + S[p(x_{t+1})]. \tag{2.17}$$

However, this inequality is saturated if and only if X_t and X_{t+1} are statistically independent—a property uniquely satisfied by maximum caliber sampling processes like those discussed in the previous section. Therefore, the information gained from individual observations is only equal to the information gained from a collection of sequential observations when the underlying samples are independent. In other words, if a sampling process is not *i.i.d.*, then it is inefficient in the sense that the underlying sequential observations are informationally redundant and *temporally correlated*. This observational redundancy is at the heart of many important questions in robotics, and motivates the need for effective exploration strategies across robot learning applications.

2.3.2. Effective Exploration and Temporal Correlations

One of the most common exploration strategies in robot learning is randomized action exploration. The underlying idea being that agents will experience a diverse set of observations if they take a diverse set of random actions. In the simplest of these methods, agents merely sample actions randomly from either uniform or Gaussian distributions to in hopes of producing effective exploration. More sophisticated methods, such as maximum entropy reinforcement learning [18–20], elaborate on this basic idea by using policy optimization to learn a distribution from which to sample random actions and improve agent outcomes. For the purpose of our analysis, these more advanced methods are functionally equivalent to the simplest methods—they assume that taking random

actions produces effective state exploration. However, from the perspective of control theory we know that this is not necessarily the case. For a system to be able to reach desired states arbitrarily, it must be controllable [21]. In this subsection we will explore the connection between temporal correlations, controllability, and exploration in systems with continuous state-space trajectories.

To illustrate how the controllability properties of agents can affect exploration outcomes, we will briefly consider randomized action exploration in linear time-varying (LTV) control systems. LTV dynamics can be expressed in terms of continuous-time deterministic trajectories in the following way:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{2.18}$$

where $A(t)$ and $B(t)$ are appropriately dimensioned matrices with state and control vectors $x(t) \in \mathcal{X}$ and $u(t) \in \mathcal{U}$, and $x(t_0) = x^*$ for $\mathcal{T} = [t_0, t]$. The general form of solutions to this system of linear differential equations is expressed in terms of a convolution with the system's state-transition matrix, $\Psi(t, t_0)$, in the following way:

$$x(t) = \Psi(t, t_0)x^* + \int_{t_0}^{t} \Psi(t, \tau)B(\tau)u(\tau)d\tau. \tag{2.19}$$

We consider these dynamics because by working with LTV dynamics we implicitly consider a very broad class of systems—all while retaining the simplicity of linear controllability analysis [22]. This is due to the fact that the dynamics of any nonlinear system that is locally linearizable along its trajectories can be effectively captured by LTV dynamics. Hence, any conclusions applicable to the dynamics in Eq. 2.18 will apply to linearizable nonlinear systems.

To understand the performance of randomized action exploration in a given LTV system, we may ask what states are reachable by this system within a finite time interval. After all, states that are not reachable cannot be explored or learned from. This is precisely what controllability characterizes:

Definition 2.3. *A system is said to be controllable over a time interval* $[t_0, t] \subset \mathcal{T}$ *if given any states* $x^*, x_1 \in \mathcal{X}$*, there exists a controller* $u(t) : [t_0, t] \to \mathcal{U}$ *that drives the system from state* x^* *at time* t_0 *to* x_1 *at time* t*.*

While this definition intuitively captures what is meant by controllability, as written it does not make for an easily verifiable property. To this end, different computable metrics have been developed that equivalently characterize the controllability properties of certain classes of systems (e.g., the Kalman controllability rank condition [23]). In particular, here we will analyze the controllability Gramian of the system.

For our class of LTV systems, characterizing controllability with this method is simple:

$$W(t_0, t) = \int_{t_0}^{t} \Psi(t, \tau) B(\tau) B(\tau)^T \Psi(t, \tau)^T d\tau, \tag{2.20}$$

where the Gramian is a symmetric positive semidefinite matrix that depends on the state-control matrix $B(t)$ and the state-transition matrix $\Psi(t, t_0)$. The Gramian is a controllability metric that quantifies the amount of energy required to actuate the different degrees of freedom of the system [24, 25]. For any given finite time interval, the controllability Gramian also characterizes the set of states reachable by the system. Importantly, when the controllability Gramian is full-rank, the system is provably controllable in the sense of Definition 2.3 [21], and capable of fully exploring its environment. However, when

the controllability Gramian is poorly conditioned, substantial temporal correlations are introduced into the agent's state transitions, which can prevent effective exploration—as we will show.

To draw the connection between random action exploration, controllability, and temporal correlations explicitly, we will now revisit the dynamics in Eq. 2.18 under a slight modification. Let us design a controller that performs naive action randomization, i.e., let $u(t) = \xi$, where $\xi \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$ and Id is an identity matrix with diagonal of the same dimension as the control inputs, and $\mathbf{0}$ is the zero vector of the same dimension. Note that the system trajectories are now collections of random variables. Then, we have:

$$\dot{x}(t) = A(t)x(t) + B(t) \cdot \xi. \tag{2.21}$$

Here, we abuse notation slightly to minimize the difference between this equation and Eq. 2.18, but we can interpret the system as having linear Langevin dynamics [26]. With these modifications in mind, we are now interested in examining the mean and covariance trajectory statistics in hopes of characterizing the structure of temporal correlations induced by the agent dynamics. We begin by taking the expectation over system trajectories described by Eq. 2.19:

$$\begin{aligned} E[x(t)|x(t_0) = x^*] &= E\Big[\Psi(t, t_0)x^* + \int_{t_0}^{t} \Psi(t, \tau)B(\tau) \cdot \xi d\tau\Big] \\ &= \Psi(t, t_0)x^* + E\Big[\int_{t_0}^{t} \Psi(t, \tau)B(\tau) \cdot \xi d\tau\Big] \\ &= \Psi(t, t_0)x^*. \end{aligned} \tag{2.22}$$

Hence, the expected sample paths of the dynamics will be centered around the autonomous paths of the system—that is, the paths the system takes in the absence of control inputs.

We may now characterize the covariance of our system's sample paths. To do so, let $\mathbf{C}[x^*] = E\big[(x(t) - E[x(t)])(x(t) - E[x(t)])^T \big| x(t_0) = x^*\big]$ be the trajectory autocovariance about an initial condition x^*. The trajectory autocovariance is a local measure of temporal correlations in stochastic processes. To see this, let $\{X(t)\}_{t\in\mathcal{T}}$ be a stochastic process defined according to Definition 2.1. Then, an autocovariance function, $K_{XX}(t_1, t_2)$, expresses the covariance of the process with itself at any two points in time $t_1, t_2 \in \mathcal{T}$, or

$$K_{XX}(t_1, t_2) = E\big[(X(t_1) - E[X(t_1)])(X(t_2) - E[X(t_2)])^T\big]. \tag{2.23}$$

Point-wise autocovariances between the random variables of a stochastic process can then be integrated over a given time interval $[t_0, t] \subset \mathcal{T}$ for a given initial condition $x^* \in \mathcal{X}$, leading to the following result:

$$\begin{aligned}\int_{t_0}^{t} K_{XX}(\tau, t)d\tau &= E\big[(X(t) - E[X(t)])(X(t) - E[X(t)])^T \big| X(t_0) = x^*\big] \\ &= \mathbf{C}[x^*].\end{aligned} \tag{2.24}$$

Thus, the trajectory autocovariance, $\mathbf{C}[x^*]$, acts as a measure of temporal correlations along a process' sample paths by integrating statistical autocorrelations between random

variables over a given time interval. With these preliminaries taken care of, we have:

$$
\begin{aligned}
\mathbf{C}[x^*] &= E\big[(x(t) - E[x(t)])(x(t) - E[x(t)])^T \big| x(t_0) = x^*\big] \\
&= E\Big[\Big(\Psi(t,t_0)x^* + \int_{t_0}^{t} \Psi(t,\tau)B(\tau)\cdot\xi d\tau - E[x(t)]\Big) \\
&\qquad \times \Big(\Psi(t,t_0)x^* + \int_{t_0}^{t} \Psi(t,\tau)B(\tau)\cdot\xi d\tau - E[x(t)]\Big)^T \Big| x(t_0) = x^*\Big] \\
&= E\Big[\Big(\int_{t_0}^{t} \Psi(t,\tau)B(\tau)\cdot\xi d\tau\Big)\Big(\int_{t_0}^{t} \Psi(t,\tau)B(\tau)\cdot\xi d\tau\Big)^T \Big| x(t_0) = x^*\Big] \\
&= E\Big[\int_{t_0}^{t} \Psi(t,\tau)B(\tau)\cdot(\xi\xi^T)\cdot B(\tau)^T\Psi(t,\tau)^T d\tau \Big| x(t_0) = x^*\Big] \\
&= \int_{t_0}^{t} \Psi(t,\tau)B(\tau)B(\tau)^T\Psi(t,\tau)^T d\tau. \qquad (2.25)
\end{aligned}
$$

By inspection of the above expression and Eq. 2.20, we arrive at the following important connection:

$$\mathbf{C}[x^*] = W(t_0,t) \qquad (2.26)$$

which tells us that for LTV dynamics (and by extension for linearizable nonlinear dynamics), a measure of temporal correlations—the trajectory autocovariance $\mathbf{C}[x^*]$—is exactly equivalent to the controllability Gramian of the system. Thus, for a broad class of systems, an agent's controllability properties introduce temporal correlations along their state trajectories. Moreover, in LTV systems these are not state-dependent properties. In other words,

$$\nabla_x \mathbf{C}[x^*] = \nabla_x W(t_0,t) = \mathbf{0}, \qquad (2.27)$$

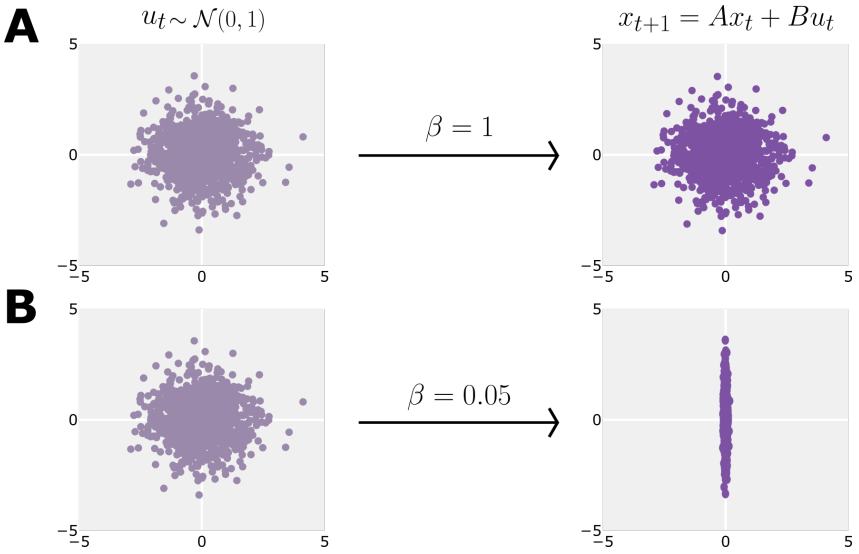


Figure 2.4. **Effect of controllability on the distribution of reachable states**. **a,** For the simple system in Eq. 2.29, we depict the effect of controllability on a naive random action exploration strategy. For a system with ideal controllabilty properties, isotropic distributions of actions map onto isotropic distributions of states. **b,** However, when the system is poorly conditioned the system dynamics distort the isotropy of the original input distribution, introducing temporal correlations, and fundamentally changing its properties as an exploration strategy.

where $\mathbf{0}$ is an appropriately dimensioned zero matrix. However, for linearizable nonlinear systems, as well as more general nonlinear systems, these properties will be state-dependent. While our controllability analysis has been restricted to the class of dynamics describable by linear differential equations with time-varying parameters, we note that the connections we observe between trajectory autocovariance and controllability Gramians have been shown to hold for even more general classes of nonlinear systems through more involved analyses [27].

From Eq. 2.21 we can describe the system's reachable states by analyzing its state probability density function, which can be found analytically by solving its associated

Fokker-Planck equation [28]. To do this, we only require the mean and covariance statistics of the process, in Eqs. 2.22 and 2.25. The system's time-dependent state distribution is

(2.28)
$$p(x,t,t_0) = \frac{1}{\sqrt{(2\pi)^d \det[W(t_0,t)]}} \exp\Big[-\frac{1}{2}\big(x-\Psi(t,t_0)x^*\big)^T W^{-1}(t_0,t)\big(x-\Psi(t,t_0)x^*\big)\Big]$$

for some choice of initial conditions at t_0, where we have substituted Eq. 2.26 to highlight the role of controllability in the probability density of states reachable by the system through naive random exploration. Figure 2.4 illustrates this in a toy dynamical system with linear dynamics:

(2.29)
$$x_{t+1} = x_t + \begin{bmatrix} \beta & 0 \\ 0 & 1 \end{bmatrix} u_t.$$

We observe that changes in β have an effect on the distribution of reachable states for the system that are consistent with Eq. 2.28. Thus, the effectiveness of action randomization as an effective exploration strategy is entirely determined by the controllability properties of the system—or, equivalently, by a measure of temporal correlations of its state trajectories. This raises the question, if action randomization is not the right idea then what is?

2.3.3. Undirected Exploration as Maximum Caliber Trajectory Sampling

As we saw in the previous subsection, action randomization is an ineffective exploration strategy in systems that take continuous paths through their state-spaces. This is the case even for sophisticated strategies that seek to maximize the entropy of the randomized action distribution. Instead, in this subsection we will investigate the following question: *What if we instead developed an exploration strategy around maximizing the entropy of*

an agent's state transitions? To this end, we will make use of the principle of maximum caliber. However, in order to ensure that our exploration strategy produces paths that are continuous, we will need to formalize a dynamics-agnostic constraint on path continuity.

What sorts of principles can such a constraint be based upon? Conservation of energy is not applicable because autonomous systems are inherently nonequilibrium systems. Nonetheless, the behavior of embodied autonomous systems is constrained by aspects of their morphology, such as actuation limits. In particular, the rates at which agent experiences or states can vary—and *co-vary*—in time are typically bounded, which prevents them from discontinuously jumping between states by limiting their local rate of exploration. In fact, this is precisely what we found in the previous subsection, where we saw that a system's ability to locally explore space is closely tied to a measure of its temporal correlations, $\mathbf{C}[x^*]$. Thus, we will choose to constrain the velocity fluctuations of our stochastic process so that they are finite and consistent with the integrated autocovariance statistics of the process, which may be determined empirically, and are related to a system's controllability properties in a broad class of systems. The use of an empirical (or learned) autocovariance estimate to quantify velocity fluctuations is important because different embodied agents have different limitations, which may additionally be spatially inhomogeneous and difficult to know a priori. Through this constraint, we can ensure that agent sample paths are continuous in time.

To formulate this path continuity constraint, we must first express the system's velocity fluctuations at each point in state space, $x^* \in \mathcal{X}$. We define the system's velocity

fluctuations along sample paths $x(t)$ in the following way:

$$\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \int_{\mathcal{X}^\mathcal{T}} P[x(t)] \int_{\mathcal{T}} \dot{x}(\tau)\dot{x}(\tau)^T \delta(x(\tau) - x^*) d\tau \mathcal{D}x(t), \tag{2.30}$$

where $\delta(\cdot)$ denotes the Dirac delta function, and we note that the $\langle \cdot \rangle$ expression is equivalent to the following expectation: $E[\int_{\mathcal{T}} \dot{x}(\tau)\dot{x}(\tau)^T \delta(x(\tau) - x^*) d\tau]$. We assume that the tensor described by Eq. 2.30 is full-rank so that the system's velocity fluctuations are not degenerate anywhere in the state space of the stochastic process. This assumption is crucial because it guarantees that our resulting path distribution is non-degenerate. If we had instead chosen to constrain the system by directly bounding the magnitude of its velocities (i.e., $E[\int_{\mathcal{T}} \dot{x}(\tau)^T \dot{x}(\tau) \delta(x(\tau) - x^*) d\tau]$) as opposed to its velocity fluctuations, we would not be able to guarantee the non-degeneracy of the resulting path distribution. Another important note is that the velocities of the trajectories of the stochastic process in this expression should be interpreted in the Langevin sense [26]. That is to say, not as expressions of the differentiability of the sample paths of the underlying stochastic process, but as a shorthand for an integral representation of the stochastic differential equations describing the evolution of the sample paths of the system.

We can now express our constraint as,

$$\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \mathbf{C}[x^*], \quad \forall x^* \in \mathcal{X}. \tag{2.31}$$

Crucially, these statistics are bounded everywhere in the exploration domain, and we assume them to satisfy Lipschitz continuity so that their spatial variations are bounded. We note that linearizability of the underlying agent dynamics is a sufficient condition to satisfy this property. Hence, we now have equality constraints on the system's velocity

fluctuations that can vary at each point in the exploration domain—as one would expect for a complex embodied system, such as a robot. We note that, as before, we require that $P[x(t)]$ integrates to 1 so that it is a valid probability density over trajectories, which introduces another constraint.

With expressions for each of our constraints in hand, we may now express the complete maximum caliber variational optimization problem using Lagrange multipliers:

$$\begin{aligned}(2.32) \qquad \underset{P[x(t)]}{\operatorname{argmax}} \; & -\int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \log P[x(t)] \mathcal{D}x(t) - \lambda_0 \Big(\int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \mathcal{D}x(t) - 1 \Big) \\ & - \int_{\mathcal{X}} Tr \Big(\Lambda(x^*)^T \big(\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} - \mathbf{C}[x^*] \big) \Big) dx^*,\end{aligned}$$

Here, we express the constraints at all points x^* by taking an integral over all points in the domain. The λ_0 is a Lagrange multiplier enforcing our constraint that ensures valid probability densities, and $\Lambda(\cdot)$ is a matrix-valued Lagrange multiplier working to ensure that the rate of exploration constraints hold at every point in the domain. By solving this optimization we can obtain an expression for the maximum entropy distribution over sample paths. As a result, the solution to this problem will determine the distribution over sample paths with the greatest support, with the most uniformly spread probability mass, and with the least-correlated sample paths—thereby specifying the statistical properties of an optimal undirected exploration strategy, subject to a path continuity constraint. Moreover, as a result of the solution's ergodic properties, we will also find our solution to violations of the *i.i.d.* sampling during embodied exploration.

2.3.4. Maximally Diffusive Trajectory Statistics

To solve this optimization, we must take the variation of the objective with respect to $P[x(t)]$ and solve for the optimal trajectory distribution. Remarkably, this can be done analytically; moreover, we find that the statistics of the maximum caliber sample paths are given by those of a diffusion process, as shown in the following theorem.

Theorem 2.1. *The statistics of the maximum caliber sample paths of a stochastic process (Definition 2.1) with continuous sample paths (in the sense of Eq. 2.32) are described by a diffusion process with spatially-varying coefficients.*

Proof. Letting $\mathcal{T} = [t_0, t]$, we begin by substituting Eq. 2.30 into Eq. 2.32, taking its variation with respect to the probability density $\delta\hat{S}[P[x(t)]]/\delta P[x(t)]$, and setting it equal to 0:

$$\frac{\delta\hat{S}}{\delta P[x(t)]} = -1 - \log P_{max}[x(t)] - \lambda_0 - \int_{\mathcal{X}}\int_{t_0}^{t} Tr\Big(\Lambda(x^*)^T(\dot{x}(\tau)\dot{x}(\tau)^T)\Big)\delta(x(\tau)-x^*)d\tau dx^* = 0.$$

Then, taking advantage of the following linear algebra identity, $a^T B a = Tr(B^T(aa^T))$, for any $a \in \mathbb{R}^m$ and $B \in \mathbb{R}^{m\times m}$; as well as the properties of the Dirac delta, we can simplify our expression to the following:

$$\frac{\delta\hat{S}}{\delta P[x(t)]} = -1 - \log P_{max}[x(t)] - \lambda_0 - \int_{t_0}^{t} \dot{x}(\tau)^T\Lambda(x(\tau))\dot{x}(\tau)d\tau = 0,$$

which allows us to solve for the maximum entropy probability distribution over the sample paths of our stochastic control process. The solution will then be of the form:

$$P_{max}[x(t)] = \frac{1}{Z}\exp\Big[-\int_{t_0}^{t} \dot{x}(\tau)^T\Lambda(x(\tau))\dot{x}(\tau)d\tau\Big], \tag{2.33}$$

where we have subsumed the constant and Lagrange multiplier, λ_0, into a normalization factor, Z. We note that even without determining the form of our Lagrange multipliers, the maximum entropy probability density in Eq. 2.33 is already equivalent to the path probability of a diffusing particle with a (possibly anisotropic) spatially-inhomogeneous diffusion tensor (see [26], Ch. 9). While there is more work needed to characterize the diffusion tensor of this process, $\Lambda^{-1}(\cdot)$, this completes our proof. $\square$

Thus, the least-correlated sample paths, which optimally sample from the exploration domain through continuous trajectories, are statistically equivalent to diffusion. This is to say that the distribution of paths with the greatest support over the state space describes the paths of a diffusion process. Hence, if the goal of some stochastic control process is to optimally explore and sample from its state space, the best strategy is to move randomly—that is, to decorrelate its sample paths. An additional benefit of our diffusive exploration strategy is that we did not have to presuppose that our agent dynamics were Markovian or ergodic. Instead, we find that these properties emerge through our derivation as intrinsic properties of the optimal exploration strategy itself. The following corollaries of Theorem 2.1 follow from the connection to diffusion processes and Markov chains, and as such more general forms of these proofs may be found in textbooks on stochastic processes and ergodic theory. Here, we assume that the diffusion tensor in Eq. 2.33, $\Lambda^{-1}(\cdot)$, is full-rank and invertible everywhere in the state space. Additionally, for now we will assume that $\Lambda^{-1}(\cdot)$ is Lipschitz and bounded everywhere on $\mathcal{X}$. We will later find that these are not in fact different assumptions from those made in Eqs. 2.30 and 2.31.

Corollary 2.1.1. *The maximum caliber sample paths of a stochastic process stochastic process (Definition 2.1) with continuous paths (in the sense of Eq. 2.32) satisfy the Markov property.*

Proof. This follows trivially from the temporal discretization of our path distribution in Eq. 2.33, or alternatively from the properties of Langevin diffusion processes. Letting x_t be the initial condition, we can see that,

$$\begin{aligned} p_{max}(x_{t+\delta t}|x_t) &= \frac{1}{Z}\exp\Big[-\int_t^{t+\delta t}\dot{x}(\tau)^T\Lambda(x(\tau))\dot{x}(\tau)d\tau\Big] \\ &\approx \frac{1}{Z_d}\exp\Big[-|x_{t+\delta t}-x_t|^2_{\Lambda(x_t)}\Big], \end{aligned} \tag{2.34}$$

where we subsumed δt into a new normalization constant Z_d for convenience, and note that the support of $p_{max}(x_{t+\delta t}|x_t)$ is infinite. Importantly, our local Lagrange multiplier $\Lambda(x_t)$ enforces our velocity fluctuation constraint within a neighborhood of states reachable from x_t for a sufficiently small time interval δt, which is guaranteed by our Lipschitz continuity assumption. In what remains of this thesis we use $\delta t = 1$ for notational convenience, but without loss of generality. Thus, our distribution in Eq. 2.34 depends only on the current state, which concludes our proof. □

Corollary 2.1.2. *The maximum caliber sample paths of a stochastic process (Definition 2.1) with continuous paths (in the sense of Eq. 2.32) in a compact and connected space $\mathcal{X} \subset \mathbb{R}^d$ are ergodic.*

Proof. To prove the ergodicity of the process described by Eq. 2.33, we use Corollary 2.1.1 and the properties of $\mathcal{X}$. We begin by discretizing our optimal stochastic control

process in time and space such that $P_{max}[x_{1:N}] = \prod_{t=1}^{N-1} p_{max}(x_{t+1}|x_t)$, which we can do without loss of generality as a result of Corollary 2.1.1 and because $\mathcal{X}$ is compact, resulting in a finite space. Importantly, since $p_{max}(x_{t+1}|x_t) > 0,\ \forall x_t, x_{t+1} \in \mathcal{X},\ \forall t \in \mathcal{T}$, and $\mathcal{X}$ is finite and connected, then all states in $\mathcal{X}$ communicate. Moreover, because for all $x^* \in \mathcal{X}$, $p_{max}(x^*|x^*) > 0$, the underlying Markov chain described by the transition kernel is aperiodic. Therefore, the Markov chain describing the stochastic control process is ergodic [29]. □

At this point, it is essential to discuss ergodicity more broadly and what it practically implies about the underlying sampling process. Ergodicity is a property of stochastic processes requiring that the statistics of samples drawn sequentially from a process match those of samples drawn *i.i.d.* from its stationary distribution—as codified by any of many "ergodic theorems" [30]. To motivate its relevance to optimization and learning with data sampled from embodied sampling processes, we will explore a simple example. Consider learning a model describing the dynamics of an embodied dynamical process (e.g., the dynamics of a boat in turbulent waters). Given observations drawn from $\{X_t\}_{t\in\mathcal{T}}$, we might be interested in finding parameters θ that allow a model $x_{t+1} = f_\theta(x_t)$ to predict the dynamics of the underlying process. Thus, we might be interested in minimizing a quadratic loss

$$l(\theta) = E[(x_{t+1} - f_\theta(x_t))^2] \tag{2.35}$$

by using gradient descent to find the optimal θ to fit our dynamics model. However, now there is an issue—doing so would require being able to easily compute $l(\theta)$ and $\nabla_\theta l$, and

the expectation in the definition of $l(\theta)$ makes this intractable in almost all application domains of interest. So, in practice, this problem is solved by optimizing a sample-based surrogate of the loss

$$\hat{l}(\theta) = \frac{1}{T}\sum_{t=1}^{T-1}(x_{t+1} - f_\theta(x_t))^2 \tag{2.36}$$

and its gradients $\nabla_\theta \hat{l}$ from a batch of observations $\{x_1, \cdots, x_T\}$. The underlying hope is that optimizing $\hat{l}(\theta)$ is interchangeable from optimizing $l(\theta)$, or more precisely that

$$\lim_{T\to\infty} \hat{l}(\theta) = l(\theta). \tag{2.37}$$

The problem is that this is explicitly *not* true unless the underlying process is *i.i.d.*, or we introduce additional assumptions. However, if the underlying process is ergodic, then Birkhoff's ergodic theorem guarantees this relationship still holds. Thus, in a sense, ergodic sampling is asymptotically "as good as" *i.i.d.* sampling. With this motivation in mind, we may now return to the derivation of our exploration strategy.

To finish our derivation and fully characterize the nature of our maximum entropy exploration strategy, we must return to Eq. 2.33 and determine the form of the matrix-valued Lagrange multiplier $\Lambda(\cdot)$. Hence, we will return to our expression for $\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*}$ in Eq. 2.30 and discretize our continuous sample paths, which we can do without loss of generality due to Corollary 2.1.1. Since Eq. 2.30 represents a proportionality, we take out many constant factors throughout the derivation. Additionally, any constant factor of $\Lambda(\cdot)$ would be taken care of by the normalization constant Z in the final expression for Eq. 2.33. We proceed by discretizing Eq. 2.30, using i and j as time indices and $p_{max}(\cdot|\cdot)$

as the conditional probability density defined in Eq. 2.34. We do this by slicing the time interval $[t_0, t]$ into time indices $\{1, \cdots, N\}$. Our resulting expression is the following:

$$\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \prod_{i=1}^{N-1} \Big[\int_{\mathcal{X}} dx_{i+1}\ p_{max}(x_{i+1}|x_i) \Big] \sum_{j=1}^{N-1} (x_{j+1} - x_j)(x_{j+1} - x_j)^T \delta(x_j - x^*), \tag{2.38}$$

where the path integrals are discretized according to the Feynman formalism [14], using the same discretization as in our proof of Corollary 2.1.1.

From this expression in Eq. 2.38, we take the following two steps. First, we switch out the order of summation and product by applying the Fubini-Tonelli theorem. Then, we factor out two integrals from the product expression—one capturing the probability flow *into* x_j and one capturing the flow *out of* it:

$$\begin{aligned} &= \sum_{j=1}^{N-1} \prod_{i \neq j, j-1} \Big[\int_{\mathcal{X}} dx_{i+1}\ p_{max}(x_{i+1}|x_i) \Big] \\ &\quad \times \int_{\mathcal{X}} p_{max}(x_j|x_{j-1}) \int_{\mathcal{X}} p_{max}(x_{j+1}|x_j)(x_{j+1} - x_j)(x_{j+1} - x_j)^T \delta(x_j - x^*) dx_{j+1} dx_j, \end{aligned}$$

where $\times$ denotes multiplication with the line above. Then we can apply the Dirac delta function to simplify our expression and get:

$$\begin{aligned} &= \sum_{j=1}^{N-1} \prod_{i \neq j, j-1} \Big[\int_{\mathcal{X}} dx_{i+1}\ p_{max}(x_{i+1}|x_i) \Big] \\ &\quad \times p_{max}(x^*|x_{j-1}) \int_{\mathcal{X}} p_{max}(x_{j+1}|x^*)(x_{j+1} - x^*)(x_{j+1} - x^*)^T dx_{j+1}. \end{aligned} \tag{2.39}$$

To simplify further we will tackle the following integral as a separate quantity:

$$I = \int_{\mathcal{X}} p_{max}(x_{j+1}|x^*)(x_{j+1} - x^*)(x_{j+1} - x^*)^T dx_{j+1}. \tag{2.40}$$

where we can substitute Eq. 2.34 into Eq. 2.40 to get:

$$I = \int_{\mathcal{X}} \frac{1}{Z_d} e^{-(x_{j+1}-x^*)^T \Lambda_{(x^*)}(x_{j+1}-x^*)} (x_{j+1}-x^*)(x_{j+1}-x^*)^T dx_{j+1}.$$

This integral can then be tackled using integration by parts and closed-form Gaussian integration. Thus far, we have not had any need to specify the domain in which exploration takes place. However, in order to evaluate this multi-dimensional integral-by-parts we require integration limits. To this end, we will assume that the domain of exploration is large enough so that the distance between x^* and x_{j+1} makes the exponential term approximately decay to 0 at the limits, which we shorthand by placing the limits at infinity:

$$\begin{aligned} I = \frac{1}{Z_d} \Lambda(x^*)^{-1} \Big[& \sqrt{\det(2\pi \Lambda^{-1}(x^*))} \\ & - (x_{j+1}-x^*)^T \mathbf{1} e^{-(x_{j+1}-x^*)^T \Lambda_{(x^*)}(x_{j+1}-x^*)} \Big|_{x_{j+1}=-\infty}^{x_{j+1}=\infty} \Big], \end{aligned} \tag{2.41}$$

where $\mathbf{1}$ is the vector of all ones, and the exponential term vanishes when evaluated at the limits. Note that our assumption on the domain of integration implies that we do not consider boundary effects, and that the quantity within the brackets is a scalar that can commute with our Lagrange multiplier matrix.

We are now ready to put together our final results. By combining Eq. 2.41 and plugging it into Eq. 2.39 we have

$$\begin{aligned} \langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \frac{1}{Z_d} \sum_{j=1}^{N-1} \prod_{i \neq j, j-1} \Big[\int_{\mathcal{X}} dx_{i+1} p_{max}(x_{i+1}|x_i) \Big] \\ \times p_{max}(x^*|x_{j-1}) \sqrt{\det(2\pi \Lambda^{-1}(x^*))} \Lambda(x^*)^{-1}. \end{aligned} \tag{2.42}$$

Since $\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*}$ is everywhere full-rank, we can see that $\Lambda(x^*)^{-1}$ must be full-rank as well. Next, we recognize that $\sqrt{\det(2\pi\Lambda(x^*)^{-1})}$ cancels out with Z_d, and that we can re-expand $p_{max}(x^*|x_{j-1})$ as an integral over $\delta(x_j - x^*)$ and fold it back into the integral product. Rearranging terms we have:

$$\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \prod_{i=1}^{N-1} \Big[\int_{\mathcal{X}} dx_{i+1}\ p_{max}(x_{i+1}|x_i) \Big] \sum_{j=1}^{N-1} \delta(x_j - x^*)\Lambda(x^*)^{-1}. \tag{2.43}$$

At this point, we note that this expression merely computes the average of $\Lambda(x^*)^{-1}$ over all possible state trajectories that pass through x^*, i.e., $E[\Lambda(x^*)^{-1}\delta(x(t) - x^*)]$. However, because $\Lambda(x^*)^{-1}$ is a constant for any given x^*, this expression reduces down to $\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = \Lambda(x^*)^{-1}$. Thus, using Eq. 2.31, we find that our Lagrange multiplier is given by:

$$\Lambda(x^*) = \mathbf{C}^{-1}[x^*]. \tag{2.44}$$

This result is significant because now we can relate a measure of temporal correlations to the sample path distribution of an optimally exploring agent. Taking this result and returning to Eq. 2.33, we now have the final form of the maximum entropy exploration sample path distribution in terms of our measure of temporal correlations:

$$P_{max}[x(t)] = \frac{1}{Z} \exp\Big[-\frac{1}{2} \int_{t_0}^{t} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)]\dot{x}(\tau) d\tau \Big], \tag{2.45}$$

where we have added a factor of one half to precisely match the path probability of diffusive spatially-inhomogeneous dynamics. This final connection can be made rigorous by noting that $\mathbf{C}[x^*]$ is an estimator of a system's local diffusion tensor through the following

relation: $\mathbf{C}[\cdot] = \frac{1}{2}\mathbf{D}[\cdot]\mathbf{D}[\cdot]^T$ for some diffusion tensor $\mathbf{D}[\cdot]$ [31, 32]. Lastly, we can discretize this distribution to arrive at the discrete-time maximum entropy sample path probability density:

$$p_{max}(x_{t+1}|x_t) = \frac{1}{Z_d} \exp\Big[-\frac{1}{2}|x_{t+1} - x_t|^2_{\mathbf{C}^{-1}[x_t]}\Big]. \tag{2.46}$$

Thus, when faced with path continuity constraints, the optimal exploration strategy is given by diffusion in state space, which concludes our derivation. In line with this, we describe systems that satisfy these statistics as *maximally diffusive*.

2.4. The Low-Rattling Selection Principle

In the previous section we largely constrained our interpretation of the maximum caliber optimization to the context of exploration. However, since the maximum caliber framework is an inference framework first and foremost, there exist alternative interpretations that are formally equivalent. Originally, we framed our derivation as an answer to the following question: What is a strategy that realizes optimal exploration in agents with continuous trajectories? Equivalently, we may ask: If all we knew about a complex system is that the local magnitude of their velocity fluctuations is bounded, then what would be our best guess as to underlying the structure of their dynamics? Thus, the statistics in Eq. 2.45 also describe the least-biased guess as to the dynamics of an arbitrary complex nonequilibrium statistical mechanical system with continuous paths through state space.

In contrast to equilibrium statistical mechanics, in nonequilibrium statistical mechanics there is no obvious analogue of energy—that is, a *local* scalar quantity capable of predicting the *global* steady-state behavior of a complex system. On the basis of Eq. 2.45, our goal

in this section will be to derive a scalar quantity that is predictive of the state occupancy statistics of a broad class of nonequilibrium stochastic processes. In other words, we are interested in deriving a quantity that is capable of predicting self-organization in a broad class of complex natural and engineered systems. Beyond the scientific value of such an endeavor, being able to predict self-organization is essential to being able to harness it towards engineering goals—especially in microsystems design, as we will see in later chapters. The results presented throughout this section are original and unpublished. However, they are heavily informed and inspired by a previous publication [6], whose results will be explored in much greater detail in the following chapter.

In order to find a scalar quantity that is predictive of the state occupancy statistics of complex nonequilibrium systems, we will explore the properties of the class of Markov chains described by Eq. 2.46. We being by defining the considering an infinite time interval, and defining the occupancy measure of a given set $\sigma \subset \mathcal{X}$ as

$$\rho(\sigma) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(x_t \in \sigma) = \sum_{t=0}^{\infty} \gamma^t \int_{\sigma} p(x_t) dx_t, \tag{2.47}$$

for some $\gamma \in (0, 1)$. The parameter γ is a nondimensional modelling parameter that weighs the relative contribution of short vs. long trajectories to the steady-state occupation statistics, as is standard in the study of Markov chains. This parameter is required in order to guarantee that the resulting occupancy statistics do not diverge. The probability density function $p(x_t)$ expresses the probability of reaching state x after t time steps. Recalling the definition of a Markovian sample path distribution in Section 2.1.3, we have

$$P[x_{0:T}] = \mu(x_0) \prod_{t=1}^{T-1} p(x_{t+1}|x_t),$$

where μ represents an initial density over the domain of the chain. We may then define $p(x_t)$ in the following way for any $t > 0$:

$$p(x_t) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \mu(x_0) \prod_{\tau=1}^{t} p(x_\tau | x_{\tau-1}) dx_0 \cdots dx_{t-1}, \tag{2.48}$$

with $p(x_0) = \mu(x_0)$. In short, an occupancy measure adds up the probability of reaching a given state through ever-increasing path lengths.

Because we are interested in finding the occupancy statistics of maximally diffusive Markov chains, let

$$p_{max}(x_t | x_{t-1}) = \frac{1}{\hat{Z}} \exp\Big[-\frac{1}{2} ||x_t - x_{t-1}||^2_{\mathbf{C}^{-1}[x_t]} \Big], \tag{2.49}$$

and define

$$\rho_{max}(\sigma) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{max}(x_t \in \sigma) = \sum_{t=0}^{\infty} \gamma^t \int_{\sigma} p_{max}(x_t) dx_t$$

in terms of the density $p_{max}(x_t)$, whose expression is the same as Eq. 2.48 except for the substitution of the expression for the maximally diffusive conditional density (i.e., Eq. 2.49). While evaluating $p_{max}(x_t)$ is nontrivial in general, we can calculate the first couple of terms in the sum of ρ_{max} analytically. Given uniform initial conditions, i.e., $\mu(x_0) = p_{max}(x_0) = 1/|\mathcal{X}|$, we have the following trivial initial measure, $\mathbb{P}_{max}(x_0 \in \sigma) = |\sigma|/|\mathcal{X}|$. Then, for $\mathbb{P}_{max}(x_1 \in \sigma)$ we have:

$$\begin{aligned} \mathbb{P}_{max}(x_1 \in \sigma) &= \int_{\sigma} \int_{\mathcal{X}} p_{max}(x_1 | x_0) p_{max}(x_0) dx_0 dx_1 = \int_{\sigma} \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} p_{max}(x_1 | x_0) dx_0 dx_1 \\ &\approx \frac{1}{|\mathcal{X}|} \int_{\sigma} \frac{1}{\hat{Z}_1} \exp\Big[-\frac{1}{2} x_1^T \mathbf{C}^{-1}[x_1] x_1 \Big] dx_1 \end{aligned}$$

where $\hat{Z}_1 = \sqrt{(2\pi)^d \det(\mathbf{C}[x_1])}$. Although the Gaussian integral formally requires no limits of integration, here we assume that for large enough $\mathcal{X}$—and for σ far enough away from $\partial\mathcal{X}$—we can ignore boundary effects. Collecting the first couple of terms, we now have:

$$\begin{aligned}\rho_{max}(\sigma) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{max}(x_t \in \sigma) \\ &= \frac{|\sigma|}{|\mathcal{X}|} + \frac{\gamma}{|\mathcal{X}|} \int_{\sigma} \frac{1}{\hat{Z}_1} \exp\Big[-\frac{1}{2} x_1^T \mathbf{C}^{-1}[x_1] x_1 \Big] dx_1 + \sum_{t=2}^{\infty} \gamma^t \mathbb{P}_{max}(x_t \in \sigma). \end{aligned} \tag{2.50}$$

To proceed further, we will assume that $\mathbf{C}[x^*]$ is bounded and positive definite for all $x^* \in \mathcal{X}$. That is, we will assume that there exist

$$\Sigma_{max} = \mathbf{C}[x^*_{max}], \quad \text{and} \quad \Sigma_{min} = \mathbf{C}[x^*_{min}] \tag{2.51}$$

with

$$x^*_{max} = \text{argmax}_{x^* \in \mathcal{X}} ||\mathbf{C}[x^*]||_2, \quad \text{and} \quad x^*_{min} = \text{argmin}_{x^* \in \mathcal{X}} ||\mathbf{C}[x^*]||_2 \tag{2.52}$$

where $||\cdot||_2$ is the spectral norm. Furthermore, we will assume that Σ_{max} and Σ_{min} are isotropic, i.e., that $\Sigma_{min} = \lambda_{min}\mathbf{I}$ and $\Sigma_{max} = \lambda_{max}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. In other words, our assumption can be summarized as stating that the transition statistics of a maximum caliber diffusion process can be bounded from above and below by two Brownian motion processes with parameters λ_{max} and λ_{min}, which set its minimum and maximum diffusivities globally. We can motivate these upper and lower bounds on system diffusivity in the following ways. The existence of an upper bound on system diffusivities is explicitly a reasonable assumption for physically-embodied systems, whose diffusivities

cannot ever be unbounded. On the other hand, the lower-bound is a reasonable assumption as a result of the full-rankness of $\mathbf{C}[x^*]$ everywhere in $\mathcal{X}$.

Now, let $\mathbb{P}^{\Sigma}_{max}(x_t \in \sigma)$ correspond to the transition measure of a Brownian motion process with constant covariance Σ in a bounded, convex set $\mathcal{X}$. Regardless of whether we choose $\Sigma = \Sigma_{max}$ or $\Sigma = \Sigma_{min}$, it is well known that the transition measure of this class of processes converges to

$$\lim_{t\to\infty} \mathbb{P}^{\Sigma}_{max}(x_t \in \sigma) = \frac{|\sigma|}{|\mathcal{X}|} \tag{2.53}$$

under reflecting boundary conditions, which we will now also assume for the purposes of this derivation. For this class of processes, it has recently been shown that the convergence rate of the transition measure is exponentially fast in t regardless of initial conditions and $dim(\mathcal{X})$ [33]. In line with these results, we make the following simplifications to our results in Eq. 2.50:

$$\begin{aligned}
\rho_{max}(\sigma) &\approx \frac{|\sigma|}{|\mathcal{X}|} + \frac{\gamma}{|\mathcal{X}|}\int_{\sigma} \frac{1}{\hat{Z}_1} \exp\Big[-\frac{1}{2}x_1^T\mathbf{C}^{-1}[x_1]x_1\Big]dx_1 + \sum_{t=2}^{\infty}\gamma^t \mathbb{P}^{\Sigma}_{max}(x_t \in \sigma) \\
&\approx \frac{|\sigma|}{|\mathcal{X}|} + \frac{\gamma}{|\mathcal{X}|}\int_{\sigma} \frac{1}{\hat{Z}_1} \exp\Big[-\frac{1}{2}x_1^T\mathbf{C}^{-1}[x_1]x_1\Big]dx_1 + \sum_{t=2}^{\infty}\gamma^t \frac{|\sigma|}{|\mathcal{X}|} \\
&= \Big(\frac{\gamma^2-\gamma+1}{1-\gamma}\Big)\frac{|\sigma|}{|\mathcal{X}|} + \frac{\gamma}{|\mathcal{X}|}\int_{\sigma} \frac{1}{\hat{Z}_1} \exp\Big[-\frac{1}{2}x_1^T\mathbf{C}^{-1}[x_1]x_1\Big]dx_1,
\end{aligned} \tag{2.54}$$

where we used the convergence properties of geometric series under the assumption that $\mathbb{P}^{\Sigma}_{max}(x_t \in \sigma)$ converges quickly to the uniform measure (i.e., for $t \geq 2$). Thus, in this approximation *long paths contribute uniformly to the background occupation measure of a*

given set. This is reminiscent to the assumption that the underlying system dynamics are "messy" in [6].

Now, we can define the occupation density of a given state (as opposed to the occupation measure of a given set) as the following limit:

$$\begin{aligned}\rho_{max}(x) &= \lim_{\epsilon \to 0^+} \frac{1}{V_\epsilon} \rho_{max}(B_\epsilon(x)) \\ &= \frac{\gamma^2 - \gamma + 1}{|\mathcal{X}|(1-\gamma)} + \frac{\gamma}{|\mathcal{X}|\hat{Z}(x)} \exp\Big[-\frac{1}{2} x^T \mathbf{C}^{-1}[x] x \Big] \\ &= \rho_{const} + \rho_{var}(x) \end{aligned} \tag{2.55}$$

where we let $\sigma = B_\epsilon(x)$ be a d-dimensional ball of radius $\epsilon > 0$ centered at x with volume V_ϵ, and $\hat{Z}(x) = \sqrt{(2\pi)^d \det(\mathbf{C}[x])}$. In short, we can see that the occupation density of any given state can be written as the sum of a constant density that assigns a "background" probability to every state in $\mathcal{X}$, and a state-dependent density that actually determines which states are more likely than others.

The relative magnitudes of ρ_{const} and $\rho_{var}(x)$ effectively determine the balance between diffusive and self-organizing behaviors in a complex nonequilibrium system. For example, if ρ_{const} is too high, then diffusive randomness will disorder the system and prevent self-organization. Because of the way that ρ_{const} and $\rho_{var}(x)$ depend on γ, we can understand the potential for self-organization in terms of the relative contribution of paths of different lengths. In the $\gamma \to 0$ limit, only short paths contribute probability mass and the system is essentially completely memoryless, with only the initial density contributing to the occupancy statistics. Interestingly, as $\gamma \to 1$ the contribution from the uniform background density increases as well (diverging at $\gamma = 1$), resulting in uniformly random occupancy

statistics again. This is because in this limit we are increasingly valuing the occupancy contributions of long paths, whose transition density converges exponentially fast to uniform, as we saw in Eq. 2.53. Thus, in order for self-organization to occur within the context of this theory, the underlying systems can neither be infinitely sensitive to every detail of every state transition, nor completely memoryless—in other words, they need to be simultaneously *dissipative* yet *adaptive* [34, 35].

For the rest of this derivation, we will now focus our attention on the $\rho_{var}(x)/\rho_{const} >> 1$ regime, where self-organization can spontaneously occur and dominate over the uniform probability background. Since we are interested in predicting the relative likelihood of states, as opposed to raw occupancy, we will also focus directly on $\rho_{var}(x)$. Because we know that $||\Sigma_{max}||_2 \geq ||\mathbf{C}[x]||_2 \geq ||\Sigma_{min}||_2$ for all $x \in \mathcal{X}$, we will look to bound the contributions of $\rho_{var}(x)$ to the occupation density through the following closely related densities:

$$\rho_{var}^{\Sigma_{min}}(x) = \frac{\gamma}{|\mathcal{X}|\hat{Z}_{min}} \exp\Big[-\frac{1}{2}x^T\Sigma_{min}^{-1}x\Big]$$
$$\rho_{var}^{\Sigma_{max}}(x) = \frac{\gamma}{|\mathcal{X}|\hat{Z}_{max}} \exp\Big[-\frac{1}{2}x^T\Sigma_{max}^{-1}x\Big].$$

Then, recalling that $\Sigma_{max} = \lambda_{max}\mathbf{I}$ and $\Sigma_{min} = \lambda_{min}\mathbf{I}$, we can write the logarithms of all relevant distributions in the following way:

$$\log\rho_{var}(x) = \log\gamma - \log|\mathcal{X}| - \frac{d}{2}\log 2\pi - \frac{1}{2}x^T\mathbf{C}^{-1}[x]x - \frac{1}{2}\log\det\mathbf{C}[x]$$
$$\log\rho_{var}^{\Sigma_{max}}(x) = \log\gamma - \log|\mathcal{X}| - \frac{d}{2}\log 2\pi - \frac{1}{2\lambda_{max}}||x||_2^2 - \frac{1}{2}\log\lambda_{max}^d$$
$$\log\rho_{var}^{\Sigma_{min}}(x) = \log\gamma - \log|\mathcal{X}| - \frac{d}{2}\log 2\pi - \frac{1}{2\lambda_{min}}||x||_2^2 - \frac{1}{2}\log\lambda_{min}^d.$$

Now, we are ready to make our final assumption and arrive at our final result. Assume that the underlying system is high-dimensional enough (i.e., $d >> 1$), and that global parameters $\lambda_{min} >> 1$ and $\lambda_{max} >> 1$ are large enough such that

$$\log \rho_{var}^{\Sigma_{max}} \approx const - \frac{1}{2} \log \lambda_{max}^{d}$$
$$\log \rho_{var}^{\Sigma_{min}} \approx const - \frac{1}{2} \log \lambda_{min}^{d}.$$

In other words, assume that the underlying system is "sufficiently complex and messy." Then, because $\frac{1}{2\lambda_{min}}||x||_2^2 \leq \frac{1}{2}x^T \mathbf{C}^{-1}[x]x \leq \frac{1}{2\lambda_{max}}||x||_2^2$ for all $x \in \mathcal{X}$, we also have

$$\log \rho_{var}(x) \approx const - \frac{1}{2} \log \det \mathbf{C}[x].$$

Lastly, let $\mathcal{R}(x) = \frac{1}{2} \log \det \mathbf{C}[x]$ be termed "rattling" and then we have

$$\rho_{var}(x) \propto e^{-\mathcal{R}(x)}. \tag{2.56}$$

Note that under this set of assumptions, the support of $\rho_{var}(x)$ is the entirety of $\mathcal{X}$, confirming the ergodicity of maximally diffusive trajectory statistics. This relationship between rattling and state occupancy is known as the "low-rattling selection principle," which was originally published in [6] under a different theoretical framing, as we will discuss in a later chapter of this thesis. As shown in Eq. 2.56, the low-rattling selection principle states that in complex systems the states that "rattle" the least are the ones with greatest occupation density. Thus, we have shown that rattling is a scalar quantity that is predictive of the state-occupancy statistics of a broad class of complex nonequilibrium statistical mechanical processes, which concludes our derivation.

Characterizing the range of settings under which this approximation holds *and is useful* is essential. Intuitively, and as we have already discussed, it holds when the system is "sufficiently messy" (i.e., when λ_{min} is large enough). However, it is also important to note that this approximation is only useful when the ratio $\lambda_{max}/\lambda_{min}$ is sufficiently large (i.e., $\lambda_{max}/\lambda_{min} >> 1$). Otherwise, $\frac{1}{2}\log\lambda_{min}^{d} \leq \frac{1}{2}\log\det\mathbf{C}[x] \leq \frac{1}{2}\log\lambda_{max}^{d}$ will not vary enough for $\rho_{var}(x)/\rho_{const} >> 1$ and meaningfully contribute to the overall occupation density $\rho_{max}(x)$. In other words, if $\lambda_{max}/\lambda_{min}$ is not much greater than 1, then ρ_{max} will be approximately uniform. Now that we are equipped with a robust theoretical understanding of the behavior of complex dynamical systems, we will explore applications of this framework in the design and control of robot collectives across scales in the following chapters—finding ways to harness and exploit randomness and self-organization towards novel task-capabilities in robotic systems.

2.5. Free Energy, Optimal Control, and Reinforcement Learning

In previous sections, we have largely focused on the role of maximum caliber as an inference framework. In order to extend from inference to *synthesis*—and thus to realize the potential of *robot thermodynamics* as a framework—we need to expand beyond entropy functional maximization and into *free energy minimization*. In this section, we will discuss free energy as means of encoding goal-directed behaviors effectively as priors within the maximum caliber framework. Then, we will discuss how one can derive results such as Pontryagin's maximum principle within our framework [21]. Lastly, we will also briefly discuss how to frame reinforcement learning (RL) problems within our framework as applications of Kullback-Leibler (KL) divergence control [36]. As in the previous section,

much of the work in this section was originally published in the supplement of [7] but original, unpublished contributions will be highlighted.

2.5.1. Directed Exploration as Maximum Caliber Trajectory Sampling

Prior to introducing control as an element into our optimizations, we will first consider introducing external notions of "importance" to system states—a prerequisite for solving optimal control and RL problems. In many exploration problems, there is an a priori understanding of what regions of the exploration domain are important or informative. For example, in RL this is encoded by the reward function [18], and in optimal control this is often encoded by a cost function or an expected information density [37, 38]. In such settings, one may want an agent to explore states while taking into account a measure of state desirability, which leads to directed (or active) exploration. In order to realize directed exploration, we require a notion of state desirability that is amenable to the statistical-mechanical construction of our approach. To this end, we can reformulate our maximum caliber objective into a "free energy" minimization objective by introducing a bounded, real-valued *potential function*, $V(\cdot)$. Across fields, potential functions are used to ascribe (either a physical or virtual) cost to system states. A potential function is then able to encode tasks in control theory, learning objectives in artificial intelligence, desirable regions in spatial coverage problems, etc. Hence, we will extend the formalism presented in the previous sections to encode goal-directed behavior by considering the effect of potential functions.

Since our maximum caliber functional is an expression over all possible trajectories, we need to adapt our definition of a potential to correctly express our notion of "free energy"

over possible system realizations. To this end, we define the trajectory averages of our potential function over $\mathcal{T} = [t_0, t]$ in the following way,

$$E_P[V[x(t)]] = \int_{\mathcal{X}^\tau} P[x(t)] \int_{t_0}^{t} V(x(\tau)) d\tau \mathcal{D}x(t), \tag{2.57}$$

which captures the average cost over all possible system paths (integrated over each possible state and time for each possible path). Note that we use the brackets to distinguish between our state-based and trajectory-based potential functions, such that $V[x(t)] = \int_{t_0}^{t} V(x(\tau)) d\tau$. Formally, we must assume that $E_P[V[x(t)]]$ is bounded, which in practice will be the case for policies and controllers derived from these principles. Our new free energy functional objective is

$$\underset{P[x(t)]}{\operatorname{argmin}} \; E_P[V[x(t)]] - \hat{S}[P[x(t)]], \tag{2.58}$$

where we use $\hat{S}[P[x(t)]]$ as a short-hand for the argument to Eq. 2.32, but note that in principle it could represent any constrained entropy functional. Thankfully, finding the optimal path distribution does not require redoing all the work carried out in Chs. 2.3.3 and 2.3.4. All that's needed is to take the variation of Eq. 2.57 with respect to $P[x(t)]$ and integrate it into the optimal path distribution. As this arithmetic is very similar to the derivation provided in the proof of Theorem 2.1, we omit it here. The resulting minimum free energy path distribution is then

$$P^V_{max}[x(t)] = \frac{1}{Z} \exp\Big[- \int_{t_0}^{t} \Big(V(x(\tau)) + \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau) \Big) d\tau \Big], \tag{2.59}$$

which corresponds to the path distribution of a diffusion process in a potential field [26]. Hence, when constrained to continuous paths the optimal *directed* exploration strategy scales the strength of local diffusion relative to the desirability of the state—lowering diffusivity if the state is desirable, and increasing it if the state is undesirable. In this sense, the net effect of the potential is to bias the diffusive exploration process. We refer to systems satisfying such statistics as *maximally diffusive with respect to the underlying potential*.

As an aside, we note that,

$$P^V_{max}[x(t)] = P_{max}[x(t)] \cdot e^{-V[x(t)]} \tag{2.60}$$

from which we can recover $P_{max}[x(t)]$ in the absence of a potential (i.e., $V[\cdot] = 0$). Moreover, we note that expression above can be applied to discrete-time settings as well:

$$P^V_{max}[x_{1:N}] = \prod_{t=1}^{N-1} p_{max}(x_{t+1}|x_t) e^{-V(x_t)}, \tag{2.61}$$

where we have discretized agent state trajectories without loss of generality. Remarkably, this path distribution resembles the form of those used in the control-as-inference literature [36]. Lastly, an interesting note is that potential functions play a similar role in free energy optimizations as priors play in maximum caliber optimizations. For example, solving

$$\underset{P[x(t)]}{\text{argmin}}\ D_{KL}(P[x(t)]||P_0[x(t)]), \tag{2.62}$$

with $P_0[x(t)] = e^{-V[x(t)]}$ is formally equivalent to solving an unconstrained free energy optimization problem. Thus, we can use potential functions to codify goal-directed priors onto maximum caliber optimizations more generally.

What are the properties of an agent that is maximally diffusive with respect to a potential? Since we already know that the sample paths of agents applying our undirected exploration strategy are Markovian, as long as the potential function and its interactions with our agent are memoryless the sample paths generated by Eq. 2.58 will continue to be as well. However, ergodicity is a more challenging property to ascertain as it depends on the properties of the underlying potential function and of our diffusion process. Nonetheless, in the following theorem we show that the trajectories of an agent successfully diffusing according to our directed exploration strategy in a non-singular potential will continue to be ergodic under some mild assumptions.

Theorem 2.2. *The minimum free energy sample paths of a stochastic process (Definition 2.1) with continuous paths (in the sense of Eq. 2.58) in a compact and connected space $\mathcal{X} \subset \mathbb{R}^d$ are ergodic.*

Proof. The proof of this theorem can be easily arrived at by extending the proof of Corollary 2.1.2. As long as $V(\cdot)$ is bounded everywhere in the domain, we may discretize the stochastic control process in space and time everywhere in the domain, as in Corollary 2.1.2. Then, we can see that $p^V_{max}(x_{t+1}|x_t) = p_{max}(x_{t+1}|x_t)e^{-V(x_t)} > 0,\ \forall x_t, x_{t+\delta t} \in \mathcal{X},\ \forall t \in \mathcal{T}$. This is because we have already shown that $p_{max}(\cdot|\cdot) > 0$ in Corollary 2.1.2, and because of the properties of the potential. Thus, the underlying Markov chain described by the

$p^V_{max}(x_{t+1}|x_t)$ transition kernel is aperiodic and all states communicate, which guarantees ergodicity and concludes our proof. □

Hence, the net effect of the potential is to reshuffle probability mass in the stationary distribution of the agent's underlying Markov chain. We note that these proofs can be carried out without discretizations by instead invoking the physics of diffusion processes, as in [39] where the authors proved that heterogeneous diffusion processes in a broad class of non-singular potentials are ergodic when the strength of the potential exceeds the strength of diffusion-driven fluctuations. However, here we limit ourselves to methods from the analysis of stochastic processes. In short, path-continuity-constrained minimum free energy exploration leads to ergodic coverage of the exploration domain with respect to the potential. We note that this is an important result when it comes to the applicability of our results in robotics and RL, as we will illustrate in future sections.

2.5.2. Optimal Control as Path Likelihood Maximization

Equipped with a means of encoding goal-directed behavior within the maximum caliber framework, we are now prepared to make the jump from inference and into *synthesis*. This is because we can largely think of cost or reward functions as potentials that ascribe preferences over states and controllers $u(t)$ (or policies $\pi(\cdot|\cdot)$). However, so far we have only sought to derive and characterize the goal-directed path distributions of autonomous stochastic processes—that is, those in which we do not explicitly model or account for the influence of control actions. In order to frame synthesis problems we will need to consider *stochastic control processes* (see Definition 2.2) instead, which lie at the heart

of the robot thermodynamics framework. We note that the results in this subsection are original, unpublished contributions of this thesis.

The primary distinguishing feature of a stochastic control process is the influence of a controller $u(t)$. In short, we think of controllers as parameters of a stochastic control process' underlying measure and, consequently, of its path distribution $P_{u(t)}[x(t)]$. For any given maximum caliber variational optimization, we fix $u(t)$ such that taking variations with respect to $P_{u(t)}[x(t)]$ is well-defined. Additionally, this allows us to treat cost and reward functions in the same way as we treat potentials, since for any fixed $u(t)$ a given cost function $L[x(t), u(t)] = \int_{t_0}^{t} l(x(\tau), u(\tau))d\tau$ is only a function of the control process' states. With this in mind, we may now outline the core procedures of the robot thermodynamics framework.

As discussed earlier, robot thermodynamics poses two key questions: "What is the structure of an optimal agent's dynamics?" and "How can such dynamics be realized?" To answer the first of these questions, we make use of the maximum caliber principle. Let $\hat{S}[P_{u(t)}[x(t)]]$ be the standard entropy functional with any given additional constraints. Then, for a given $u(t)$ and cost function $L[x(t), u(t)]$ over a time interval $[t_0, t]$, we can infer the structure of the optimal agent's path distribution by solving the following optimization:

$$\underset{P_{u(t)}[x(t)]}{\operatorname{argmin}} \; E_{P_{u(t)}}[L[x(t), u(t)]] - \hat{S}[P_{u(t)}[x(t)]]. \tag{2.63}$$

We may find the solution to this problem by taking the variation of the objective function with respect to $P_{u(t)}[x(t)]$ and setting it to zero, leading to some optimal path distribution, $P^{*}_{u(t)}[x(t)]$. However, knowing the structure of the optimal dynamics does not mean that we know how to synthesize controllers that are likely to realize desirable trajectories (i.e.,

we still need $P^*_{u^*(t)}[x(t)]$). Thus, we require a means of answering the second of our two key questions.

Throughout the rest of this subsection, we will interpret this question in the following way: *How can we choose controllers that increase the likelihood of sampling desirable trajectories?* While we explore alternative formulations in the following subsection, for now we approach this question from the perspective of the variational principle of maximum likelihood [40]. Thus, given an optimal path distribution, $P^*_{u(t)}[x(t)]$, we are interested in finding a controller that maximizes the log-likelihood of desirable agent trajectories, $\log P^*_{u(t)}[x(t)]$. That is, a controller that maximizes the following objective,

$$\underset{u(t)}{\text{argmax}} \ \log P^*_{u(t)}[x(t)], \tag{2.64}$$

whose optimizer $u^*(t)$ maximizes the likelihood of desirable agent trajectories. While we have presented this procedure as two separate questions, we note that we could also interpret them as a min-max optimization:

$$\underset{u(t)}{\text{argmax}} \left(\underset{P_{u(t)}[x(t)]}{\text{argmin}} \ E_{P_{u(t)}}[L[x(t), u(t)]] - \hat{S}[P_{u(t)}[x(t)]] \right). \tag{2.65}$$

As an example of this procedure in action, we will apply this procedure to an agent with dynamics given by $\dot{x}(t) = f(x(t), u(t))$. First, we must infer the structure of the agent's goal-directed trajectory statistics. To this end, we will formulate a minimum free energy optimization with respect to an objective $L[x(t), u(t)] = \int_{t_0}^{t} l(x(\tau), u(\tau)) d\tau$ over an interval $[t_0, t]$. Since we can only consider dynamically feasible trajectories, we must include a constraint on the agent dynamics in our optimization. This can be done in the

following way,

$$E_{P_{u(t)}}\Big[\int_{t_0}^{t}\lambda^T(\tau)\big(f(x(\tau),u(\tau))-\dot{x}(\tau)\big)d\tau\Big]=0 \tag{2.66}$$

by introducing a time-varying Lagrange multiplier $\lambda(t)$. As usual, we also include a normalization constraint. The complete optimization problem is then the following:

$$\begin{aligned}\underset{P_{u(t)}[x(t)]}{\operatorname{argmin}}\ & E_{P_{u(t)}}[L[x(t),u(t)]]+\int_{\mathcal{X}^{\mathcal{T}}}P_{u(t)}[x(t)]\log P_{u(t)}[x(t)]\mathcal{D}x(t)+\\ & E_{P_{u(t)}}\Big[\int_{t_0}^{t}\lambda^T(\tau)\big(f(x(\tau),u(\tau))-\dot{x}(\tau)\big)d\tau\Big]+\lambda_0\Big(\int_{\mathcal{X}^{\mathcal{T}}}P_{u(t)}[x(t)]\mathcal{D}x(t)-1\Big).\end{aligned} \tag{2.67}$$

Now, let $\delta F/\delta P_{u(t)}$ represent the variational derivative of the free energy optimization objective with respect to the path distribution, which we can evaluate analytically and set to zero:

$$\frac{\delta F}{\delta P_{u(t)}}=-1-\log P^*_{u(t)}[x(t)]-\lambda_0+\int_{t_0}^{t}l(x(\tau),u(\tau))+\lambda^T(\tau)(f(x(\tau),u(\tau))-\dot{x}(\tau))d\tau=0. \tag{2.68}$$

With some simple arithmetic we arrive at the optimal path distribution,

$$P^*_{u(t)}[x(t)]=\frac{1}{Z}\exp\Big[-\int_{t_0}^{t}l(x(\tau),u(\tau))+\lambda^T(\tau)(f(x(\tau),u(\tau))-\dot{x}(\tau))d\tau\Big]. \tag{2.69}$$

While we do not yet have an analytical expression for the Lagrange multiplier $\lambda(t)$, we may still formulate the maximum likelihood optimization, which leads to

$$\underset{u(t)}{\operatorname{argmin}}\int_{t_0}^{t}l(x(\tau),u(\tau))+\lambda^T(\tau)(f(x(\tau),u(\tau))-\dot{x}(\tau))d\tau, \tag{2.70}$$

where we note that we changed the sign on the objective function and changed the maximization into a minimization. Instead of using Lagrange multipliers, we could have equivalently written the following optimization

$$\begin{aligned} &\underset{u(t)}{\operatorname{argmin}} \int_{t_0}^{t} l(x(\tau), u(\tau))d\tau, \\ &s.t.\ \dot{x}(\tau) = f(x(\tau), u(\tau)), \forall \tau \in [t_0, t], \end{aligned} \tag{2.71}$$

which we recognize as the standard form of an optimal control problem. Thus, maximizing the log-likelihood of desirable maximum caliber sample paths is equivalent to optimal control.

Returning to Eq. 2.70, we will proceed by deriving first-order optimality conditions for our objective. That is, taking the variations of our objective function with respect to $x(t)$ and $u(t)$. To this end, let

$$\begin{aligned} \mathcal{L}(x(t), u(t), \lambda(t)) &= \int_{t_0}^{t} l(x(\tau), u(\tau)) + \lambda^T(\tau)(f(x(\tau), u(\tau)) - \dot{x}(\tau))d\tau \\ &= \int_{t_0}^{t} l(x(\tau), u(\tau)) + \lambda^T(\tau) f(x(\tau), u(\tau)) - \lambda^T(\tau)\dot{x}(\tau)d\tau \end{aligned} \tag{2.72}$$

be a Lagrangian. Using integration by parts on the last of these terms:

$$-\int_{t_0}^{t} \lambda^T(\tau)\dot{x}(\tau)d\tau = \lambda^T(t_0)x(t_0) - \lambda^T(t)x(t) + \int_{t_0}^{t} \dot{\lambda}^T(\tau)x(\tau)d\tau. \tag{2.73}$$

This can then be substituted back into Eq. 2.72 to get

$$\mathcal{L}(x(t), u(t), \lambda(t)) = \int_{t_0}^{t} \Big(l(x(\tau), u(\tau)) + \lambda^T(\tau) f(x(\tau), u(\tau)) + \dot{\lambda}^T(\tau) x(\tau) \Big) d\tau \\ + \lambda^T(t_0) x(t_0) - \lambda^T(t) x(t). \tag{2.74}$$

Lastly, to satisfy first-order optimality conditions we know that $\frac{\delta \mathcal{L}}{\delta x(t)} = 0$ and $\frac{\delta \mathcal{L}}{\delta u(t)} = 0$. Performing these operations we arrive at the following system of ordinary differential equations equations:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ \dot{\lambda}(t) &= -\frac{\delta f}{\delta x} \lambda(t) - \frac{\delta l}{\delta x}^T \\ 0 &= \frac{\delta l}{\delta u} + \lambda^T(t) \frac{\delta f}{\delta u}. \end{aligned} \tag{2.75}$$

These differential equations (and their boundary conditions) describe the conditions that the maximum likelihood sample paths of our stochastic control process must satisfy in order for our controller to be optimal. These conditions are also known as *Pontryagin's maximum principle* [21], and they form the basis of a broad swath of the field of optimal control. Thus, the framework of robot thermodynamics is capable of reproducing canonical results in control theory. As we will show in the following section, this framework is also capable of leading us in novel directions.

2.5.3. Reinforcement Learning as Path-Based KL-Control

In the previous subsection, we formulated an optimization problem in order to find a controller that maximizes the likelihood of desirable agent trajectories, and in doing so we

recovered some key results in optimal control. However, in doing so we implicitly assumed that the underlying controller was deterministic. Here, we consider the alternative scenario in which we cannot think of controllers as parameters, and instead we use stochastic policies to specify the statistics of agent behavior. In doing so, we reformulate the second question of robot thermodynamics into: *How choose policies that match the trajectory statistics of desirable trajectories?* We note that much of the contents of this subsection contents are original, unpublished contributions of this thesis.

As a result of this question's connection to policies and MDPs, throughout this subsection we switch to working with discrete-time representations of path distributions. In turn, this means that our maximum caliber inference procedure changes into the following form,

$$\underset{P[x_{1:T},u_{1:T}]}{\operatorname{argmin}} \; E_P[L[x_{1:T},u_{1:T}]] - \hat{S}[P[x_{1:T},u_{1:T}]], \tag{2.76}$$

where the path distribution is now a joint distribution that treats control sequences as sample paths of an separate stochastic process. Now, let $\hat{S}[P[x_{1:T},u_{1:T}]] = S[P[x_{1:T},u_{1:T}]] + \mathcal{C}[P[x_{1:T},u_{1:T}]]$, where $\mathcal{C}[P[x_{1:T},u_{1:T}]]$ represents any constraints we may impose on the path distribution. Then, as previously discussed in Eq. 2.10, we may write Eq. 2.76 equivalently as

$$\underset{P[x_{1:T},u_{1:T}]}{\operatorname{argmin}} \; E_P[L[x_{1:T},u_{1:T}]] + D_{KL}(P[x_{1:T},u_{1:T}]||P_0[x_{1:T},u_{1:T}]) + \mathcal{C}[P[x_{1:T},u_{1:T}]] \tag{2.77}$$

with $P_0[x_{1:T},u_{1:T}] = P_{uniform}$, which we may then attempt to solve in the usual way. However, because policies and state transition models are explicitly probabilistic objects, incorporating constraints as Lagrange multipliers is less natural. Instead, we may impose

requirements on path distributions by using priors other than $P_{uniform}$ in Eq. 2.77 and letting $\mathcal{C}$ be zero, i.e.,

$$\underset{P[x_{1:T},u_{1:T}]}{\operatorname{argmin}} E_P[L[x_{1:T},u_{1:T}]] + D_{KL}(P[x_{1:T},u_{1:T}]||P_0[x_{1:T},u_{1:T}]). \tag{2.78}$$

In general, solutions to such optimizations will take the form

$$P_0^*[x_{1:T},u_{1:T}] = P_0[x_{1:T},u_{1:T}]e^{\sum_{t=1}^{T}\gamma^t r(x_t,u_t)}, \tag{2.79}$$

where the influence of the prior factors and where we let $-L[x_{1:T},u_{1:T}] = \sum_{t=1}^{T}\gamma^t r(x_t,u_t)$ to match the discounted MDP setting where objectives are typically framed in terms of reward functions $r(x_t,u_t)$ instead of cost functions.

Despite differences in how constraints are taken into account, the end result of this inference procedure is the same—we infer an optimal path distribution $P_0^*[x_{1:T},u_{1:T}]$. And, as before, *we still require a means of choosing control actions such that the underlying process satisfies the statistics specified by* P_0^*. However, since we cannot directly choose control actions to maximize the likelihood of desirable behaviors, we will instead optimize agent policies such that the statistics of agent trajectories match the statistics of a desired path distribution. We can pose an optimization that achieves this from the perspective of KL-control [36]. First, recall that for a given discounted MDP $(\mathcal{X},\mathcal{U},p,r,\gamma)$ over $\mathcal{T} = \{1,\cdots,T\}$ an agent's path distribution takes the form,

$$P_\pi[x_{1:T},u_{1:T}] = \prod_{t=1}^{T-1} p(x_{t+1}|x_t,u_t)\pi(u_t|x_t). \tag{2.80}$$

Then, given some P_0^* whose statistics we want our agent to match, we may frame a policy optimization problem as follows

$$\underset{\pi}{\text{argmin}}\ D_{KL}(P_\pi[x_{1:T}, u_{1:T}]||P_0^*[x_{1:T}, u_{1:T}]), \tag{2.81}$$

whose solutions may drastically vary as a function of our choice of priors P_0.

While this KL-control procedure may seem idiosyncratic, in what follows of this section we will illustrate how it can connect more broadly to stochastic optimal control, as well as well-known reinforcement learning techniques. To begin to see this, let $P_0^*[x_{1:T}, u_{1:T}] = \prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)e^{\gamma^t r(x_t,u_t)}$ be the target distribution in our optimization in Eq. 2.81. By placing the agent's policy and state transition model directly into the target distribution, we ask the optimization to seek policies that preserve the underlying structure of the system's path statistics. Instead, the optimization will seek policies that steer the system's predetermined path structure towards highly rewarding regions, as desired. To see this more clearly, we may use P_0^* in Eq. 2.81 and rearrange:

$$\begin{aligned} D_{KL}(P_\pi[x_{1:T}, u_{1:T}]||P_0^*[x_{1:T}, u_{1:T}]) &= E_{P_\pi}\left[\log \frac{P_\pi[x_{1:T}, u_{1:T}]}{P_0^*[x_{1:T}, u_{1:T}]}\right] \\ &= E_{P_\pi}\left[\log \frac{\prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)}{\prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)e^{\gamma^t r(x_t,u_t)}}\right] \\ &= E_{P_\pi}\left[-\sum_{t=1}^{T-1} \gamma^t r(x_t, u_t)\right]. \end{aligned}$$

Once we change the sign of this expression and reformulate the problem as a maximization, we have the following objective

$$\underset{\pi}{\text{argmax}}\ E_{P_\pi}\left[\sum_{t=1}^{T-1}\gamma^t r(x_t,u_t)\right], \tag{2.82}$$

which we note is the standard form of stochastic optimal control problems, providing an important link between robot thermodynamics and the control-as-inference literature [36]. Moreover, because RL problems are typically framed in the formalism of stochastic optimal control, we also have the ability to frame RL in terms of the robot thermodynamics formalism, which we will explore in what remains of this subsection, as well as throughout this thesis.

Before we conclude this section, we will consider one additional choice of prior as an example. Let $P_0^*[x_{1:T},u_{1:T}]=\prod_{t=1}^{T-1}p(x_{t+1}|x_t,u_t)\pi_{uniform}e^{r(x_t,u_t)}$ in Eq. 2.81, noting that we set $\gamma=1$ for convenience but without lack of generality. Hence, our optimization will seek goal-directed policies that preserve the underlying structure of the system's state transition dynamics while also seeking to match the statistics of a uniformly random policy. In other words, our optimization will seek to *maximize policy entropy*. Plugging P_0^* into Eq. 2.81 we see the following:

$$\begin{aligned}
D_{KL}(P_\pi[x_{1:T},u_{1:T}]||P_0^*[x_{1:T},u_{1:T}]) &= E_{P_\pi}\left[\log\frac{\prod_{t=1}^{T-1}p(x_{t+1}|x_t,u_t)\pi(u_t|x_t)}{\prod_{t=1}^{T-1}p(x_{t+1}|x_t,u_t)\pi_{uniform}e^{r(x_t,u_t)}}\right]\\
&= E_{P_\pi}\left[\sum_{t=1}^{T-1}\log\frac{\pi(u_t|x_t)}{\pi_{uniform}}-r(x_t,u_t)\right]\\
&= E_{P_\pi}\left[\sum_{t=1}^{T-1}\log\pi(u_t|x_t)-r(x_t,u_t)+const\right].
\end{aligned}$$

Disregarding the constant term and flipping the sign once again, we arrive at the following optimization objective:

$$\underset{\pi}{\text{argmax}}\ E_{P_\pi}\left[\sum_{t=1}^{T-1} r(x_t, u_t) - \log \pi(u_t|x_t)\right]. \tag{2.83}$$

Noting that the expectation of the logarithm of the policy is an entropy term, we may also rewrite this objective as

$$\underset{\pi}{\text{argmax}}\ E_{P_\pi}\left[\sum_{t=1}^{T-1} r(x_t, u_t) + \alpha S[\pi(\cdot|x_t)]\right], \tag{2.84}$$

where we introduced $\alpha > 0$ as a parameter. As expected, we have arrived at a policy optimization that maximizes policy entropy while optimizing agent rewards. Importantly, we recognize this objective as maximum entropy reinforcement learning (MaxEnt RL) objective [18, 19, 41, 42], which constitutes a major family of state-of-the-art algorithms in RL that seeks to improve exploration during learning by maximizing policy entropy. Thus, using robot thermodynamics we were once again able to reproduce canonical results in stochastic optimal control and RL—namely, the KL-control and stochastic optimal control duality, as well as the derivation of maximum entropy methods in RL. We were able to do this with remarkable ease by simply incorporating constraints as priors during our maximum caliber inference procedure. In the final section of this chapter, we will combine this same procedure with all of the technical machinery we have been building up throughout this chapter in order to derive a novel RL framework built with embodied agents in mind.

2.6. The Maximum Diffusion Reinforcement Learning Framework

In what remains of this chapter, we will use the tools we developed in previous sections to derive a novel approach to embodied learning and decision-making. Throughout this thesis, we are interested in examining the role that embodiment plays in the resulting behavior and performance of autonomous systems. As we alluded to in Ch. 2.3.1, for an agent to be embodied is for them to be localized in space and time, requiring them to take continuous paths through space and time. Then, as we saw in Ch. 2.3.2, when an agent is required to take continuous paths through state-space their experiences become correlated, leading to violations of the *i.i.d.* property, which creates major issues for any learning process relying on these experiences as data. In Chs. 2.3.4 and 2.5.1, we found a solution to this issue: Violations of the *i.i.d.* property can be avoided when the underlying agent is *maximally diffusive*—or, ergodic, more generally. As before, many results in this section are drawn from the supplement of [7].

2.6.1. Deriving the MaxDiff RL Objective

Much like how the MaxEnt RL framework seeks to encourage better exploration through policy entropy maximization, in this section we are interested in deriving an exploration strategy for embodied RL agents capable of overcoming violations of the *i.i.d.* property. To this end, we will return to the KL-control-based optimization framework discussed in the previous section. In particular, we consider the following objective,

$$\underset{\pi}{\text{argmin}}\ D_{KL}(P_\pi[x_{1:T}, u_{1:T}]||P^*_{max}[x_{1:T}, u_{1:T}]), \tag{2.85}$$

where $P^*_{max}[x_{1:T}, u_{1:T}] = \prod_{t=1}^{T-1} p_{max}(x_{t+1}|x_t)\pi_{uniform}e^{r(x_t,u_t)}$ or, equivalently, we have $P^*_{max}[x_{1:T}, u_{1:T}] = \prod_{t=1}^{T-1} p_{max}(x_{t+1}|x_t)e^{r(x_t,u_t)}$. Thus, our policy optimization will seek to find a goal-directed policy that forces the system's dynamics to match maximally diffusive state transition statistics while also maximizing the underlying policy's entropy. Proceeding from Eq. 2.85 we have

$$\begin{aligned} D_{KL}(P_\pi[x_{1:T}, u_{1:T}]||P^*_{max}[x_{1:T}, u_{1:T}]) &= E_{P_\pi}\left[\log \frac{\prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)}{\prod_{t=1}^{T-1} p_{max}(x_{t+1}|x_t)e^{r(x_t,u_t)}}\right] \\ &= E_{P_\pi}\left[\sum_{t=1}^{T-1} \log \frac{p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)} - r(x_t, u_t)\right], \end{aligned}$$

which we can rearrange into the following optimization problem

$$\underset{\pi}{\text{argmax}}\ E_{P_\pi}\left[\sum_{t=1}^{T-1} r(x_t, u_t) + \alpha \log \frac{p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)}\right], \tag{2.86}$$

where we introduced $\alpha > 0$ as a temperature-like parameter to balance between the drive to optimize rewards and the drive to match maximally diffusive trajectory statistics. As currently expressed in terms of state-transitions models and policies, we refer to Eq. 2.86 as the *MaxDiff RL* objective. As we will see in Ch. 5, MaxDiff RL is a novel reinforcement learning framework that resolves issues with the *i.i.d.* property and can provide formal guarantees on the behavior and performance of embodied RL agents. For now, we only mean to motivate its derivation from the principles of robot thermodynamics we have laid out in this chapter.

2.6.2. Deriving the MaxDiff Trajectory Synthesis Objective

Before concluding, we will derive an alternative form of the objective in Eq. 2.86 that is more amenable to applications outside of RL, such as trajectory optimization and sampling-based control. Instead of using the KL-control approach shown above, we can make use of the fact that in Ch. 2.3.4 we proved that $P_{max}[x(t)]$ is the unique probability distribution that optimizes $S[P[x(t)]]$ under a path continuity constraint. For convenience, we will ignore cost or reward functions in this derivation. We note we can do this without loss of generality because we also proved analogous results in Ch. 2.5.1 for optimizations in the presence of potentials, costs, and rewards as long as these functions satisfy some assumptions. Given

$$u^*(t) = \underset{u(t)}{\operatorname{argmax}}\ S[P_{u(t)}[x(t)]], \tag{2.87}$$

we know that $S[P_{max}[x(t)]] \geq S[P_{u(t)}[x(t)]]$ in general, with $S[P_{max}[x(t)]] = S[P_{u^*(t)}[x(t)]]$ if and only if $P_{max}[x(t)] = P_{u^*(t)}[x(t)]$ and attaining the global optimum is feasible. Thus, since there is no path distribution with greater entropy we may use control to directly maximize the entropy of our embodied agent's path distribution, and—if we succeed—to realize maximally diffusive state transition statistics.

With this in mind, in order to proceed further we will make an optimistic assumption: Assume that the underlying agent's path statistics are within a *local* variational neighborhood of the optimal path statistics. We can formalize this assumption by asserting that our agent's path probability densities are of the following functional form:

$$P^L_{u(t)}[x(t)] = \frac{1}{Z} \exp\Big[-\frac{1}{2} \int_{t_0}^{t} \dot{x}(\tau)^T \hat{\mathbf{C}}^{-1}_{u(t)}[x(\tau)] \dot{x}(\tau) d\tau \Big] \tag{2.88}$$

where it is still the case that $S[P_{max}[x(t)]] \geq S[P^L_{u(t)}[x(t)]]$, and that the optimum can only be reached if and only if we can find $u^*(t)$ such that $P_{max}[x(t)] = P^L_{u^*(t)}[x(t)]$. In other words, since there is no path distribution with greater entropy than $P_{max}[x(t)]$, *and* because $P^L_{u(t)}[x(t)]$ matches its functional form, we may optimize

$$\underset{u(t)}{\text{argmax}}\ S[P^L_{u(t)}[x(t)]], \tag{2.89}$$

instead and reach find the same optimum as Eq. 2.87. The matrix $\hat{\mathbf{C}}_{u(t)}[x^*]$ is an empirical estimate of the local temporal correlations (as we defined them in Ch. 2.3.2). Importantly, the optimum of the optimization is reached if and only if $\mathbf{C}[x^*] = \hat{\mathbf{C}}_{u(t)}[x^*]$ for all $x^* \in \mathcal{X}$. Computing $\hat{\mathbf{C}}_{u(t)}[x^*]$ is simple: Using $u(t)$, we forward simulate short rollouts of the system initialized at x^* and evaluate their covariance across system trajectories. Alternatively, if we do not have access to predictive system rollouts we may evaluate the covariance along an individual trajectory. However, this is only equivalent to the previous procedure if the underlying dynamics are ergodic. Thus, by optimizing $S[P^L_{u(t)}[x(t)]]$, we merely change the direction from which our system approaches the same variational optimum.

Why go through this trouble? Because we can find an analytical expression for $S[P^L_{u(t)}[x(t)]]$ that is concave and can be efficiently optimized. To do so, we will discretize our trajectory distribution into $P^L_{u_{1:T}}[x_{1:T}]$ and make use of the chain rule of conditional entropies, which is

$$S[P[x_{1:T}]] = \sum_{t=1}^{T-1} S[p(x_{t+1}|x_{1:t})] \tag{2.90}$$

in general. However, when the process is Markovian the right hand side of the expression simplifies significantly. Then, applying this rule to $P^L_{u_{1:T}}[x_{1:T}]$ we have,

$$S[P^L_{u_{1:T}}[x_{1:T}]] = \sum_{t=1}^{T-1} S[p^L_{u_{t:T}}(x_{t+1}|x_t)] \propto \sum_{t=1}^{T-1} \frac{1}{2} \log\det \hat{\mathbf{C}}_{u_{t:T}}[x_t], \tag{2.91}$$

where we made use of the Markov property to simplify our sum over conditional entropies, and then the analytical form of the entropy of a Gaussian distribution (up to a constant offset) to reach our final expression.

Equipped with this result in hand, we may now rewrite the MaxDiff RL objective as

$$\underset{\pi}{\text{argmax}}\ E_{P_\pi}\left[\sum_{t=1}^{T-1} r(x_t, u_t) + \frac{\alpha}{2} \log\det \hat{\mathbf{C}}_\pi[x_t]\right], \tag{2.92}$$

where we used $\hat{\mathbf{C}}_\pi[x_t]$ instead because of our use of a policy instead of a controller. Nonetheless, $\hat{\mathbf{C}}_\pi[x_t]$ is computed in the same way as before for a given policy. Equivalently, we may write this objective as an optimal control problems in terms of cost functions,

$$\underset{u(t)}{\text{argmin}} \int_{t_0}^{t} l(x(\tau), u(\tau)) - \frac{\alpha}{2} \log\det \hat{\mathbf{C}}_{u(\tau)}[x(\tau)] d\tau. \tag{2.93}$$

We refer to both of these objectives as the *MaxDiff Trajectory Synthesis* objectives.

2.6.3. Examples of MaxDiff Trajectory Synthesis

In what remains of this section, we implement MaxDiff trajectory synthesis across handful of applications *outside of reinforcement learning* that require both directed and undirected exploration. This is because reinforcement learning will be discussed at length in Ch. 5. These examples should illustrate the sense in which maximally diffusive trajectories can

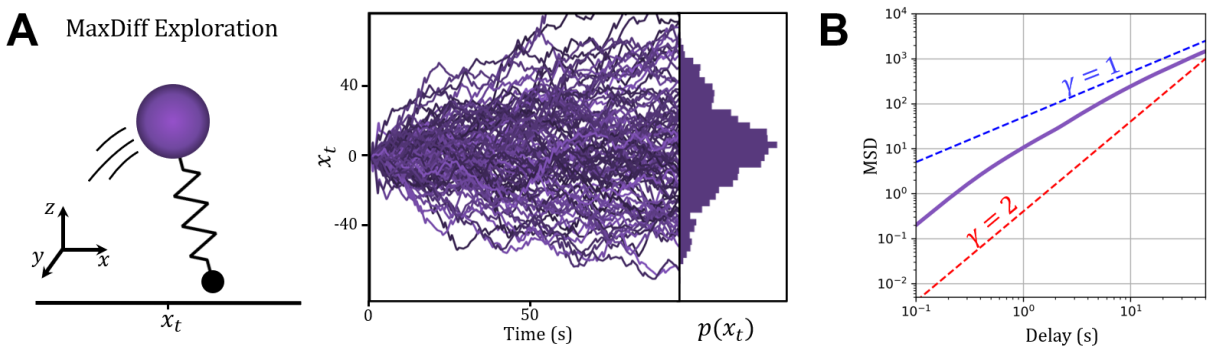


Figure 2.5. **Maximally diffusive trajectories of a spring-loaded inverted pendulum (SLIP)**. (A) The SLIP model (left panel) is a 9-dimensional nonlinear and nonsmooth second-order dynamical system, which is used as a popular model of human locomotion. (right panel) We choose this system because it is far from the ideal assumptions under which our theory is formulated, and yet its sample paths behave as we expect. The sample paths of the SLIP model with MaxDiff trajectories in the one dimensional space determined by its x-coordinate approximately match the statistics of pure Brownian motion in one dimension. (B) Mean squared displacement (MSD) plots give the deviation of the position of an agent over time with respect to a reference position. We can distinguish between diffusion processes by comparing the growth of their MSD over time. In general, we expect them to follow a relationship described by $\text{MSD}(x) \propto t^{\gamma}$, where γ is an exponent that determines the different diffusion regimes (normal diffusion $\gamma = 1$, superdiffusion $1 < \gamma < 2$, ballistic motion $\gamma \geq 2$). As we can see, the behavior of the diffusing SLIP model is superdiffusive at short time-scales, but gradually becomes more like a standard diffusion process as we coarse-grain. Similar short-delay superdiffusion regimes have been observed in systems with nontrivial inertial properties [43], such as those of our macroscopic SLIP agent.

be useful across problem settings in robotics. Moreover, here we will analyze the behavior of various dynamical systems made to follow maximally diffusive trajectories through the lens of statistical mechanics.

We begin by studying MaxDiff trajectory synthesis in the undirected exploration of a nontrivial control system—a spring-loaded inverted pendulum (SLIP) model. The SLIP model is a popular dynamic model of locomotion and encodes many important properties

of human locomotion [44]. In particular, we will implement the SLIP model as in [45], where it is described as a 9-dimensional nonlinear nonsmooth control system. The SLIP model is shown in Fig. 2.5(A) and consists of a "head" which carries its mass, and a "toe" which makes contact with the ground. Its state-space is defined by the 3D velocities and positions of its head and toe, or $x = [x_h, \dot{x}_h, y_h, \dot{y}_h, z_h, \dot{z}_h, x_t, y_t, q]^T$, where $q = \{c, a\}$ is a variable that tracks whether the system is in contact with the ground or in the air. The SLIP dynamics are the following:

$$\dot{x} = f(x,u) = \begin{cases} f_c(x,u), & \text{if } l_c < l_0 \\ f_a(x,u), & \text{otherwise} \end{cases},$$

$$f_c(x,u) = \begin{bmatrix} \dot{x}_h \\ \frac{(k(l_0-l_s)+u_c)(x_h-x_t)}{ml_c} \\ \dot{y}_h \\ \frac{(k(l_0-l_c)+u_c)(y_h-y_t)}{ml_c} \\ \dot{z}_h \\ \frac{(k(l_0-l_c)+u_c)(z_h-z_t)}{ml_c} - g \\ 0 \\ 0 \end{bmatrix}, \; f_a(x,u) = \begin{bmatrix} \dot{x}_h \\ 0 \\ \dot{y}_h \\ 0 \\ \dot{z}_h \\ -g \\ \dot{x}_h + u_{t_x} \\ \dot{y}_h + u_{t_y} \end{bmatrix}, \tag{2.94}$$

where $f_c(x,u)$ captures the SLIP dynamics during contact with the ground, and $f_a(x,u)$ captures them while in the air. During contact the SLIP can only exert a force, u_c, by pushing along the axis of the spring, whose resting length is l_0 and its stiffness is k. During flight the SLIP is subject to gravity, g, and is capable of moving the x, y-position of its toe by applying u_{t_x} and u_{t_y}, respectively. To finish specifying the SLIP dynamics, and

determine whether or not the spring is in contact with the ground, we define,

$$l_c = \sqrt{(x_h - x_t)^2 + (y_h - y_t)^2 + (z_h - z_G)^2},$$

which describes the distance along the length of the spring to the ground, and z_G is the ground height. Rather than explore diffusively in the entirety of the SLIP model's 9-dimensional state-space, we will first only try to explore the 1-dimensional subspace described by its x-coordinate, starting from an initial condition of $x(0) = 0$. We can think of this as a projection to a 1-dimensional subspace of the system, or equivalently as a coordinate transformation with a constant Jacobian matrix. In general, we may define our covariance matrices in terms of other coordinates as $\mathbf{C}[y^*] = \mathbf{J}_\psi[x^*]\mathbf{C}[x^*]\mathbf{J}_\psi[x^*]^T$, where $\mathbf{J}_\psi[\cdot]$ is the Jacobian matrix corresponding to the coordinate transformation $y^* = \psi(x^*)$. We note that the system's nonsmoothness should break the path continuity constraint that our approach presumes to hold. However, since we use a coordinate transformation to formulate the exploration problem in terms of the system's x-coordinate we do not violate the assumptions of MaxDiff trajectory synthesis. This is because, while the system's velocities experience discontinuities, its position coordinates do not. In general, the use of coordinate transformations can extend the applicability of MaxDiff trajectory synthesis to even broader classes of systems. Formalizing this, however, would require a formal analysis of observability, which is outside the scope of this thesis.

In order to practically implement MaxDiff trajectory synthesis across the following examples, we make use of Model-Predictive Path Integral Control (MPPI) [46] in conjunction with the objective in Eq. 2.93. Figure 2.5(A) depicts the sample paths generated by the maximally diffusive exploration of the SLIP model's x-coordinate. The sample paths

of the SLIP agent resemble the empirical statistics of Brownian particle paths despite the fact that the SLIP model is far from a non-inertial point mass. In Fig. 2.5(B), we study the fluctuations of maximally diffusive exploration from the lens of statistical mechanics. Here, we analyze the mean squared displacement (MSD) statistics of undirected maximally diffusive exploration and compare to the statistics of standard and anomalous diffusion processes. MSD plots capture the deviations of a diffusing agent from some reference position over time. In standard diffusion processes, the relationship between MSD and time elapsed is linear on average. That is, we expect the squared deviation of a diffusing agent from its initial condition to grow linearly in proportion to the time elapsed (see blue line in Fig. 2.5(B)). However, in general there exist other diffusion regimes characterized by the growth of MSD over time. These regimes are typically determined by fitting the exponent γ in $\text{MSD}(x) \propto t^{\gamma}$, where normal diffusion has $\gamma = 1$, superdiffusion has $1 < \gamma < 2$, and ballistic motion has $\gamma \geq 2$. The purple line in Fig. 2.5(B) depicts the MSD statistics of the SLIP model. The diffusion generated by the SLIP model's maximally diffusive exploration has superdiffusive displacements over short-time scales owing to the the inertial properties of the system. However, as we consider longer time-scales, the behavior of the SLIP model becomes indistinguishable from standard diffusion processes with $\gamma = 1$. This difference in scaling exponents has been shown to be a general property of diffusion with inertial particles and should be expected in macroscopic systems [43].

Keeping with the SLIP dynamical system, in Fig. 2.6 we study the behavior of MaxDiff trajectory synthesis across various standard robotics applications. In Fig. 2.6(A), a single SLIP agent is performing undirected MaxDiff exploration within the bounds of an **N**-shaped environment. In this task, the agent must be able to explore its x-y plane by hopping

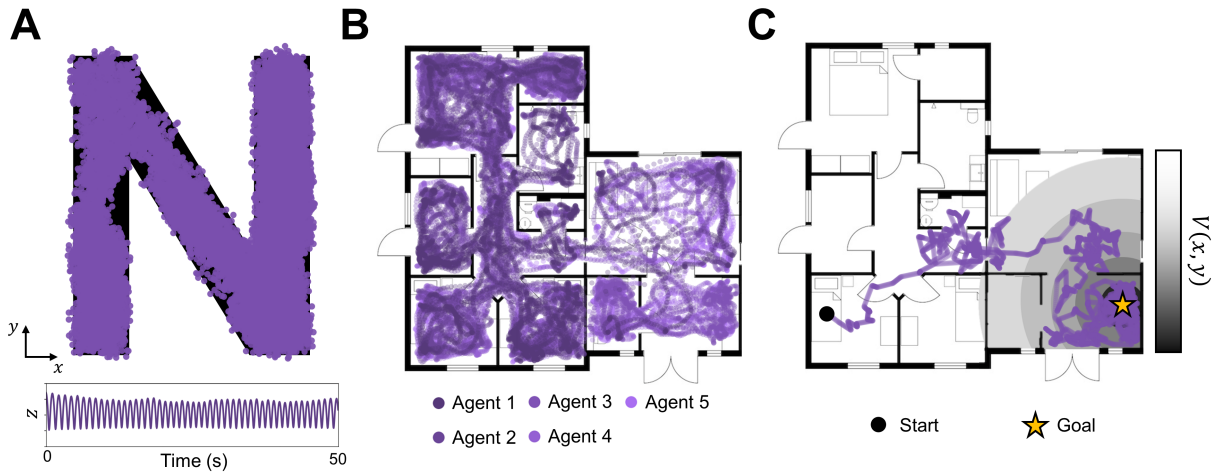


Figure 2.6. **SLIP maximally diffusive exploration in various settings**. (A) Undirected maximally diffusive exploration in a constrained **N**-shaped environment. The boundaries of the environment, as well as safety constraints, are established through the use of control barrier functions, which enable safe and continuous maximally diffusive exploration without modifications to our approach. (B) Undirected multiagent maximally diffusive exploration of more complex environment: a house's floor plan. Here, five agents with identical objectives perform maximally diffusive exploration. Because maximally diffusive exploration is ergodic, many tasks are inherently distributable between agents with linear scaling in complexity. (C) Directed maximally diffusive exploration in a complex environment. Here, a single agent in a complex environment performs directed exploration in a potential that encodes a navigation goal.

along, without falling or exiting the bounds of the exploration domain. To ensure the SLIP model's safety, as well as establish the bounds of the environment, we made use of control barrier functions (CBFs) [47]—a standard technique in the field for guaranteeing safety. Then, to illustrate another application application of the ergodicity guarantees of our method, in Fig. 2.6(B) we apply MaxDiff trajectory synthesis to multiagent exploration in a complex environment—a house floor plan—in conjunction with CBFs. Since maximally diffusive exploration is ergodic, the outcomes of a multiagent execution and a single agent execution are asymptotically identical. In this way, distributed maximally diffusive

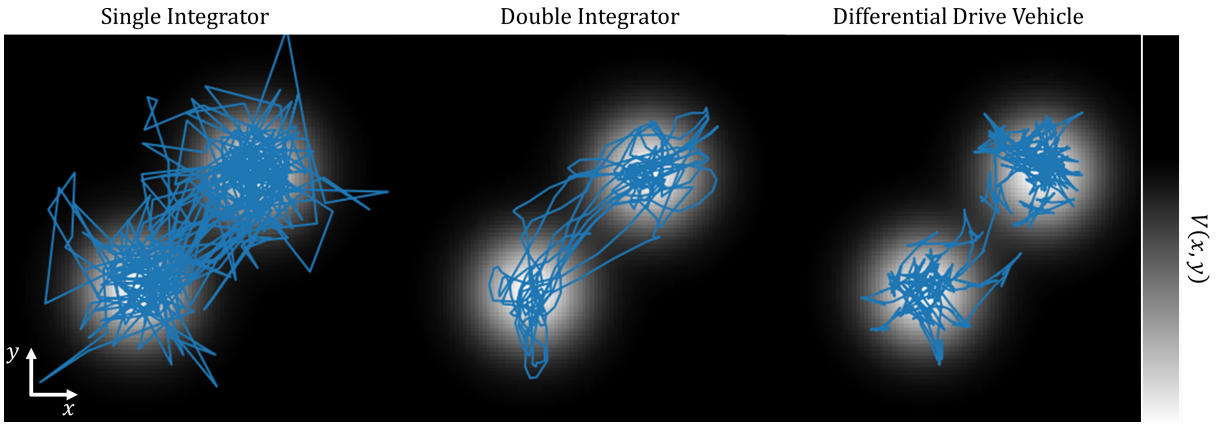


Figure 2.7. **Directed maximally diffusive exploration of bimodal potential across systems**. (left panel) The single integrator is a linear system whose velocities are directly determined by the controller. Hence, its sample paths behave exactly as free Brownian particles in a potential. (middle panel) The double integrator is the second-order equivalent of the single integrator system. In this system, the controller inputs acceleration commands that the system then integrates subject to its inertial properties. Despite being an inertial system, its interactions with the potential approximately follow the behavior of a Brownian particle in a potential. (right panel) The differential drive vehicle is a car-like system with simple nonlinear and nonholonomic dynamics with more complex controllability properties. Nonetheless, when we subject the differential drive vehicle to directed maximally diffusive exploration it traverses the potential as desired.

exploration only incurs a linear scaling in computational complexity as a function of the number of agents. Finally, in Fig. 2.6(C) we return to the single agent case to illustrate directed maximally diffusive exploration in the same complex environment as before. Here, a potential function encoding a goal destination is flat beyond a certain distance, which leads to undirected exploration initially. However, as the agent nears the goal, it can detect variations in the potential and follows its gradients diffusively towards the goal.

Now, we will highlight how the underlying properties of an agent's dynamics can affect the trajectories generated during maximally diffusive exploration. To this end, we consider a simple planar exploration task subject to a bimodal Gaussian potential ascribing a cost

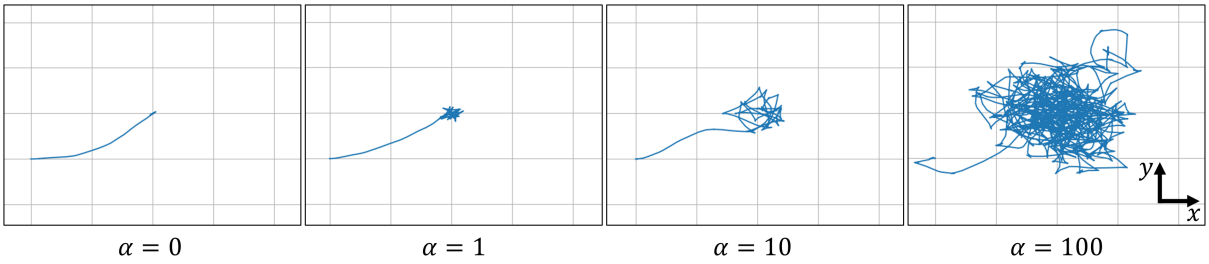


Figure 2.8. **Varying the α parameter of directed MaxDiff exploration**. Here, we are making a differential drive vehicle explore a quadratic potential centered at the origin under varying choices of α modulating the strength of the diffusive exploration within the potential. As we increase α the strength of the diffusion increases as well, leading to greater exploration of the basin of attraction of the quadratic potential well.

to system states far away from the distribution means. In Fig. 2.7, we explore the planar domain with three different systems. First, exploration over the bimodal potential is shown with a single integrator system, which is a controllable first-order linear system. Since this system is effectively identical to a non-inertial point mass, its sample paths are formally the same as those of Brownian particles in a confining potential. In the middle panel of Fig. 2.7, we consider a double integrator system, which is a controllable, linear, second-order system. However, for this system its diffusion tensor is degenerate because the noise only comes into the system as accelerations. Nonetheless, the system realizes ergodic coverage with respect to the underlying potential (in agreement with the theory of degenerate diffusion [48, 49]). Finally, we consider the differential drive vehicle, which is a simple first-order nonlinear dynamical system with nontrivial controllability properties. Yet, the differential drive vehicle realizes ergodic coverage in the plane, as predicted by the properties of maximally diffusive systems.

As a final look into the properties of directed maximally diffusive exploration, we examine the role that the temperature parameter α plays on the behavior of the agent

in a simpler setting. To this end, we revisit the differential drive vehicle dynamics and make use of MPPI once again to optimize our objective. However, instead of a bimodal Gaussian potential, we consider a quadratic potential centered at the origin with the system initialized at $(x, y) = (-4, -2)$. Quadratic potentials such as these are routinely implemented as cost functions throughout robotics and control theory. In Fig. 2.8, we depict the behavior of the system as a function of the temperature parameter. Initially, with the temperature set to zero the agent's paths are solely determined by the solution to the optimal control problem, smoothly driving towards the potential's minimum at the origin. Then, as we tune up α, we increase diffusivity of our agent's sample paths. While at $\alpha = 1$ the position of the system fluctuates very slightly at the bottom of the quadratic potential, at $\alpha = 100$ the agent diffuses around violently by overcoming its energetic tendency to stay at the bottom of the well. If we were to continue increasing α to larger and larger values, we would observe that directed maximally diffusive exploration would cease to be ergodic, as predicted by [39]. This occurs as a result of the strength of diffusive fluctuations (here set by our α parameter) dominating the magnitude of the drift induced by the potential's gradient. This is to say that for a given problem, system, and operator preferences, there should be a range of α values that best achieve the task.

Throughout this chapter, we have sought to lay down the foundations of *robot thermodynamics* as a general decision-making framework grounded in the statistical mechanics of complex embodied systems. As we discussed, robot thermodynamics is concerned with two questions: "What is the structure of an optimal agent's dynamics?" and "How can such dynamics be realized?" Across the many sections of this chapter, we formalized our approach to answering these questions. In short, we developed inference procedures for

answering the former of these questions, as well as control and policy synthesis procedures to answer the latter. The result is a flexible set of tools for inferring and specifying goal-directed behavior. In the following chapters, we will explore applications of the mathematical foundations we have laid down in a broad range of applications such as nonequilibrium statistical mechanics (in Ch. 3), microsystem design (in Ch. 4), and robot learning (in Ch. 5). These applications will highlight the capabilities of our framework in modelling, designing, controlling, and learning with embodied autonomous systems.

CHAPTER 3

Predicting Self-Organization in Active and Robotic Matter

In Ch. 2.3, we asked ourselves: *If all we knew about a complex system is that they explore their configurations in continuous paths, then what would be our best guess as to their dynamics?* Our exploration of this question led to a principle for predicting the steady-state occupancy statistics of a class of "sufficiently messy" stochastic processes in Ch. 2.4. In this chapter, we will investigate this principle from the perspective of nonequilibrium statistical mechanics. In other words, as a physical mechanism and explanation for many *far-from-equilibrium self-organization* phenomena in nature. Most of the work presented in this chapter was previously published in [6]. We note that the contributions of this thesis to that work include (but are not limited to): Developing the experimental platform, performing experimental validations and data analyses, and deriving control techniques to manipulate and engineer nonequilibrium steady-states.

Self-organization is frequently observed in active nonequilibrium collectives, from ant rafts to molecular motor assemblies and beyond. However, general principles describing self-organization in far-from-equilibrium settings have been challenging to identify. Here, we will offer a unifying perspective that views the behavior of complex systems as largely random—except for their configuration-dependent responses to external perturbations. Taking this perspective enables the derivation of a nonequilibrium Boltzmann-like principle (as in Ch. 2.4), which allows one to understand, predict, and manipulate nonequilibrium self-organization in a broad class of complex systems. Throughout this chapter, we validate

our predictions experimentally in a shape-changing robot collective capable of emulating the diverse properties of active matter systems across engineered and natural settings. Additionally, we outline multiple methodologies for steering and controlling nonequilibrium collective behavior based on these principles. Our findings highlight how emergent order depends sensitively on the matching between external patterns of forcing and internal dynamical response properties, pointing towards future approaches for design and control of collectives of autonomous agents and active particle collectives, as we will go on to show in Ch. 4.

3.1. Introduction

Self-organization in nature is surprising because getting a large group of separate particles to act in an organized way is often difficult. By definition, arrangements of matter we call "orderly" are special, making up a tiny minority of all allowed configurations. For example, we find each unique, symmetrical shape of a snowflake visually striking, in contrast with any randomly-rearranged clump of the same water molecules. Thus, any theory of emergent order in many-particle collectives must explain how a small subset of configurations are spontaneously selected among the vast set of disorganized arrangements.

Spontaneous many-body order is well-understood in thermal equilibrium cases such as crystalline solids or DNA origami [50], where the assembling matter is allowed to sit unperturbed for a long time at constant temperature T. The statistical mechanical approach proceeds by approximating the complex deterministic dynamics of the particles with a probabilistic "molecular chaos," positing that the law of conservation of energy governs otherwise random behavior [51]. What follows is the Boltzmann distribution

for the steady-state probabilities, $p_{ss}(q) \propto \exp[-E(q)/T]$, which shows that the degree to which special configurations q of low energy $E(q)$ have a high probability $p_{ss}(q)$ in the long-term depends on the amplitude of the thermal noise. Orderly configurations can assemble and remain stable, so long as inter-particle attractions are strong enough to overcome the randomizing effects of thermal fluctuations. For the remainder of this chapter, we use q instead of x to denote the state of the underlying physical stochastic process to align with field norms.

However, there are also many examples of emergent order outside of thermal equilibrium. From "random organization" in sheared colloids [52], to phase separation in multi-temperature particle mixtures [53], and dynamic vortices in protein filaments [54], a variety of ordered behaviors arise far from equilibrium that cannot be explained in terms of simple inter-particle attraction or energy gradients [55–58].

In all of these examples, the energy flux from external sources allows different system configurations to experience fluctuations of different magnitude [59, 60]. We suggest that the emergence of such configuration-dependent fluctuations, *which cannot happen in equilibrium*, may be key to understanding many nonequilibrium self-organization phenomena. In particular, we introduce a measure of driving-induced random fluctuations, which we term "rattling" $\mathcal{R}(q)$, and argue that it could play a similar role in many far-from-equilibrium systems as energy does in equilibrium. For a derivation of this quantity in the context of stochastic processes, we refer readers to Ch. 2.4.

We test our claim in a number of systems, including a flexible active matter system of simple robots we call "smarticles" (smart active particles) [8] as a convenient

test-platform (see movie S1[1]) inspired by similar robo-physical emulators of collective behavior [61–63]. Despite their purely repulsive inter-robot interactions, we find that smarticles spontaneously self-organize into collective "dances," whose shape and motions are matched to the temporal pattern of external driving forces (see movies S2[2] and S3[3]). This platform and others [64–66], including the nonequilibrium ordering examples mentioned above, all exhibit *low-rattling* ordered behaviors that echo low-energy structures emergent at equilibrium. We thus motivate and test a predictive theory based on rattling that may explain a broad class of nonequilibrium ordering phenomena.

3.2. Results

3.2.1. Rattling Theory

In devising our approach, we take inspiration from the phenomenon of thermophoresis, which is the simplest example of purely nonequilibrium self-organization, and is characterized by the diffusion of colloidal particles from hot regions to cold regions [67]. If non-interacting particles in a viscous fluid are subject to a temperature $T(q)$ that varies over position q, their resulting density in the steady-state $p_{ss}(q)$ will concentrate in the regions of low temperature. Particles diffuse to regions where thermal noise is weaker and become trapped there. With the diffusivity landscape set by thermal noise locally according to the fluctuation-dissipation relation $D(q) \propto T(q)$ [68], the steady-state diffusion equation $\nabla^2 \left(D(q)p_{ss}(q)\right) = 0$ is satisfied by the probability density $p_{ss}(q) \propto 1/D(q)$. Hence, a low-entropy, "ordered" arrangement of particles can be stable when the diffusivity

[1] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s1.mp4
[2] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s2.mp4
[3] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s3.mp4

landscape has a few locations q that are strongly selected by their extremely low $D(q)$ values.

We seek to extend this intuition to explain nonequilibrium self-organization more broadly. However, a straightforward mathematical extension of the idea encounters challenges in only slightly more complicated scenarios. For an arbitrary diffusion tensor landscape $\mathbf{D}(q)$, in which diffusivity can depend on the direction of motion, one can no longer find general solutions for the steady-state. Moreover, the steady-state density $p_{ss}(q)$ at configuration q may depend on the diffusivity $\mathbf{D}(\tilde{q})$ at arbitrarily distant configurations $\tilde{q}$. Nonetheless, we suggest that for most typical diffusion landscapes, the local magnitude of fluctuations $|\mathbf{D}(q)|$ should statistically bias $p_{ss}(q)$, and hence be approximately predictive of it. This insight, which is central to our theory, is illustrated to hold numerically in Fig. 3.1(A) for a randomly constructed two-dimensional anisotropic landscape.

The key assumption underlying our approach is that the complex system dynamics are so messy that only the amplitude of local drive-induced fluctuations governs the otherwise random behavior—an assumption inspired by molecular chaos at equilibrium, and motivated mathematically in Ch. 2.4. We expect this to apply when the system dynamics are so complex, nonlinear, and high-dimensional that no global symmetry or constraint can be found for its simplification. Although one cannot predict a configuration's nonequilibrium steady-state probability from its local properties in the general case [69, 70], the feat becomes achievable in practice for "messy" systems. To illustrate how local drive-induced fluctuations may be predictive of steady-state occupancy in complex dynamical systems, here we consider a discrete dynamical system with random transition rates between a large number of states. In this context, it has been shown analytically that the

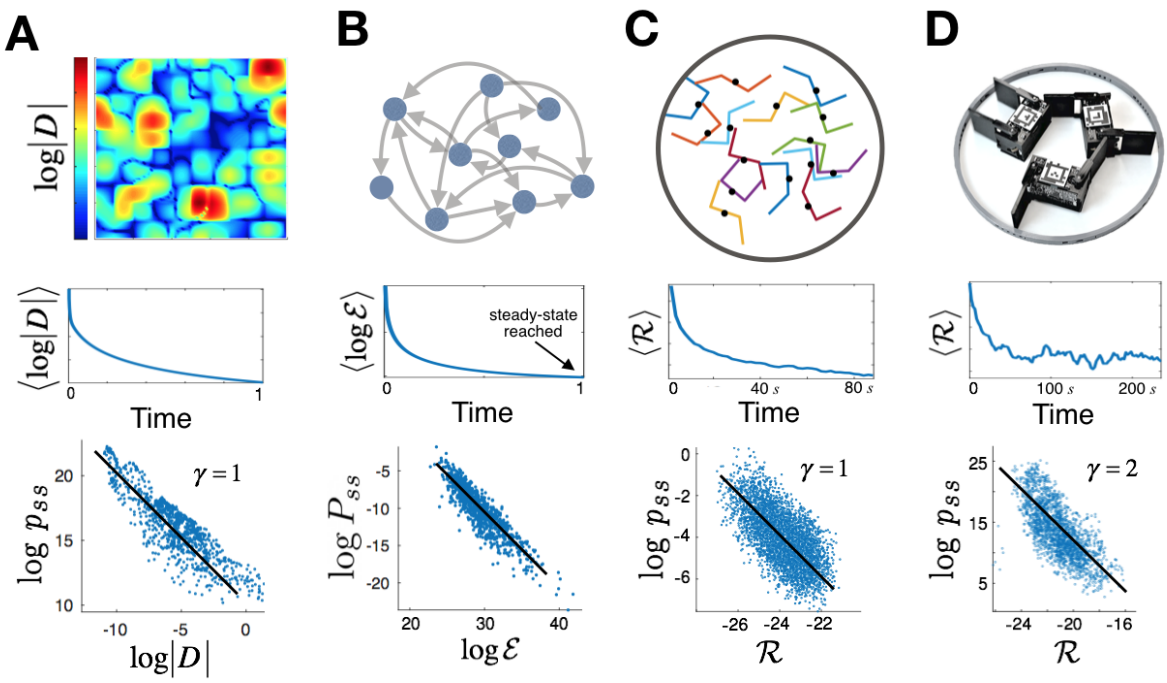


Figure 3.1. **Rattling $\mathcal{R}$ is predictive of steady-state occupancy across far-from-equilibrium systems.** (A) shows inhomogeneous anisotropic diffusion in 2D, where the steady-state density $p_{ss}(q)$ is seen to be approximately given by the magnitude of local fluctuations $\log|\mathbf{D}(q)| \propto \mathcal{R}(q)$ ($|\mathbf{D}|$—determinant of the diffusion tensor). (B) shows a random walk on a large random graph (1000 states), where P_{ss}—the probability at a state—is approximately given by $\mathcal{E}$—that state's exit rate. (C) shows an active matter system of shape-changing agents: an enclosed ensemble of 15 "smarticles" in simulation. (D) realizes similar agents experimentally with an enclosed three-robot smarticle ensemble. The middle row shows that relaxation to the steady-state of a uniform initial distribution is accompanied by monotonic decay in the average rattling value in all cases—analogous to free energy in equilibrium systems. The bottom row shows the validity of the nonequilibrium Boltzmann-like principle in Eq. 3.3, where the black lines in (A, B, and C) illustrate the theoretical correlation slope for a sufficiently large and complex system (see supplementary materials). The mesoscopic regime in (D) provides the most stringent test of rattling theory (where we observe deviations in γ from 1), while also exhibiting global self-organization. In (A and B, middle) time units are arbitrary, and for (C and D, middle) time is in seconds, where the drive period is 2 s.

net rate at which we exit any given state predicts its long-term probability approximately in most settings, even though the exact result requires global system knowledge in general(see Fig. 3.1(B)) [71, 72]. This result may be related to the above discussion of thermophoresis by noting that the discrete state exit rates are determined by the continuum diffusivity if our dynamics are built by discretizing the domain of a diffusion process.

To formulate our random dynamics assumption explicitly, we represent the complex system evolution as a trajectory in time $q(t)$, where the configuration vector q captures the properties of the entire many-particle system. Our messiness assumption amounts to approximating the full complex dynamics between two points $q(t)$ and $q(t+\delta t)$ by a random diffusion process. To this end, we take the amplitude of the noise fluctuations $D(q)$ to locally reflect the amplitude of the true configuration dynamics: $|q(t+\delta t)-q(t)|^2 \propto D(q)\delta t$ for short rollouts $q(t \to t+\delta t)$ (i.e., samples of system trajectories) of duration δt initialized in configuration $q(t) = q$. Through this approximation, our dynamics are effectively reduced to diffusion in q-space, which then allows us to locally estimate the steady-state probability of system configurations from $D(q)$ as in thermophoresis. Hence, the global steady-state distribution may be predicted from the properties of short-time, *local* system rollouts.

For rare orderly configurations to be strongly selected in a messy dynamical system, the landscape of local fluctuations must vary in magnitude over a large range of values. While in thermophoresis these fluctuations are directly imposed by an external temperature profile, in driven dynamical systems the effect results from the way a given pattern of driving can affect system configurations differently. The $D(q)$ landscape is emergent from the interplay between the pattern of driving, and the library of possible q-dependent system response properties. In practice, we observe that the amplitudes of system responses

to driving do often vary over several orders of magnitude (see Fig. 3.1). We see this phenomenology in many well-known examples of active matter self-organization [52, 60, 73]. For example, the crystals that form in suspensions of self-propelled colloids in [74] may be seen as the collective configurations that respond least diffusively to driving by precisely balancing the propulsive forces among individual particles. This illustrates how the low $D(q)$ configurations are selected in the steady-state by an exceptional matching of their response properties to the way the system is driven.

We apply these ideas in real complex driven systems whose response to driving we cannot predict analytically, such as our robotic swarm of smarticles. In this case, we require an estimator for the local value of $D(q^*)$ based on observations of short rollouts of system behavior when initialized at some configuration q^*. The estimator of the local diffusion tensor that we choose here is the covariance matrix [31]:

$$\mathbf{C}[q^*] = Cov[\tilde{v}_{q^*}, \tilde{v}_{q^*}] \tag{3.1}$$

where $\tilde{v}_{q^*}$ is seen as a random variable with samples drawn from $\{(q(t) - q(0))/\sqrt{t}\}_{q(0)=q^*}$ at various time-points t along one or several short system trajectories $q(t)$ rolled out from $q(0) = q^*$. We assume these rollouts $q(t)$ to be long enough to capture fluctuations in the configuration variables under the influence of a drive, but short enough to have $q(t)$ stay near q^*. We note that this is closely related to our discussion of the relative contributions of paths of different lengths to the steady-state occupancy statistics of a system in Ch. 2.4.

While the covariance matrix reflects the amplitude of local fluctuations, in estimating effective diffusivity we are instead interested in a measure of their disorder. This follows from the observation that high-amplitude ordered oscillations do not contribute to the rate

of stochastic diffusion [59]. We suggest that the degree of disorder of fluctuations may be captured by the entropy of the distribution of $\tilde{v}_{q^*}$ vectors, which is how we define "rattling" $\mathcal{R}(q^*)$. Physically, vectors $\tilde{v}_{q^*}$ capture the statistics of the force fluctuations experienced in configuration q^*, and so rattling measures the disorder in the system's driven response properties at that point. By approximating the distribution of $\tilde{v}_{q^*}$ as Gaussian, we can express its entropy (up to a constant offset) simply in terms of $\mathbf{C}[q^*]$ as:

$$\mathcal{R}(q^*) = \frac{1}{2}\log \det \mathbf{C}[q^*]. \tag{3.2}$$

Separately, we note that this definition is consistent with our derivation in Ch. 2.4.

With this definition, we generalize the thermophoretic expression for the steady-state density $p_{ss}(q^*) \propto 1/D(q^*)$ and express it in a Boltzmann-like form:

$$p_{ss}(q^*) \propto e^{-\gamma \mathcal{R}(q^*)}, \tag{3.3}$$

where γ is a system-specific constant of order 1. We note that when energy varies on the same scale as rattling, the interaction between the two landscapes can generate strong steady-state currents and may break this relation [59]. This way, using rattling we are able predict the long-term global steady-state distribution based on empirical measurements of short-term local system behavior, which suggests that probability density accumulates over time in low-rattling configurations.

3.2.2. Smarticle Experiments

We study the collective behavior of a simple ensemble of smarticles, aligning ourselves within the tradition of using robotic systems as flexible, physical emulators for self-organizing

natural systems [61–64]. Each smarticle (shown in Fig. 3.2(A)) is composed of three 5.2 cm long links, with two hinges actuated by motors programmed to follow a driving pattern specified by a micro-controller. When a smarticle sits on a flat surface, its arms do not touch the ground, so an individual robot cannot move. However, a group of them can achieve complex motion by pushing and pulling each other (see movie S1[4]) [75]. The relative coordinates of the middle link of each robot in the ensemble (x, y, θ) may be thought of as the internal system configurations that dynamically respond to an externally-determined driving force arising from the time-variation of arm angles (α_1, α_2).

This robotic active matter system offers substantial flexibility in choosing both the programmed patterns of driving, and the properties of internal system dynamics, such as friction coefficients, weights, etc. Additionally, the smarticle system has a flat potential energy landscape, allowing one to focus on the contributions of the drive-induced fluctuations to the collective behavior, making our findings broadly applicable to other strongly-driven systems. When the smarticles are within contact range (as ensured by a confining ring, Fig. 3.2(C)), the forces experienced throughout the collective for a given pattern of arm movement are an emergent function of all system coordinates. This configuration-dependent forcing gives rise to varying rattling values, which we refer to as the "rattling landscape," and which we see to be a hallmark property in many far-from-equilibrium examples. The rattling landscape then leads to some system configurations being dynamically selected over others and allowing for self-organization, just as the diffusivity landscape does in thermophoresis. Finally, the combined effects of impulsive inter-robot collisions, nonlinear boundary interactions, and static friction lead to a large

[4] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s1.mp4

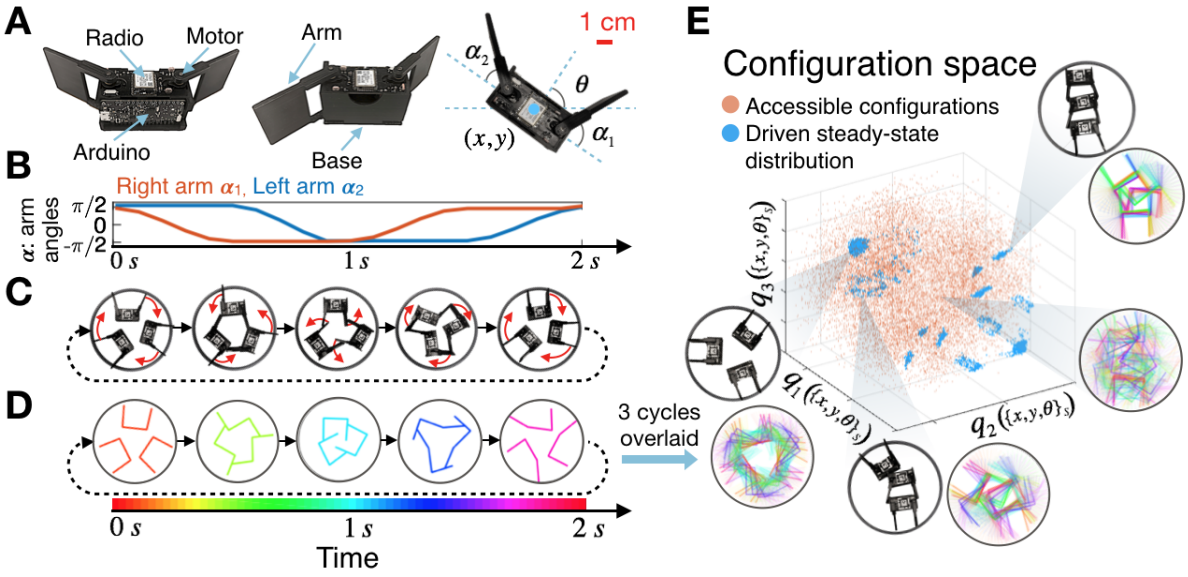


Figure 3.2. **Self-organization in a smarticle robotic ensemble.** (A) Front, back and top view of a single smarticle. Of its five degrees of freedom, we consider the time-varying arm angles (α_1, α_2) as "external" driving, since these are controlled by a pre-programmed microcontroller, while the robot coordinates (x, y, θ) are seen as "internal" system configuration, since these respond interdependently to the arms. (B) An example periodic arm motion pattern. (C) Top view of three smarticles confined in a fixed ring, all programmed to synchronously execute the driving pattern shown in (B). The video frames, aligned on the time-axis of (B), show one example of dynamically ordered collective "dance" that can spontaneously emerge under this drive. (D) Simulation video, showing agreement with experiment in (C). We color-code simulated states periodically in time, and overlay them for 3 periods to illustrate the dynamical order over time. (E) shows the system's configuration space, built from nonlinear functions of the three robots' body coordinates (x, y, θ). The steady-state distribution (blue) illustrates the few ordered configurations that are spontaneously selected by the driving out of all accessible system states (orange).

degree of quasi-random motion [75], making this a promising candidate system for exploring our theory.

Reasoning that our fundamental assumption of quasi-random configuration dynamics would be most valid in systems with many degrees of freedom, we also built a simulation

that would allow us to study the properties of larger smarticle groups and explore different system parameters (see Fig. 3.3). In this regime, we used simulations to gather enough data to sample the high-dimensional probability distributions for our analysis. In a simulation of 15 smarticles, we observed the tendency of the ensemble to reduce average rattling over time after a random initialization. For this 45-dimensional system (x, y, θ for 15 robots), the configuration-space dynamics are well-approximated by diffusion, and so Eq. 3.3 holds, which is also shown in Fig. 3.1(C). In addition, in these simulations we noted the emergence of metastable pockets of local order when groups of 3-4 nearby smarticles self-organized into regular motion patterns for several drive cycles (movie S2[5]).

The transient appearance of dynamical order in subsets of smarticle collectives raises the question of whether our rattling theory continues to hold for smaller ensembles. For the remainder of the paper we focus on ensembles of three smarticles (as in Fig. 3.1(D)), which allows for exhaustive sampling of configurations experimentally, and easier visualization of the configuration space (as in Fig. 3.2(E)). Both in simulation and experiment, we found that this regime exhibits a variety of low-rattling behaviors that manifest as distinct, orderly collective "dances" (movie S3[6] and Fig. 3.2, (C) and (D)). Despite its small size, this system is well-described by rattling theory, as evidenced by the empirical correlation between rattling and the steady-state likelihood of configurations (see Fig. 3.1(D), bottom).

We consider self-organization as a consequence of a system's landscape of rattling values over configuration space. This rattling landscape is specific to the particular drive forcing the system out of equilibrium, since different drives will generally produce different dynamical responses in the same system configuration. When the three-smarticle ensemble

[5] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s2.mp4
[6] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s3.mp4

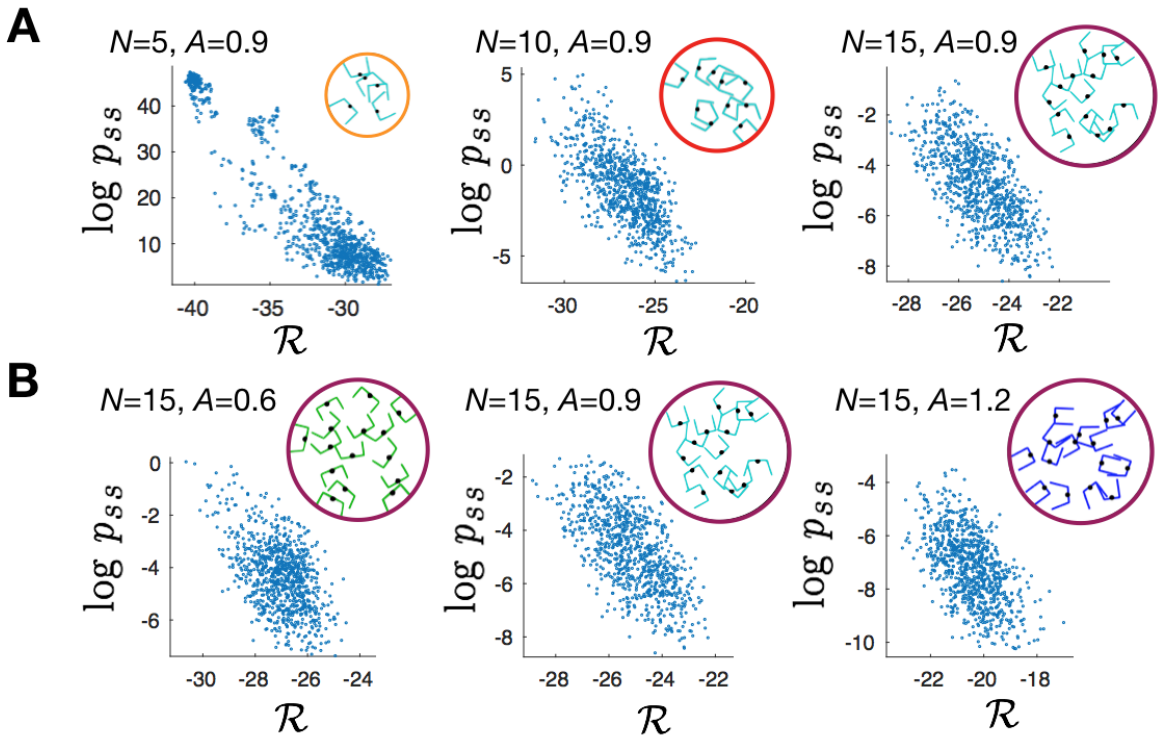


Figure 3.3. **Rattling prediction is robust across system parameters**. (A) illustrates that for larger numbers of smarticles N, the correlations between p_{ss} and $\mathcal{R}$ given by Eq. 3.3 persists in simulation. (B) similarly shows robustness to varying arm-lengths A, shown in relative units (where the middle link is of length 1).

is driven (under the pattern in Fig. 3.2(B)), the range of observed rattling values is so large that the lowest-rattling configurations—and consequently highest likelihood—account for most of the steady-state probability mass. Over 99% of probability accumulates in these spontaneously selected configurations, which represent only 0.1% of all accessible system states (Fig. 3.2(E)). Moreover, in these configurations the smarticles exhibit an orderly response to driving (see movie S4[7], and Fig. 3.2, (C) and (D)). In practice, the ensemble spends most of its time in or nearly in one of several distinct dances, with

[7] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s4.mp4

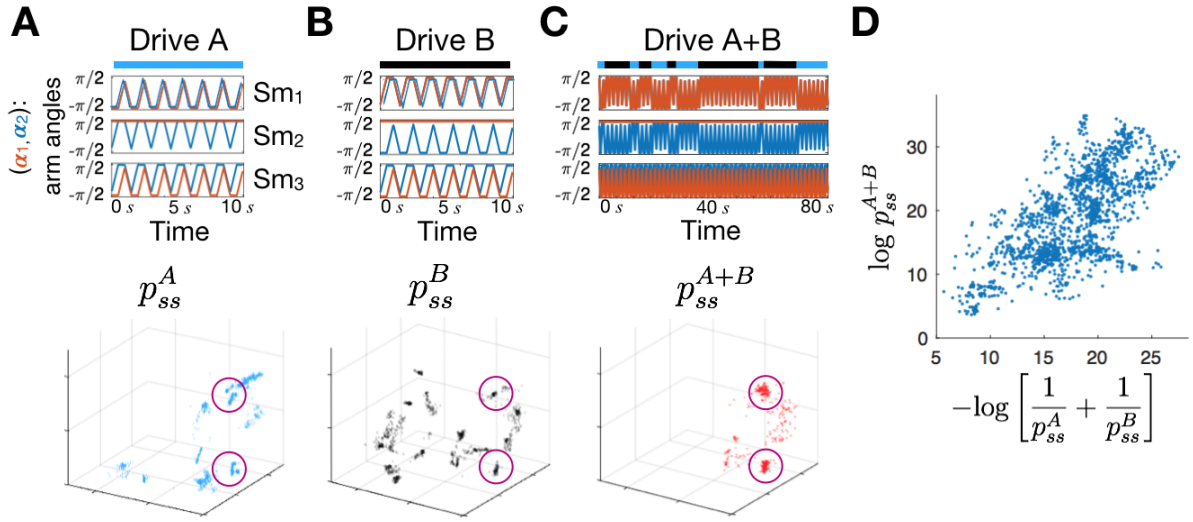


Figure 3.4. **Self-organized behaviors are fine-tuned to drive pattern**. (A) and (B) show that changing the arm motion pattern slightly (top) affects which configurations self-organize in the steady-state (bottom, same 3D configuration space as in Fig. 3.1(E)). (C) By mixing drives A and B as shown (top), we can isolate only those configurations selected in both the steady-states (circled in purple). This is an analytical prediction of the theory, and (D) further verifies its quantitative formulation.

occasional interruptions by stochastic flights from one such dynamical attractor to another (movie S5[8]).

From the above observations, we can begin to understand how self-organization emerges in driven collectives. In equilibrium, order arises when its entropic cost is outweighed by the available reduction of energy. Analogously, a sufficiently large reduction in rattling can lead to dynamical organization in a driven system. Moreover, such a reduction can require matching between the system dynamics and the drive pattern.

Through rattling theory we can predict how self-organized states are affected by changes in the features of the drive. We expect the structure of the self-organized dynamical

[8] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s5.mp4

attractors to be specific to the driving pattern, as each drive induces its own rattling landscape. To test this, we programmed the three smarticles with two distinct driving patterns (Fig. 3.4, (A) and (B), top), which we ran separately. The two resulting steady-state distributions, while each being highly localized to a few configurations, are largely non-overlapping (Fig. 3.4, (A) and (B), bottom). This indicates that by tuning the drive pattern, it may be possible to design the structure of the resulting steady-state, and hence control the self-organized dynamics (see also [76–78]).

As a proof of principle for such control, we developed a methodology for selecting particular steady-state behaviors by combining drives. By randomly switching back and forth between drives A and B in Fig. 3.4, we define a compound drive A+B (Fig. 3.4(C) and movie S6[9]). We predicted that this drive would select only those configurations common to both A and B steady-states (Fig. 3.4, (A) and (B), bottom), since having low rattling under this mixed drive requires having low rattling under both constituent drives. Our experiments confirmed this (Fig. 3.4(C)), and we were further able to quantitatively predict the probability that a configuration would appear under the mixed drive based on its likelihood in each constituent steady-state according to

$$1/p_{ss}^{A+B} \propto 1/p_{ss}^{A} + 1/p_{ss}^{B}, \tag{3.4}$$

as shown in Fig. 3.4(D). This simple relationship suggests that by composing different drives in time, one can single out desired configurations for the system steady-state, which provides an essential operational primitive on the road towards more complex control strategies for nonequilibrium collectives.

[9] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s6.mp4

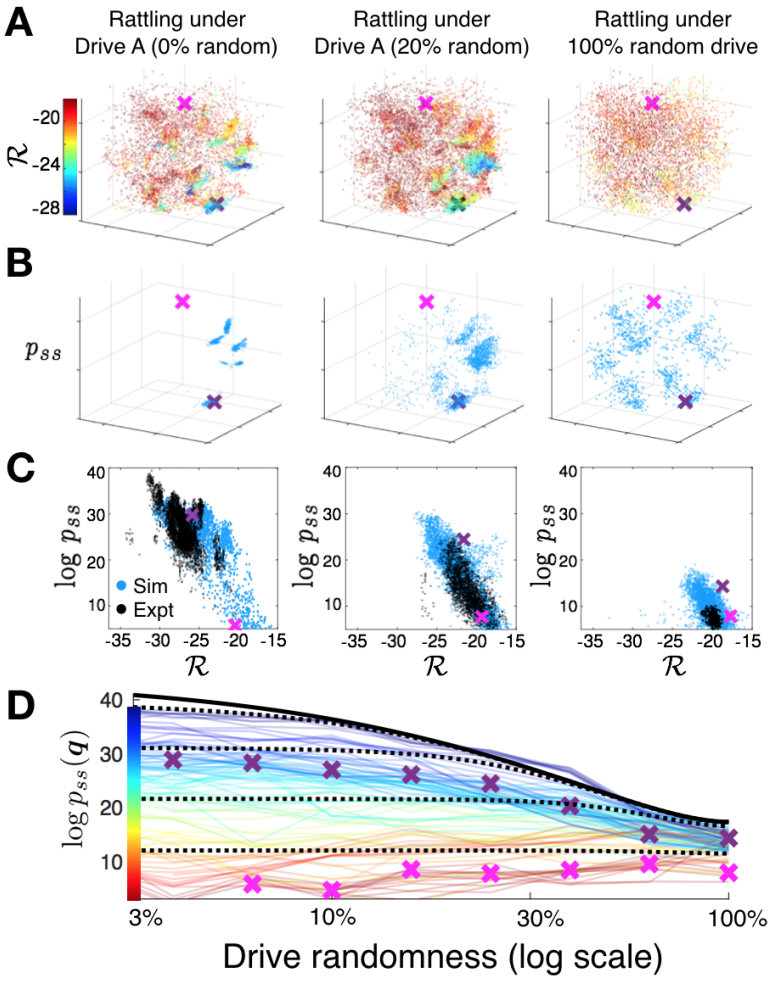


Figure 3.5. **Tuning self-organization by modulating drive randomness**. Self-organization relies on the degree of predictability in its driving forces, in a way that we can quantify and compute analytically. As the drive becomes less predictable (left to right, all panels), (A) low-rattling configurations gradually disappear. (B) The corresponding steady-states, reflecting the low-rattling regions of (A), become accordingly more diffuse (panels (A) and (B) show simulation data, and use the same 3D configuration space as Fig. 3.2(E)). (C) verifies that our central predictive relation Eq. 3.3 holds for all drives here, as all three correlations fall along the slope of the same line (blue: simulation, black: experiment). The diminishing range of rattling values thus precludes strong aggregation of probability, and with it self-organization. (D) shows our theoretical prediction (solid black line) indicating how the most likely configurations are destabilized by drive randomness. Colored lines track the probability p_{ss} at 100 representative configurations q in simulation, and dashed black lines analytically predict their trends. Two specific configurations marked by $\times$-s are tracked across analyses.

Moreover, we show that we can analytically predict and control the degree of order in the system by tuning drive randomness (see Fig. 3.5), as well as internal system friction (see movie S7[10] and Fig. 3.6). We note that the derivations of analytical curves in both of these figures can be found in the supplement of [6], but we omit them here as they are not essential to the subject matter of this thesis. As driven self-organization arises when the system has access to a broad range of rattling values, tuning it requires modulating the rattling of the most ordered behaviors relative to the background high-rattling states.

We can directly manipulate the rattling landscape by modulating the entropy of the drive pattern. This is done by introducing a probabilistic element to the programmed arm motion. At each move, we introduce a probability of making a random arm movement not included in the prescribed drive pattern. Increasing this probability results in flattening the rattling landscape: ordered states experience an increase in rattling due to drive entropy, while states whose rattling is already high do not (see Fig. 3.5(A)). Correspondingly, the steady-state distributions become progressively more diffuse (see Fig. 3.5(B)), causing localized pockets of order to give way to entropy and "melt" away—just as crystals might in equilibrium physics (movie S8[11], see also [79]).

Even as the range of accessible rattling values in the system shrinks, the predictive relation of Eq. 3.3 continues to hold (Fig. 3.5(C)), enabling quantitative prediction of how self-organized configurations are destabilized. By calculating the entropy of the drive pattern as we tune its randomness, we derive a lower-bound on rattling for the system. Thus, we can analytically predict how steady-state probabilities change as a function of drive randomness, as shown in Fig. 3.5(D) (up to normalization and γ). This result

[10] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s7.mp4
[11] https://www.science.org/doi/suppl/10.1126/science.abc6182/suppl_file/abc6182s8.mp4

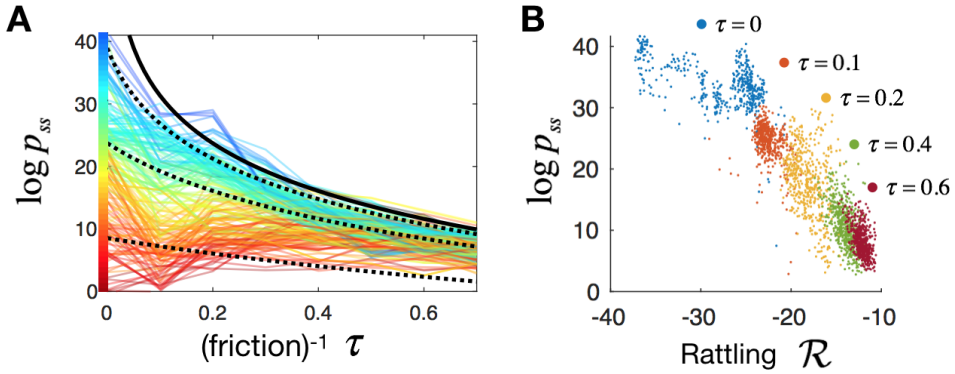


Figure 3.6. **Destroying self-organization by reducing friction**. In (A), we plot the steady-state probabilities at 200 different configurations under drive A (shown in Fig. 3.4) as we gradually reduce smarticle friction in simulation (τ is the velocity decay time-scale). Lines are colored according to state likelihood in the over-damped regime ($\tau \sim 0$). The solid black curve is the analytical prediction for the decay of low-rattling states, which also serves as an upper bound for probabilities of other configurations, which are predicted by dotted curves (with fitting parameter $\gamma = 3.1$). (B) illustrates the robustness of the relation between probability and rattling as friction in the system is changed. This, along with panel A, shows that measuring the overdamped dynamics $\tau = 0$ is sufficient to predict system behaviors for all lower friction values.

confirms the simple intuition that more predictably-patterned driving forces offer greater opportunity for the system to find low-rattling configurations, and self-organize.

3.3. Discussion

Our findings suggest that the complex dynamics of a driven collective of nonlinearly interacting particles may give rise to a situation in which a new kind of simplicity emerges. We have shown that when quasi-random transitions among configurations dominate the dynamics, the steady-state likelihood can be predicted from the entropy of local force fluctuations, which we refer to as rattling. In what we term a "low-rattling selection

principle," configurations are selected in the steady-state according to their rattling values under the given drive.

More significantly, low-rattling provides the basis for self-organized dynamical order that is specifically selected by the choice of driving pattern. We see analytically and experimentally that the degree of order in the steady-state distribution reflects the predictability of patterns in driving forces. Thus, driving patterns with low entropy pick out fine-tuned configurations and dynamical trajectories to stabilize. This makes it possible for one collective to exhibit different modes of ordered motion depending on the fingerprint of the external driving. These modes differ in their emergent collective properties, which suggests "top-down" alternatives to control of active matter and metamaterial design, where ensemble behaviors are dynamically self-selected by the choice of driving, rather than microscopically engineered [78, 80].

Throughout this chapter, we re-evaluated our mathematical results in Ch. 2.4 from the perspective of nonequilibrium statistical mechanics. In doing so, we discovered and experimentally validated a new principle of self-organization in nature. Moreover, by elucidating the relationship between important drive and system parameters we outlined strategies for control of nonequilibrium collectives. In the following chapter, we will extend these ideas and explore how to make use of rattling theory in the design of a complex system in hopes of harnessing self-organization towards microrobotic task-capabilities.

CHAPTER 4

Designing for Emergence in Robotic Microsystems

In Ch. 3, we developed a theory that explains the emergence of self-organization phenomena in broad classes of "sufficiently messy" dynamical processes. Rattling theory is built around a local scalar quantity—*rattling*, $\mathcal{R}$—that we have proven to be predictive of steady-state occupancy in Ch. 2.4. On the basis of these results, we also derived operational primitives for engineering and specifying the steady-state behaviors of systems far-from-equilibrium. We outlined how properties such as friction and drive entropy can be used as control parameters over the system's steady-state, which we used to steer the system into (and out of) highly fine-tuned and self-organized configurations.

In this chapter, we demonstrate how *design parameters* can be used to the same effect in a system of active colloidal microparticles. We explore how to induce self-organization in a complex system by carefully choosing its design parameters. Moreover, we show how such self-organization can be exploited towards rudimentary task-capabilities. If the previous chapter exemplifies the inference potential of robot thermodynamics, this chapter exemplifies its *synthesis* potential in a setting outside of control or policy optimization. From the perspective of robot thermodynamics, there is no difference between optimizing controller parameters or design parameters—they both present a means of reshaping an agent's path distribution towards desired outcomes. And, in this chapter, our desired outcome will be to induce orderly low-frequency oscillations in a microparticle collective. Much of the work in this chapter was previously published in [9], except for a few additional

results that are original contributions of this thesis. We note that the contributions of this thesis to that work include (but are not limited to): Theoretical modelling of the microparticle collective, development of computer vision algorithms, and processing of experimental data.

Spontaneous oscillations on the order of several hertz are the drivers of many crucial processes in nature. From bacterial swimming to mammal gaits, converting static energy inputs into slowly oscillating power is key to the autonomy of organisms across scales. However, the fabrication of slow micrometre-scale oscillators remains a major roadblock towards fully-autonomous microrobots. In this chapter, we will study a low-frequency oscillator that emerges from a collective of active microparticles at the air-liquid interface of a hydrogen peroxide droplet. Their interactions transduce ambient chemical energy into periodic mechanical motion and on-board electrical currents. Surprisingly, these oscillations persist at larger ensemble sizes *only when a particle with modified reactivity is added to intentionally break permutation symmetry*. We explain such emergent order through the discovery of a thermodynamic mechanism for asymmetry-induced order. The on-board power harvested from the stabilized oscillations enables the use of electronic components, which we will demonstrate by cyclically and synchronously driving a microrobotic arm. This work highlights a new strategy for achieving low-frequency oscillations at the microscale, paving the way for future microrobotic autonomy.

4.1. Introduction

The ability to produce low-frequency oscillations is central to the autonomy of living beings, and is essential to key biological processes such as heartbeats, neuron firings,

breathing, and locomotion [81–83]. While complex electronics operate at ever-increasing clock rates of many gigahertz, the frequency of many important biological oscillations seldom exceeds 100Hz. The slow rate of these oscillations stems from a need to be commensurate with both the energy budget and the natural timescales of underlying biological processes, as in the transport of CO_2 in plants [84] and in the galloping of horses [85]. Unlike oscillations arising from external periodic forcing [86–89], these self-oscillations emerge spontaneously from the balancing of competing dynamical processes driving systems away from equilibrium [90–92]—a signature of living systems [93].

In artificial microsystems, however, the production of slow self-sufficient self-oscillations is counterintuitively difficult [94, 95]. Generating self-sustaining mechanical oscillations at the microscale typically requires the transduction of complex chemical oscillators (e.g., Belousov-Zhabotinsky reaction [96]) into periodic changes to a system's physical configuration [88, 97–101]. Alternative mechanisms for producing self-sufficient mechanical oscillations based on carefully designed dynamic coupling between responsive elastic materials and thermal [92, 102], chemical [91, 92, 103], or moisture stimuli [104] have typically been demonstrated in millimetre-scale (and larger) devices. In contrast, generating slow periodic electrical signals remains prohibitively challenging aboard untethered microscale devices, given the limited downward scalability of capacitors and inductors [105, 106], as well as the power and footprint demands of CMOS oscillators, frequency dividers, and energy modules [107–109]. Despite these challenges, recent progress has shown that self-sustaining electrical oscillations can be produced by modulating electrical resistance with mechanical feedback loops in carefully designed devices, presenting a promising mechanism for sub-500μm electrical self-oscillators [94].

In this work, instead of relying on complex chemistries, integrated electronics, or elaborate mechanical microstructures, we produce robust electromechanical oscillations aboard a collective of deceptively simple microparticles by exploiting the self-organized properties of their far-from-equilibrium dynamics. By breaking the permutation symmetry of a homogeneous particle collective situated at an air-liquid interface, we reliably control their dynamics to realize simultaneous chemomechanical and electrochemical periodic energy transduction. In line with our robot thermodynamics framework, *we deliberately exploit their design parameters in order to manipulate the structure of their path distribution*, which leads to novel self-organized dynamical behaviors. We achieve this by introducing a particle with an enhanced reaction rate, whose stabilizing effect on the system behaviour we analyze through the lens of *asymmetry-induced order* and rattling theory. In turn, through a simple bimetallic on-board fuel cell design, we transduce the system's self-oscillations into periodic electrical work to power state-of-the-art microrobotic components, without the need for batteries or external sources of energy.

4.2. Results

4.2.1. Emergent Low-Frequency Oscillation

Figure 4.1 presents a system of simple microparticles where low-frequency chemomechanical self-oscillations emerge from the coupling of otherwise self-limiting catalytic reactions easily trapped at equilibrium. Figure 4.1(A) shows that each of these microparticles, composed of nothing more than a nanometre-thick Pt patch of radius 125μm microfabricated beneath a polymeric microdisc, generates a gas bubble when placed at the curved air-liquid interface

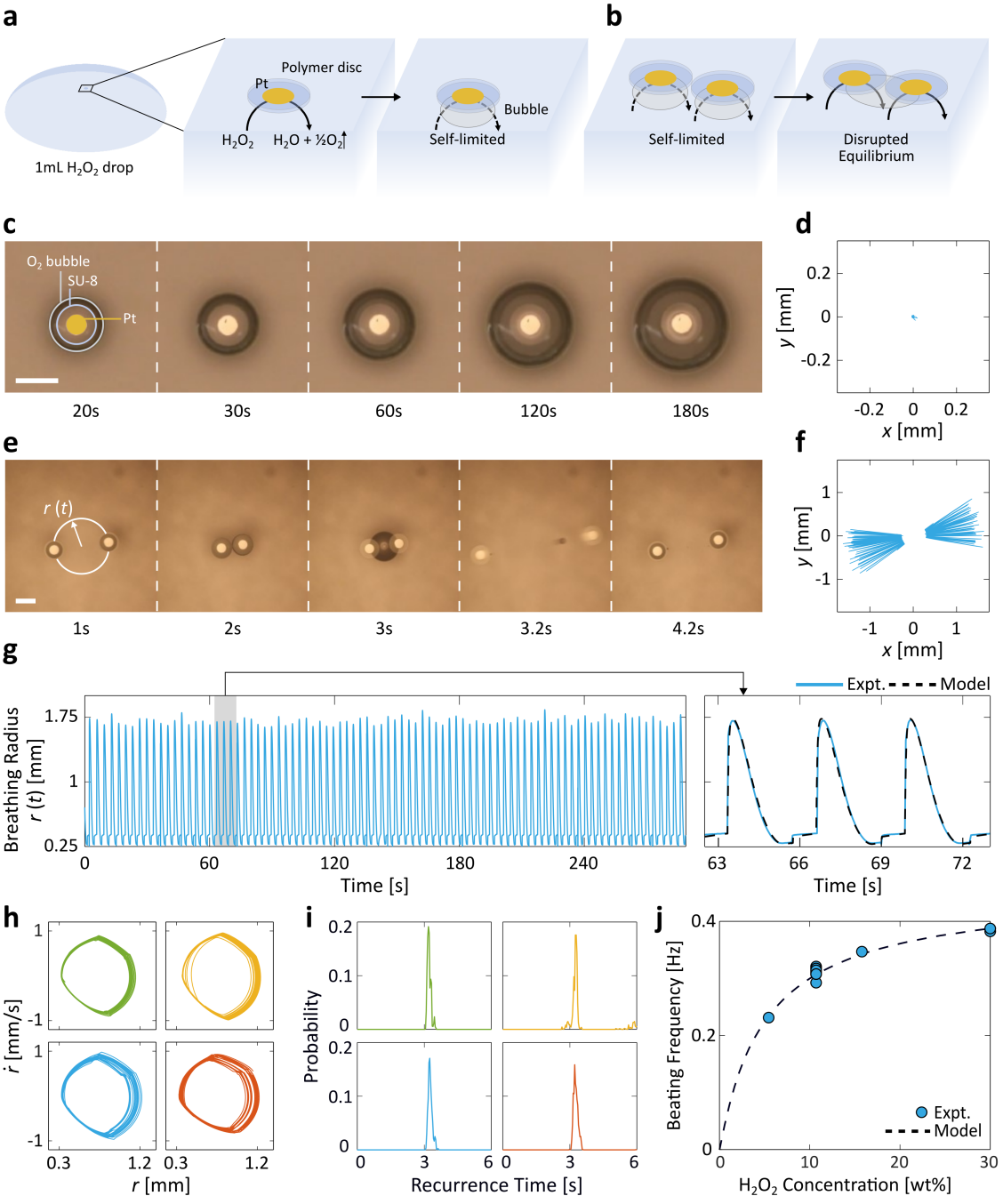

a
b
Polymer disc
Pt
Bubble
1mL H2O2 drop
H2O2
H2O + ½O2
Self-limited
Self-limited
Disrupted Equilibrium
c
O2 bubble
SU-8
Pt
20s
30s
60s
120s
180s
d
y [mm]
x [mm]
e
r (t)
1s
2s
3s
3.2s
4.2s
f
y [mm]
x [mm]
g
Breathing Radius r (t) [mm]
Time [s]
Expt.
Model
h
ṙ [mm/s]
r [mm]
i
Probability
Recurrence Time [s]
j
Beating Frequency [Hz]
H2O2 Concentration [wt%]
Expt.
Model

Figure 4.1. **Emergence of chemomechanical microparticle self-oscillation.** (A), Schematic of a self-limited system of a single particle resting still at the air-liquid interface of a H_2O_2 drop. The particle is composed of a catalytic patch of Pt (yellow) underneath a polymeric disc (blue). The O_2 formation slows down asymptotically over time as the gas bubble restricts the available catalytic surface area. (B), A 2-particle system, in contrast, exhibits an emergent and self-sustained beating behaviour as the bubble merger restores the previously hindered reactivity, thus disrupting the equilibrium state. (C),(D), Micrograph sequence (in (C)) and tracked particle coordinates (in (D)) of a 1-particle system that remains still for an extended period of time. (E),(F), Micrograph sequence (in (E)) and tracked coordinates (in (F)) of a 2-particle system with emergent beating. The breathing radius, $r(t)$, is the distance from the collective's centroid to each particle, averaged over all particles. (G), The long-term breathing radius trajectory of the same system as in (E) and (F) demonstrates the robustness of the beating behaviour. The shaded portion is magnified in the right panel, where mechanistic model simulations (black) are shown to match the experimental curve (blue). (H), The phase portraits of 4 independent 2-particle experiments demonstrate reproducible limit cycles with closed-loop orbits, confirming the periodicity of collective beating. Note that to calculate the phase portraits the system's bubble-driven discontinuities were processed through a standard finite-impulse response filter. All phase portraits share the same axes. (I), The recurrence histograms of the same 4 experiments all display a narrow peak centred at a period of 3.2s, consistent with visual evidence in (E). All histograms share the same axes. (J), The beating frequency can be tuned with the concentration of H_2O_2. The dependence predicted by the mechanistic simulations on the basis of a Langmuir-Hinshelwood kinetics (black curve) matches the experimental measurements (blue markers). Scale bars, 500μm.

of a H_2O_2 drop via

$$H_2O_2 \xrightarrow{\text{Pt}} H_2O + \frac{1}{2}O_2.$$

This well-studied catalytic reaction has been a long-time favourite in both micro- [110–113] and macroscopic robotics [92, 114], noted for the fuel's high energy density and simple chemistry [114].

For a single microparticle situated at the interface, the chemical reaction in Fig. 4.1(A) is self-limiting as the bubble grows and gradually blocks off the fuel's access to the catalyst. Consequently, the single-particle system reaches its equilibrium state promptly: The microparticle remains motionless for a prolonged time (Fig. 4.1(D), Movie S1[1]) and the bubble asymptotically reaches a terminal radius without rupture (Fig. 4.1(C)). However, a drastic change occurs when a second identical particle is introduced to the system. Figure 4.1(B) shows that as the microparticles enter each other's proximity, the separately-formed gas bubbles merge. The freed-up catalytic surface area then disrupts the self-limiting chemistry, destabilizing the original single-particle steady-state. This allows the merged bubble to grow beyond its threshold, leading to its rupture (Fig. 4.1(E), $t = 3.2$s). The collapse imparts an impulse onto the microparticles and propels them in opposite directions, at which point the particles are drawn back towards one another by the restorational forces: First, the radial component of buoyancy, $\mathbf{F}_{\mathrm{g}}$, globally directs the particles towards the apex of the concave air-liquid interface [89]. Second, the local interfacial deformations result in a mutual attractive capillary force $\mathbf{F}_{\mathrm{c}}$, affectionately known as the "Cheerios effect" [115, 116]. The combination of this Cheerios effect and catalytic bubble generation has been observed to produce repetitive back-and-forth motion [117, 118] in swarms of tubular swimmers [119, 120]. All of these factors sum up to a repeatable cycle of mutual approach, contact, bubble merger, and bubble collapse that we refer to as particle beating (Fig. 4.1(E)). The robustness of this self-sustained cycle is evidenced by the tracked coordinates of the two particles over a course of 280s (Fig. 4.1(F) and Movie

[1] `https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM4_ESM.mp4`

S2[2]), which contrast the single particle scenario where practically no motion was observed. Notably, while the central challenge in self-oscillatory systems is to keep them away from equilibria [91, 95], such states are virtually eliminated from our system by the effectively instantaneous nature of bubble collapse.

We monitored the oscillatory behaviour of the system by tracking its breathing radius $r(t)$ over time, defined as:

$$r(t) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i(t)-\bar{x})^2 + (y_i(t)-\bar{y})^2} \tag{4.1}$$

for a collection of N particles each with coordinate $(x_i(t), y_i(t))$ at time t. In other words, $r(t)$ is the Euclidean distance from the collective's centroid $(\bar{x}, \bar{y})$ to each particle, averaged over all particles (see annotations in Fig. 4.1(E)). The system's periodic beating is evident in the time evolution of $r(t)$ (Fig. 4.1(G), left panel), the limit cycle of its $r(t)$ phase portrait (Fig. 4.1(H)), as well as the narrow peak in the recurrence time histogram (Fig. 4.1(I)). The phase portraits were then constructed by plotting the coordinates of $v(t) = [\dot{r}(t), r(t)]$ after applying a low-pass filter, and the time-derivative of the breathing radius was estimated via finite differencing. Equipped with these dynamical observables, we analyzed the recurrence properties of the system by finding how often and how quickly the system returns to a neighborhood of $v(t)$. Hence, for a given experiment comprised of K samples we collect data at times $t_i = i\Delta t,\ \forall i \in \{0, \cdots, K-1\}$ with sampling rate Δt. While in principle this is all one needs in order to quantify recurrence statistics [121], an additional step must be taken in order make the calculation robust. We augmented our $v(t_i)$ vectors by "embedding" the time-series according to an integer parameter m [122].

[2] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM5_ESM.mp4

This resulted in a modified set of coordinates, $v_m(t_i) = [v(t_i), \cdots, v(t_{i+m-1})]^T$, from which to robustly calculate our recurrence statistics. Finally, to derive the recurrence properties of a system from an experimental dataset we calculated its recurrence set

$$R_s = \{|t_i - t_j| : \ ||v_m(t_i) - v_m(t_j)|| < \epsilon, \forall i, j\}, \tag{4.2}$$

over all valid indices. Note that m and ϵ are a fixed choice of positive non-zero embedding dimension and neighborhood size parameters, respectively. We may then calculate recurrence time histograms from the set R_s using any standard scientific computing package. Taken together, the results drawn from these analyses serve as conclusive evidence of the long-term stability of system oscillations. The analysis in Fig. 4.1(I) shows a period of 3.2s for the two-particle system in 10.7wt% H_2O_2, consistent with Fig. 4.1(G) and Movie S2[3].

Lastly, we developed a mechanistic Newtonian model based on analytical derivations of the forces acting on the particles: $\mathbf{F}_\mathrm{g}$, $\mathbf{F}_\mathrm{c}$, and the non-Stokesian drag force $\mathbf{F}_\mathrm{d}$. The net effect on these forces on particle motion is given by

$$\frac{d\mathbf{v}_i}{dt} = \frac{1}{m_\mathrm{eff}}(\mathbf{F}_{\mathrm{g},i} + \mathbf{F}_{\mathrm{c},i} + \mathbf{F}_{\mathrm{d},i}), \tag{4.3}$$

where i is a given particle and m_eff is its effective mass, which is affected by its bubble and the displaced liquid during acceleration. For details on the analytical expressions for these forces, we refer readers to the supplement of [9], where these quantities are explicitly derived at length. Remarkably, we found that our mechanistic model was so accurate as to capture even the detailed dynamics of the breathing radius' time evolution

[3] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM5_ESM.mp4

(Fig. 4.1(G), right panel). We verified the consistency of the beating frequency across 8 sets of independent experiments with 10.7wt% H_2O_2 in Fig. 4.1(J). Additionally, the beating frequency's dependence on H_2O_2 concentrations points to a mechanism for exerting fine control over the beating frequency, as predicted by our mechanistic model based on a Langmuir-Hinshelwood kinetics of the catalytic surface (Fig. 4.1(J)) [123, 124].

4.2.2. Rattling as a Mechanism for Asymmetry-Induced Order

Our findings in Fig. 4.2 show that the stable emergent self-oscillation can be extended well beyond $N = 2$, although curiously *only when the system's permutation symmetry is broken* and not in a homogeneous system of identical particles. We extracted the bubble burst interarrival time statistics by tracking the time that transpires between each pair of consecutive bursts in recorded experiments (Fig. 4.2(A)). In homogeneous systems of identical particles (Fig. 4.2(B)), we show that the likelihood of periodic beating dwindles gradually with rising particle counts N, reflected in the progressive decay in the sharpness and amplitude of the initial 3.2s peak corresponding to periodic beating. The decay of collective periodicity is accompanied by an increase in the probability mass of frequent and unpredictable bubble bursts taking place less than a second from one another—a result of bubble mergers and collapses among subsets of particles (see representative $N = 5$ and 8 micrographs in Fig. 4.2(B)). Interestingly, we find that the interarrival time distributions of systems beyond $N = 7$ become statistically indistinguishable from those of a Poisson process (Fig. 4.2(B), bottom panel) [125]. This shows that our system's phenomenology can remarkably vary from coordinated and reliable periodic beating to independent and effectively stochastic bubble bursts merely as a function of N. The breathing radius

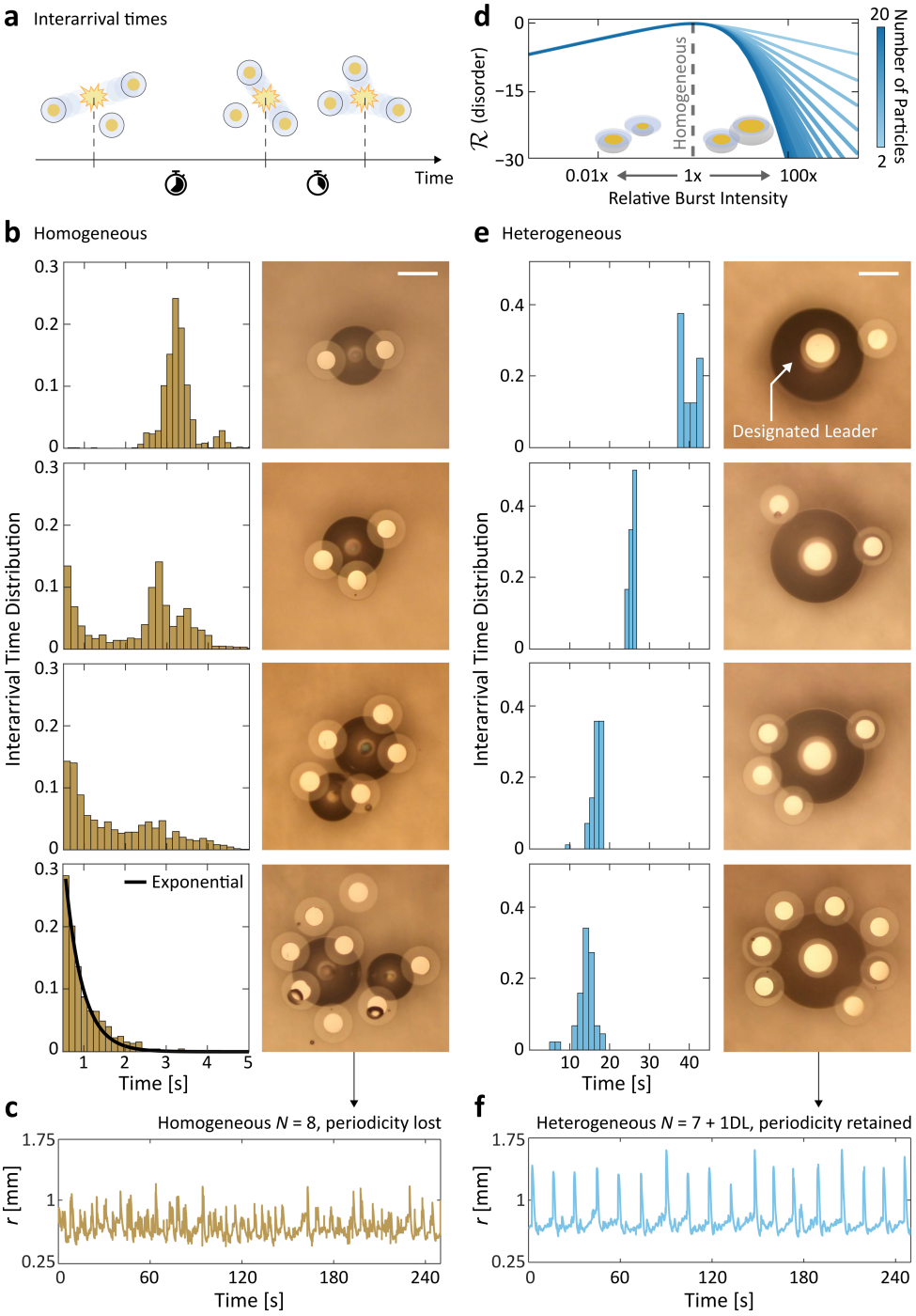

a
Interarrival times
Time
d
0
−15
−30
$\mathcal{R}$ (disorder)
Homogeneous
20
2
Number of Particles
0.01x
1x
100x
Relative Burst Intensity
b
Homogeneous
e
Heterogeneous
Interarrival Time Distribution
Exponential
Time [s]
Designated Leader
c
Homogeneous N = 8, periodicity lost
f
Heterogeneous N = 7 + 1DL, periodicity retained
r [mm]
Time [s]

Figure 4.2. **Observations of emergent order via symmetry-breaking.** (A), Schematic of interarrival times in a system of beating microparticles, defined as the time that transpires between two consecutive bubble collapses. The interarrival time distribution should be tight (i.e., a single peak) in a perfectly periodic system, and broad in an aperiodic system. (B), (top to bottom) Interarrival time distributions and optical micrographs for homogeneous systems of $N = 2, 3, 5$, and 8 identical particles. As N increases, the collective system periodicity gradually decays and transitions to an exponential interarrival distribution at $N = 8$ (bottom, black curve). Scale bar, 500μm. (C), Indeed, we observe that the breathing radius of a homogeneous $N = 8$ system is not periodic. (D), Asymmetry-induced order across N predicted by Rattling Theory. A quantification of collective disorder, the system's Rattling $\mathcal{R}$ is predicted to be lower (i.e. more orderly) if the relative burst intensity of one particle is increased beyond or decreased below 1x, which signifies homogeneity. This is experimentally realized by modulating the Pt patch size on a "designated leader" (DL) particle relative to the others. The curves are offset to make all $\mathcal{R} = 0$ at 1x intensity to highlight the effect of system heterogeneity on Rattling. (E), Same as (B), but for heterogeneous systems of equal particle numbers, where the DL broke the permutation symmetry. In contrast to the homogeneous systems (B), they remain robustly periodic across N. It is important to recognize that the polymeric disc size of a DL is unchanged. Scale bar, 500μm. (F), Breathing radius for an 8-particle DL system (i.e., $N = 7 + 1$DL), which reliably beats periodically. The period of 14.2s extracted from $r(t)$ coincides with the most probable interarrival time in ((E), bottom).

trajectory in Fig. 4.2(C) confirms the loss of periodicity, as no structure can be discerned from the noisy low-amplitude fluctuations.

The gradual transition towards aperiodicity in Figs. 4.2(B) and (C) points to the nominal fragility of periodic beating as the system size increases. Reasoning that the deliberate introduction of heterogeneity has been shown to produce asymmetry-induced order [126] in complex networked systems [127–129], we will investigate the effect of symmetry-breaking on the robustness of particle beating across system sizes. Asymmetry-induced order is a process by which explicit symmetry-breaking leads to the emergence

of ordered states in a system [126–129]. Hence, asymmetry-induced order requires both a symmetry whose breaking can be observed, and a clear notion of "degree of order." Which symmetry to break is inherently a system-dependent question, and as such there are no general means of choosing between symmetry groups to achieve a desired outcome. However, the degree of order of a system is a challenging property to specify in general. For one, what is meant by order is often ill-posed. Secondly, even when provided with a means to metricize order, metrics are often analytically and computationally intractable because they require global knowledge of system states—as is the case for calculating entropy. This is further complicated by the fact that, far from equilibrium, entropy is not sufficient to establish the robustness, stability, or persistence of system configurations (all of which are attributes often ascribed to "orderly" states) [70]. To this end, physicists have made use of order parameters to establish more narrowly-construed notions of order on a case-by-case basis for particular systems [130, 131].

Recent work in nonequilibrium statistical mechanics has made strides towards describing the emergence of order more generally in broader classes of complex systems. Rattling theory—a contribution of this thesis (in Ch. 3)—is a novel theory capable of describing the emergence of order and self-organization in "messy" nonequilibrium dynamical systems [6, 59]. The rattling ansatz sees the behavior of complex systems as stochastic diffusion processes taking place in high-dimensional configuration spaces in the presence of energy influxes driving them out of equilibrium. At the heart of the theory lies a local and computable notion of "degree of order"—*rattling*. As discussed in the previous chapter, rattling measures the way in which system configurations respond to external force fluctuations: Rapid, uncorrelated configurational changes produce high rattling values,

and slow, correlated changes produce low rattling values. Thus, in what follows we will use rattling as a general means of measuring and metricizing "degree of order" in a broad class of systems.

Equipped with a precise way to quantify order in a broad class of complex systems, we may now develop a system-specific understanding of the ways in which symmetry-breaking affects the rattling of our system of beating particles in hopes of finding strategies to stabilize periodic system beating for $N > 2$. In order to elucidate the role that symmetry-breaking may play in the self-organized states of our system of active microparticles, we must now consider specific system symmetries and their relationship to the magnitude of system-level fluctuations. While our system is not invariant to the action of any obvious continuous symmetry groups, it is *permutation-symmetric* [132]. This is to say that our collection of microparticles are all dynamically identical (up to fabrication tolerances). Hence, one promising avenue to investigate is the different ways in which permutation-symmetry breaking may lead to order in our system. Based on results from our mechanistic modelling of particle beating, we know that there are two ways in which the dynamics of individual microparticles can be made distinct from one another. First, we know that changing the volumetric shape of particles will lead to different local hydrodrynamic drag properties. Second, we know that changing the buoyancy of particles also produces local changes to individual microparticle dynamics through its effect on capillary forces. However, changing the shape of our microparticles requires major changes to their fabrication, as well as nontrivial modifications to the mechanistic model. In contrast, we can easily modify a particle's buoyancy by modulating the volume of the bubble forming underneath the particle, which we can in turn control through the size of their Pt patch.

To explore the role of permutation-symmetry breaking on our system, we constructed a simple model that we can work with analytically from the perspective of rattling theory. In line with the rattling ansatz, our model considers the configurational dynamics of collectives of beating particles as a diffusion process. We incorporate the effect of heterogenous particle buoyancies through the inclusion of a parameter modulating the size of bubbles in analogy to the role of the Pt patch. Our beating particles are perfectly suited for this sort of analysis, even more so than others, due to the physics of fluid dynamics at low-Reynolds numbers ($\sim$0.25 Re for our system) [133]. In this regime, inertia ceases to influence the behavior of systems, leaving viscous forces and stochastic thermal fluctuations to affect their dynamics substantially—thereby making a diffusive approximation natural.

Our model elucidates the role of design parameters on the structure of the system-level fluctuations on the basis of two primary assumptions. First, we assume that the behavior of each individual particle i is monotonically modulated by some real-valued design parameter U_i from a set $U = \{U_1, \cdots, U_N\}$ for a system of N particles. These design parameters correspond by analogy to the Pt patch size. Second, we assumed that particle i's bubble burst only affects the other members of the collective and not itself, which broadly matches experimental observations. We can think of the U_i parameters as implicitly determining the strength of the impulse imparted by particle i's bubble burst onto its neighbors. In particular, we model the effect of this parameter and the bubble burst strength a_i according to the following Boltzmann-like monotonic relationship,

$$a_i = \frac{1}{Z} e^{-U_i} \tag{4.4}$$

where Z is a normalization factor given by $Z = \sum_{i=1}^{N} e^{-U_i}$. In other words, the a_i parameters can be thought of in analogy to the size (and strength) of bubbles that a given particle can support. Hence, we can motivate this modeling choice by envisioning the gas in bubbles distributing itself according to an energy landscape specified by our U_i parameters, and thusly influencing the bubble popping strength a_i. The normalization factor Z arises from the fact that we are not interested in the absolute magnitude of the bubble bursts but rather the effect of their relative magnitudes on the collective behavior.

Now, let us consider the statistical properties of the dynamics of a breathing-radius-like observable, $\bar{r}(t)$, under a simple diffusive model. As before, $\bar{r}(t)$ is an averaged quantity over particles: $\bar{r}(t) = \frac{1}{N}\sum_{i=1}^{N} r_i(t)$. By assumption, a bubble burst at particle i leaves particle i stationary, but a burst from some neighbor j exerts an impulse of random direction onto particle i. In this case, the dynamics of $r_i(t)$ evolve according to

$$\dot{r}_i(t) = \sum_{j \neq i} a_j \cdot \xi_j \tag{4.5}$$

where ξ_j is normally-distributed delta-correlated multiplicative noise in the Itô convention. Note that this construction results in an anisotropic diffusion tensor without spatial dependence, as we are not modelling the geometry of interparticle interactions but rather the structure of their parameter-induced statistical fluctuations. From this specification of the system's diffusive dynamics, we can apply rattling theory to understand the effect of our design parameters U_i on the self-organized collective behavior of the system.

Given this formulation of the system dynamics, we proceed by calculating the effect of parameter changes on the magnitude of system-level fluctuations. Letting $r(t) = [r_1(t), \cdots, r_N(t)]^T$, the correlation structure of the system can be computed analytically

without empirical covariance matrices:

$$\langle \dot{r}_i(t), \dot{r}_j(t) \rangle = \sum_{k \neq i} \sum_{l \neq j} a_k a_l \delta_{kl} = \sum_{k \neq i,j} a_k^2 = \frac{1}{Z^2} \sum_{k \neq i,j} e^{-2U_k} \tag{4.6}$$

where δ_{kl} is the Kronecker delta. We note that the correlation structure of the system has no dependence on time (i.e., it has infinite temporal correlation) and no dependence on configuration $r(t)$, leaving the design parameters U_i as the only variables with an effect on the system behavior. Finally, to express rattling in terms of the system's design parameters we require an analytical expression for the determinant of its covariance tensor, which is challenging in general. Fortunately, for this particular correlation structure there exists such a closed-form expression, which enables us determine the system's rattling as a function of its design parameters:

$$\mathcal{R}(U) = \frac{1}{2} \log \det \langle \dot{r}_i(t), \dot{r}_j(t) \rangle = \log \left(\frac{(N-1) \prod_{i=1}^{N} e^{-U_i}}{Z^N} \right). \tag{4.7}$$

Equipped with an understanding of how the system's design parameters affect its rattling, we can now use the model as a tool to guide our experimental design. We will take two different approaches to system parameter design: First, we will explore the introduction of a "designated leader" particle, and then we will explore an "imprecision engineering" approach wherein poor manufacturing tolerances lead to more robust system behaviors.

While there are infinitely many parameter combinations for a given collection of N particles, one of the simplest design alterations to study is the effect of a single particle differing from the rest—for reasons that we will see shortly, we term this particle a *designated leader*. In this setting, one particle will have its parameter be U_{DL} while the

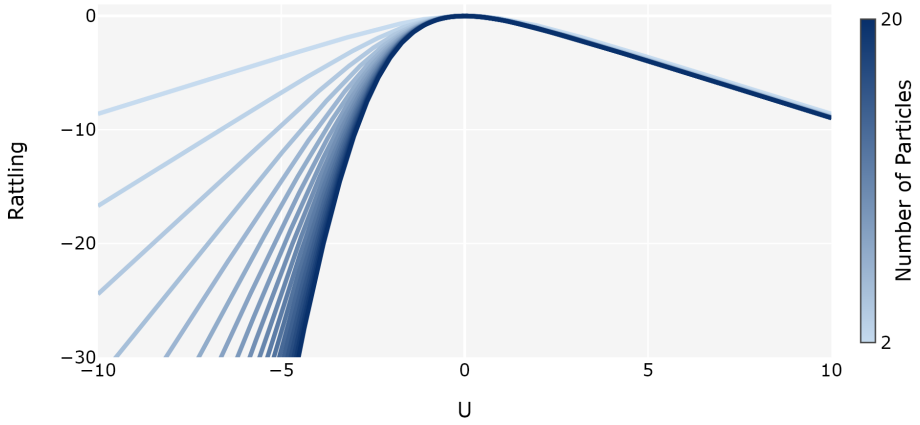


Figure 4.3. **Rattling as a function of patch size in diffusive model.** Here, we study the effect of a given particle's U parameter (in analogy to Pt patch size) on the rattling of collectives of varying sizes. Note that we subtract the constant offset in rattling due to system size so that $\mathcal{R} = 0$ at $U = 0$ for all N. We find that any variability in the size of the particle's patch produces a drop in rattling, leading to asymmetry-induced order. When a particle becomes inert as U increases, it stops contributing to system-level fluctuations, leading to a modest drop in rattling independent of N. However, as U decreases the modified particle's bubble bursts dominate and effectively become the sole source of variance in the system's configurational degrees of freedom. Such coordination among degrees of freedom leads to a sharp drop in rattling dependent on N.

rest of the $N-1$ particles will have it be $\bar{U}$ (which we take to be a constant fixed a priori). Rearranging the expression in Eq. 4.7, we have the following expression

$$\mathcal{R}(U_{DL}, N) = -U_{DL} + \log\left(\frac{(N-1)e^{-(N-1)\bar{U}}}{\left(e^{-U_{DL}} + (N-1)e^{-\bar{U}}\right)^N}\right),$$

which allows us to make predictions about the behavior of a collection of N beating particles with a single designated leader. However, much in the same way that entropy can trivially depend on system size (e.g., number of microstates), Eq. 4.2.2 does as well. Thus, to focus on the dependence of $\mathcal{R}(U_{DL}, N)$ on U_{DL}, we subtract the constant bias that system size contributes to the value of rattling. To do this, we calculate $\mathcal{R}(U_{DL}, N) - \mathcal{R}(\bar{U}, N)$

for a choice of $\bar{U}$ that we fix across all system sizes, where we note that $\mathcal{R}(\bar{U}, N)$ is merely a constant that offsets the value of rattling to be zero when $U_{DL} = \bar{U}$. Since $\mathcal{R}(\bar{U}, N)$ is exclusively a function of the number of particles for a given $\bar{U}$, subtracting it from $\mathcal{R}(U_{DL}, N)$ precisely removes the constant contribution of system size to the overall magnitude of rattling. As detailed in [6], constant offsets to the rattling values of a system do not affect its behavior. Only changes to the rattling landscape—that is, changes to the relative rattling values between configurations (or parameters)—have an effect on system behavior, which motivates our approach.

In Fig. 4.3 we show the results of varying the parameters of the designated leader for collectives of various sizes, while fixing $\bar{U} = 0$ and subtracting the bias in rattling due to system size. Crucially, we observe that any deviation from the parameter values of the rest of members of the collective (i.e., away from $U_{DL} = 0$) results in a reduction in rattling. Thus, our model predicts that any amount of heterogeneity will lead to increasingly ordered system states. Such asymmetry-induced order has been studied in networked systems of oscillators [126–129], but its emergence as a low-rattling phenomenon is a novel finding. We note that our results in Fig. 4.2(D) are also based on Eq. 4.2.2, except for the fact that the x-axis is rescaled to be in line with the more intuitive notion of "relative burst intensity."

Through this mechanism, order arises in one of two distinct ways. First, as U_{DL} increases, the designated leader particle becomes effectively inert. This is to say that the strength of its bubble bursts a_{DL} asymptotically approach zero, as though it were a patchless particle. As a result, the leader particle acts as dead weight and does not contribute to system-level fluctuations, leading to a modest decrease in rattling—independent of the total

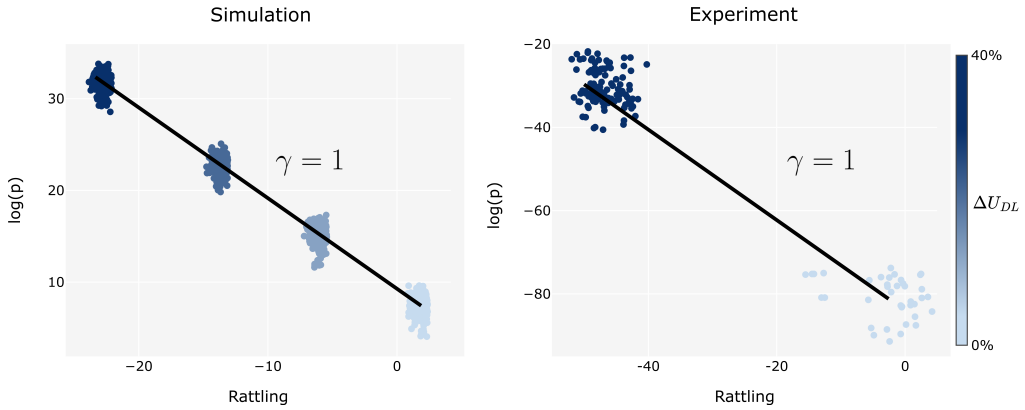


Figure 4.4. **Effect of designated leader on self-organization.** On the left panel, we simulate the dynamics in Eq. 4.5 and calculate their rattling and steady-state densities numerically. On the right panel, we consider experimental data from an 8 particle collective in both standard ($\Delta U_{DL} = 0\%$) and designated leader configurations ($\Delta U_{DL} = 40\%$), which we then process using the same procedure as for the left panel. While the absolute magnitudes of parameter values for the simulation are arbitrary, the ΔU_{DL} values are determined from the actual Pt patch sizes used on the experimental systems. For both the simulated and the experimental data, the results are consistent with rattling theory with $\gamma = 1$).

number of particles—that matches experimental observations. Second, as U_{DL} decreases, the designated leader particle's bubble bursts become stronger and its contribution to the magnitude of system-level fluctuations dominates over those of other particles. In turn, this effectively leads to a concentration of all variability and randomness in the system into a single of its many degrees of freedom, thereby leading to significant correlations in the behavior of all particles and a resulting drop in rattling. Note that as more particles are added more degrees of freedom become correlated, leading to sharper drops in rattling as a function of N. Remarkably, these outcomes fall in line with the predictions of rattling theory in both simulations and experiments (see Fig. 4.4). More precisely,

$$p(q) \propto e^{-\gamma(\mathcal{R}(q)+S(q))}, \tag{4.8}$$

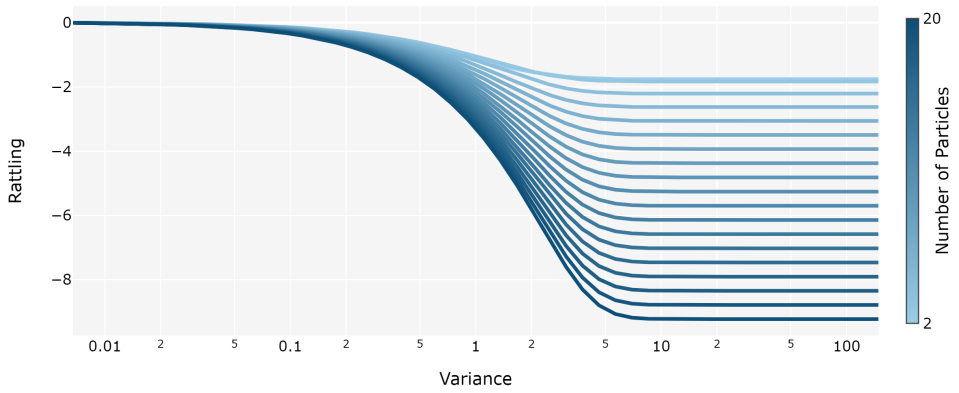


Figure 4.5. **Rattling as a function of patch variance in diffusive model.** Here, we study the effect of randomly assigning according to a log-normal distribution. Using the Fenton-Wilkinson approximation [134], we are able to derive an analytical expression for rattling as a function of the mean and variance of the U_i parameters in Eq. 4.9. For a fixed choice of mean, this figure depicts how variance in the distribution of U_i affects rattling across ensembles of different sizes.

where decreases of U_{DL} increase effective drive entropy $S(q)$, thereby lowering $p(q)$. We note that this result from [6] was originally derived for Fig. 3.5. Hence, on the basis of these results and other studies of asymmetry-induced order we chose to study the influence of designated leaders on the collective behavior experimentally by producing leader particles with larger Pt patches.

Prior to concluding this subsection, we will briefly consider one additional approach to the selection of microparticle design parameters. We may think of the designated leader approach to parameter selection as being in line with the broader doctrine of "precision engineering." In some sense, the introduction of a designated leader is similar to asking: What is the most precise perturbation one can make to the underlying system structure in order to realize some desired goal? In what remains of this subsection, we will explore an entirely different approach—one that is much more philosophically in line

with the imprecision-driven, stochastic roots of this thesis. To this end, let all $U_i \in U$ be normally distributed random variables, i.e., $U_i \sim N(\mu, \sigma^2)$. We may motivate this setting by interpreting μ as a desired Pt patch size and σ^2 being the result of imprecision in the manufacturing process.

Now, we will investigate how rattling varies in expectation as a function of μ, σ^2, and the number of particles N. However, the analytical expression for rattling in Eq. 4.7 depends on exponentials of U_i's, as opposed to the random variables themselves. To make progress, we must reason about log-normal random variables instead (i.e., $Y_i = e^{-U_i}$). Then, making use of the Fenton-Wilkinson approximation for sums of log-normal random variables [134] we arrive at the following expression for rattling:

$$\mathcal{R}(\mu, \sigma^2, N) = \log(N-1) - N\left(\mu - \log\left(\frac{1}{2}\left(\sigma^2 - \log\left(1 + \frac{e^{\sigma^2}-1}{N}\right)\right) + \log(Ne^{\mu})\right)\right). \tag{4.9}$$

Fixing μ, we proceed to evaluate this expression as a function of σ^2 across different values of N in Fig. 4.5. As expected, at low values of σ^2 there is no drop in rattling because all particles effectively have identical patch sizes in this regime. However, as we increase σ^2 across ensemble sizes we see a transition point beyond which there is a large drop in expected rattling values. Notably, the size of the drop increases as the ensemble grows in size. This is similar to the effect we saw in Fig. 4.3, where larger numbers of particles created more opportunities for coordination and for drops in rattling. However, unlike Fig. 4.3, in this setting the drop in rattling becomes saturated. This is because, unlike the designated leader setting, even at high variances there may be multiple particles with enhanced reaction rates, leading to competition between particles and hampering self-organization. That being said, we note that the Fenton-Wilkinson approximation of

sums of log-normals tends to fail at the tails of the distribution, which may in fact result in even stronger self-organization.

Thus, *instead of choosing an individual particle and carefully modulating its patch size, we may have been able to achieve a similar effect by instead manufacturing the particles with poor tolerancing*, which is reminiscent of Shannon's reliable circuits made out of unreliable relays [135] and in line with the principles of this thesis. We note that this is an original, unpublished contribution of this thesis. However, in what follows we will proceed by taking the designated leader approach.

4.2.3. Persistent Periodicity via Symmetry-Breaking

Using the analytical model we derived in the previous subsection, we were able to connect a bubble's relative size with its contribution to system-level fluctuations, and in turn collective order. The model's predictions in Fig. 4.2(D) suggest that any deviation in a single particle's bubble size relative to the rest of the ensemble (i.e., with relative burst intensity away from 1x) results in a more orderly system as quantified by lower $\mathcal{R}$. Interestingly, the reduction in $\mathcal{R}$ is found to be particularly significant when a bubble larger (and stronger) than its peers is introduced, which we confirmed with experiments. We note that this novel mechanism for asymmetry-induced order applies to a broad class of complex systems wherein parametric heterogeneities control the fluctuations of strongly interacting elements.

In line with these results, we broke the permutation-symmetry of the original system experimentally by adding a "designated leader" (DL) particle with an enlarged Pt patch of radius 175μm (Fig. 4.2(E)). Note that since the nanometre-scale thickness of the Pt layer

is negligible compared to that of the unchanged 10μm-thick polymeric microdisc, the DL design does not alter the particle's volumetric geometry. However, the heterogeneity among the catalytic surface areas translates directly to unequal bubble growth rates between the DL and its neighbours, which in turn drastically affects their collective dynamics in accordance with our theoretical predictions in Fig. 4.2(D): We observe robustly periodic bubble collapses across N in the sharp peaks of the interarrival distributions in Fig. 4.2(E), suggesting that DLs are able to sustain the periodicity of particle beating even at high particle counts. Figure 4.2(F) depicts the time evolution of the breathing radius for a system of $N = 7 + 1$DL particles (see also Movie S4[4]). In contrast to the homogeneous $N = 8$ system (Fig. 4.2(C)), the heterogeneous DL system exhibits a stable long-term self-oscillation with a period of 14.2s, owing to the broken permutation symmetry.

Figures 4.6(A,i-vii) and (B,i-vii) explain the microscale physics arising from the intentionally broken symmetry (see also Movie S3[5]). When a DL particle with an enlarged Pt patch is paired with a non-DL particle, the heterogeneity in bubble sizes leads to the subsumption of the non-DL particle bubble into the DL bubble upon contact (Figs. 4.6(A,ii-v) and (B,ii-v)). This coalescence behaviour is distinct from that of equal-sized bubbles previously shown in Fig. 4.1(B), where an unstable merged bubble forms halfway between the particles. Instead, the merged bubble sticks to the former location of the large parent bubble underneath the DL particle, seen in Figs. 4.6(A,iii) and (A,v). This behaviour falls under the sticking bubble regime in the literature, a phenomenon long observed in experiments [136, 137] but only recently thoroughly studied and theorized in a catalytic

[4] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM7_ESM.mp4

[5] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM6_ESM.mp4

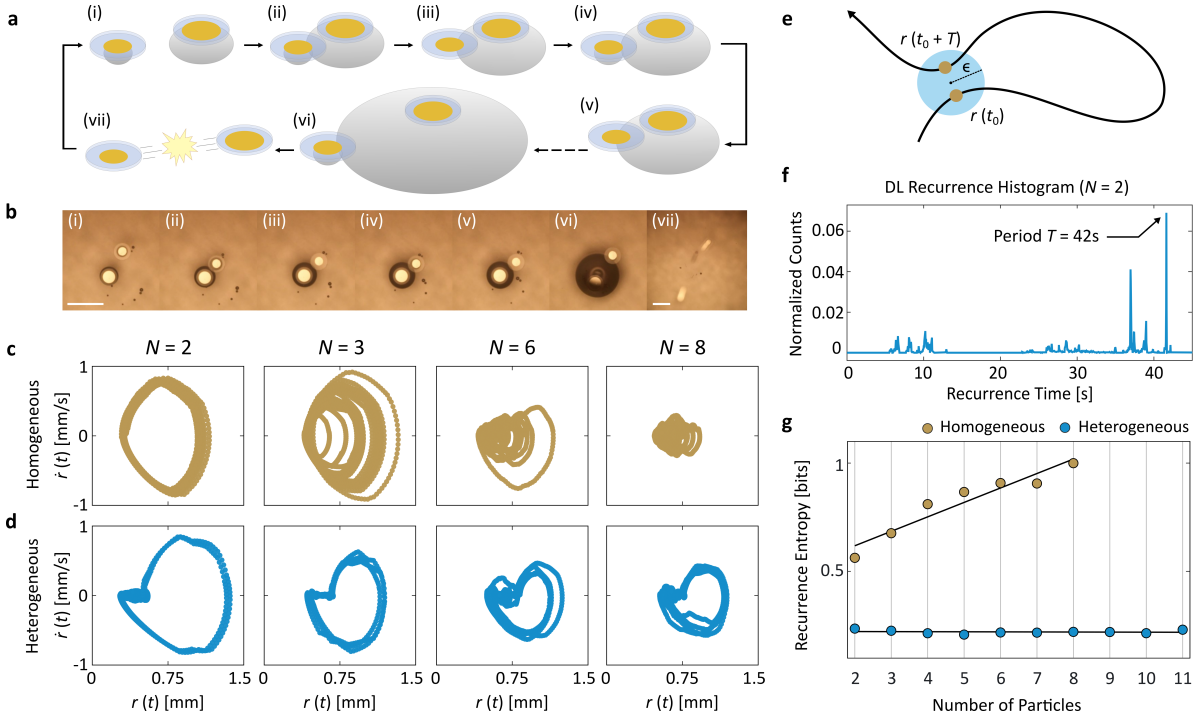


Figure 4.6. **Designated leaders induce periodic limit cycles.** (A),(B), Features of DL beating explained with schematic (A) and micrograph sequence (B) of a 2-particle heterogeneous system. The leader particle is able to grow a large bubble promptly and subsume the smaller bubbles of neighbouring particles across several rounds of bubble coalescence. Scale bars, 1mm. (C),(D), Phase portraits of homogeneous (C) and heterogeneous (D) systems of $N = 2, 3, 6$, and 8. Only the latter is able to maintain the closed-loop orbits at high particle counts. (E), Schematic of recurrence time calculation. The recurrence time is the time it takes to return from a given system configuration to the neighborhood of said configuration. (F), Recurrence histogram compiling all of the recurrence times observed across experiments of the 2-particle heterogeneous system ($N = 1 + 1$DL). (G), Recurrence entropy as a function of N for both homogeneous (yellow) and heterogeneous/DL (blue) systems. Low recurrence entropy is a quantitative indicator of periodic behaviour. The homogeneous system's recurrence entropy trends upward, suggesting a decay in periodicity, while the DL system's entropy remains low in accordance with its observed periodicity even at high N.

H_2O_2 bubble system [138]. Importantly, contrary to the more intuitive moving bubble regime where the merged bubble sits at the centre of mass of its parents [139, 140], the coalescence behaviour transitions into the sticking regime only as the parent bubbles differ sufficiently in size [138], or, in other words, with sufficient particle heterogeneity. As shown in the rest of Figs. 4.6(A) and (B), the two particles in the system undergo several rounds of small-scale bubble coalescence, eventually causing the DL bubble to collapse. We find that the bubble's rupture radius is approximately 1.7 times larger than that for a homogeneous system shown in Fig. 4.1(F), stabilized by the particle sitting directly on top. This contributes to an even lower-frequency chemomechanical oscillation (Figs. 4.2(F) and 4.6(F)) than that previously observed in homogeneous systems (Fig. 4.1(I) and 4.2(B)).

Figures 4.6(C) and (D) contrast the breathing radius phase portraits between homogeneous and heterogeneous systems of different N. We observe that the homogeneous systems experience a decay of periodicity evidenced by the gradual collapse of limit cycle orbits in its phase portraits as a function of N, consistent with trends in Fig. 4.2(B). In contrast, the heterogeneous systems' limit cycles are robust to variations in N, retaining their closed-loop phase-space orbits. To rigorously quantify the effect that DLs have on collective periodicity, we analysed the recurrence structure of the dynamical trajectories across system sizes [121]. As previously discussed and sketched in Fig. 4.6(E), recurrence analyses capture the dynamical properties of system behaviours by measuring the time the system takes to return to a given state's neighbourhood. The set of all such time intervals is compiled into a recurrence histogram (Fig. 4.6(F)) whose recurrence entropy can be used to quantify the complexity of dynamical trajectories [122], with perfect periodicity corresponding to zero entropy.

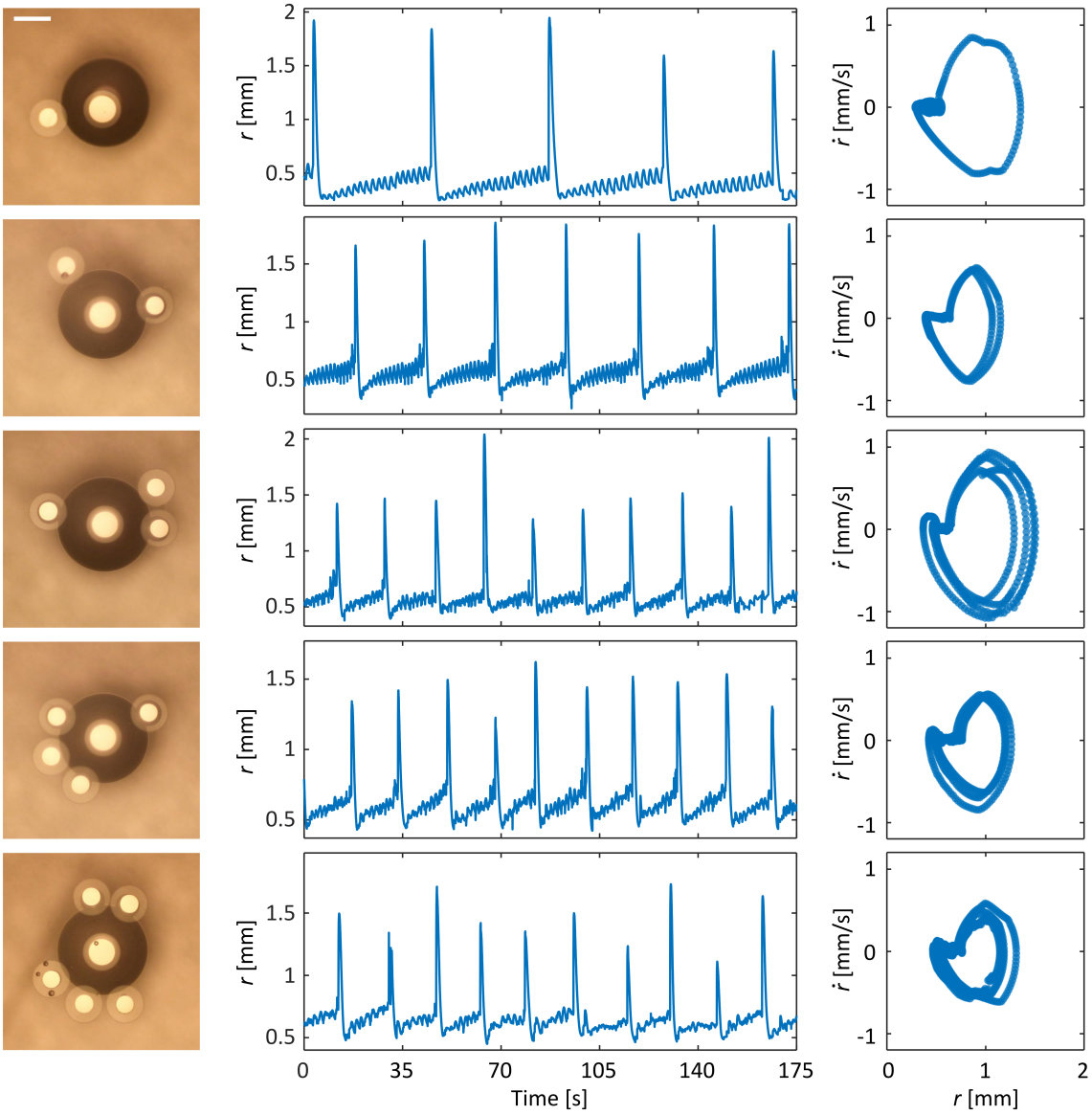


Figure 4.7. **Designated leaders induced limit cycles in $N = 2 - 6$.** Master plots associated with additional phase portrait experiments.

The linear entropy increase for homogeneous systems as a function of N (Fig. 4.6(G)) corresponds to the increasing disorder in the system's recurrences that is consistent with the progressive loss of periodicity observed in Figs. 4.2(C) and 4.6(C). Also in accordance

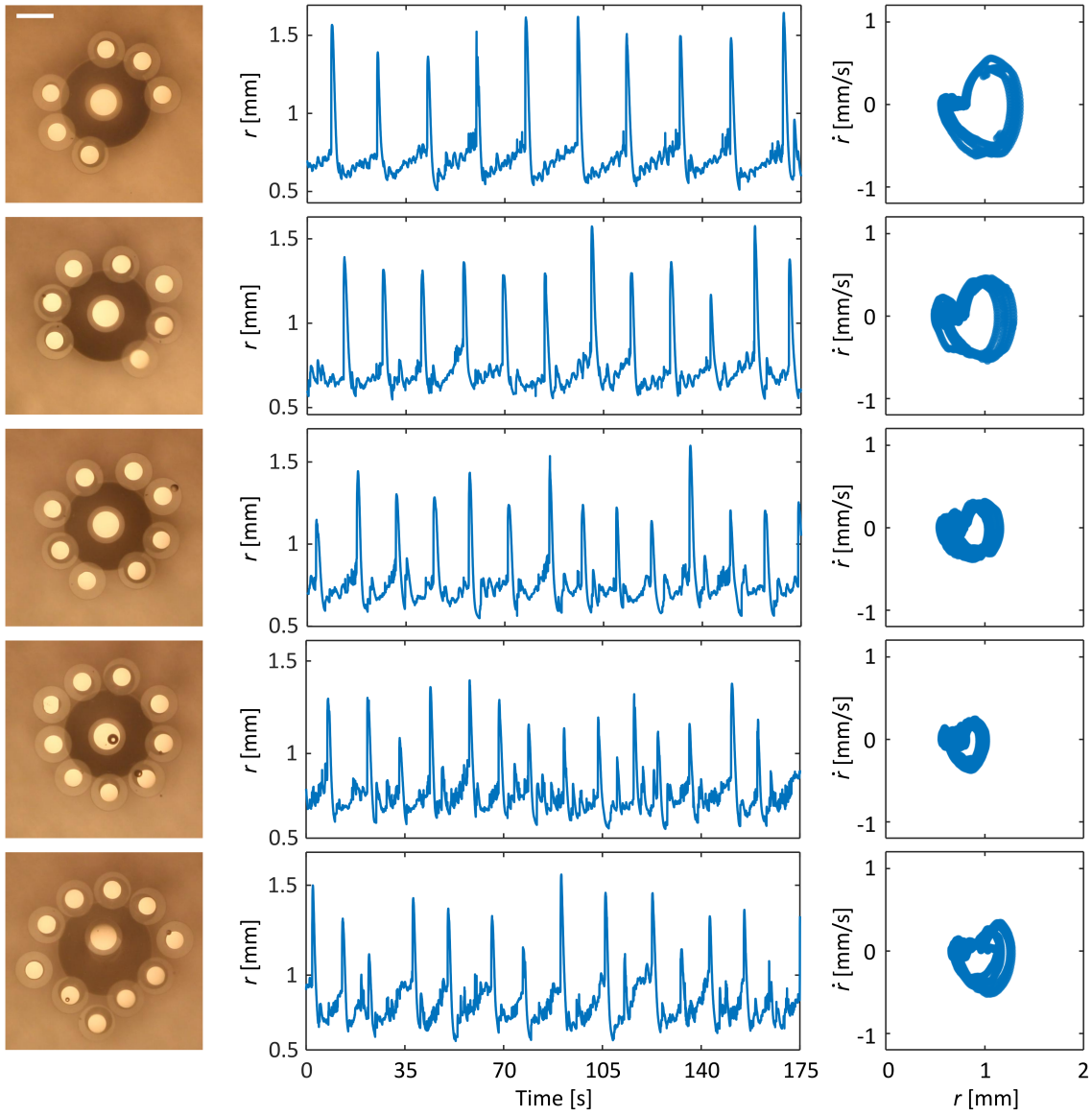


Figure 4.8. **Designated leaders induced limit cycles in $N = 7 - 11$.** Master plots associated with additional phase portrait experiments.

with earlier qualitative trends in Figs. 4.2(F) and 4.6(D), the recurrence entropy of the DL system is locally invariant to changes in N, thereby providing quantitative evidence of the robustness of the periodic beating induced via symmetry-breaking. While we find

that the system's invariance to particle number holds up to $N = 11$, we leave the study of larger particle systems for future work (Figs. 4.7 and 4.8).

4.2.4. Self-Oscillating Microgenerators

Through a simple modification to the particle design, we are able to harness the robust chemomechanical beating to generate an oscillatory electric signal. As illustrated in Fig. 4.9(A) and (B), we fabricated particles with a Pt pattern closely lined up with (though spatially separate from) an additional metal patch of either Au or Ru. With the bimetallic design, the previously auto-redox catalytic decomposition of H_2O_2 on Pt is in part separated into an oxidation half-reaction on Pt and a reduction half-reaction on Ru (Au) [110, 111, 141]:

$$\text{Pt:}\ \ H_2O_2 \longrightarrow O_2 + 2H^+ + 2e^-$$

$$\text{Ru (Au):}\ \ H_2O_2 + 2H^+ + 2e^- \longrightarrow 2H_2O$$

$$\text{Overall:}\ \ 2H_2O_2 \longrightarrow 2H_2O + O_2.$$

Consequently, a potential difference is established at the two electrodes that essentially transforms the particle into an on-board fuel cell. These same principles have been previously used to generate voltages in nanomotors, where bimetallic rods and nanoparticles are propelled electrokinetically by the accompanying electric field [142–144]. A micrograph of our fabricated prototype is displayed in Fig. 4.9(B). Note that the metallic leads extending outwards were added to facilitate electrical characterization of the devices and are not necessary to their operation. The leads were passivated and hence do not

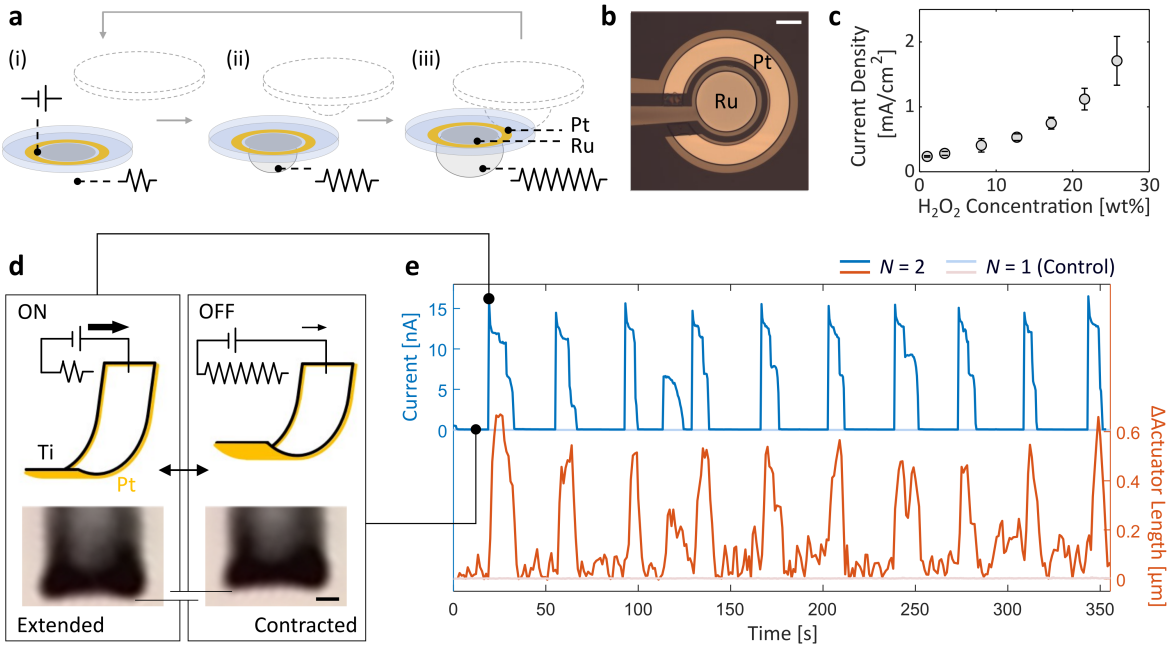


Figure 4.9. **Self-organized oscillation powers a microrobotic arm.** (A), Schematics of the generation of an oscillatory electrical current from chemomechanical beating. The pair of metals (Pt-Ru or Pt-Au) patterned on a polymer base constitute the electrodes of a H_2O_2 fuel cell, which serves as an on-board voltage source. The periodic bubble growth and collapse in a beating system separately modulates the electrical resistance between the electrodes, leading to an oscillatory current. (B), Optical micrograph of a typical Pt-Ru fuel cell particle. The entire surface, less the electrode area, is passivated with a thin layer of insulating SU-8 polymer (shaded). The metallic leads on the left are not necessary for device operation and are included to facilitate measurement. Scale bar, 100μm. (C), Short-circuit current density as a function of H_2O_2 concentration for a Pt-Ru device. (D),(E), Cyclic motion of a microrobotic actuator driven by the oscillatory current. The schematics and micrographs in (D) show the extended and contracted states of the actuator respectively under the ON and OFF current conditions, as modulated by the bubble size. The current measurement over time and the actuator length change (E) closely match, confirming that the cyclic actuation is driven by the oscillatory current, which itself is emergent from the particle beating. Scale bar, 2μm.

participate in any electrochemical reactions. The Pt-Ru and Pt-Au fuel cell devices measured open-circuit voltages of 144.9mV ± 2.4 and 21.4mV ± 3.5, respectively, in a 25.8wt% H_2O_2 solution with 0.075M KNO_3 added for conductivity. We note these values are in line with prior mechanistic studies [110, 111]. Under the same conditions, the Pt-Ru fuel cell delivers a short-circuit current density of 1.71mA/cm^2± 0.38 and a current of 56.7nA ± 12.4. As a benchmark, a significantly larger 1.5 × 6cm thermo-mechano-electrical self-oscillator reported recently recorded a peak current of ∼47nA [145]. The dependence of the current density on H_2O_2 concentration is summarized in Fig. 4.9(C).

As before, the system's collective beating drives the synchronized formation and collapse of bubbles on each particle. However, unlike previous experiments, here the instantaneous size of the bubble also modulates the electrical conductance from one electrode to the other (Fig. 4.9(A), $N = 2$ for demonstration). This effect, in conjunction with the fuel cell's voltage, enables the onboard generation of oscillatory currents that are in phase with the mechanical beating. In a Pt-Ru device, we observe that the ON/OFF ratio between maximal and minimal currents can exceed 10^6, corresponding to when the bubble is absent and at its threshold size. Importantly, the same chemical energy harnessed from the environment is used to simultaneously drive the mechanical oscillation, generate the electrical voltage, and modulate the electrical conductance. Multifunctionality of this kind is emblematic of emerging paradigms such as embodied energy [146], and is crucial to the development of efficient microsystems.

Figures 4.9(D) and (E) exemplify the beating system's capability to cyclically drive a microrobotic load with its self-generated oscillatory electrical current. In this proof-of-concept demonstration, we wired the Ru electrode of a fuel cell particle to a state-of-the-art

Pt-Ti electrochemical microactuator (see Fig. 4.9(D)), originally invented for a tethered sub-100μm walker [147]. In our experimental configuration, charged species from the electrolyte is desorbed from the Pt surface of the bimorph microactuator as current passes through, causing it to deswell and its curvature/length to change. Evident in Fig. 4.9(D), the periodic actuation of the bimorph (red curve, representative snapshots in Fig. 4.9(D), also Movie S5[6]) is driven by the periodic spikes in the current signal (blue curve), which in turn is modulated by the chemomechanical beating of two particles. Because the outer radius of the Pt electrode (Fig. 4.9(B)) exceeds the 125μm patch radius of a standard particle, the system is stabilized by the added heterogeneity, which also explains the observed sub-0.03Hz beating frequency. In contrast, the control experiments in Fig. 4.9(E) show the actuator idling in the absence of a second particle and hence any mechanical beating. By harnessing the emergent power generation of an ensemble of microparticles, we have demonstrated the design and modular interoperability of key microrobotic components—energy sources and locomotive elements—based on the physics of self-organization.

4.3. Discussion

Through the discovery of physical mechanisms for asymmetry-induced order, we constructed self-oscillating electrical generators capable of powering on-board microrobotic components from the interactions of simple microparticles. Our results stand in contrast to more traditional microrobotic approaches focusing on the design of intricate electromechanical assemblies to produce alternating electrical currents [94]. By relying on our

[6] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-022-33396-5/MediaObjects/41467_2022_33396_MOESM8_ESM.mp4

system's self-organized behaviours, we circumvented the design of complex contraptions to harvest and transduce chemical energy into periodic electrical and mechanical work—a crucial step towards fully-autonomous microrobots [146, 148]. The use of on-board electrical currents will enable the integration of sensors and computational elements to enrich physical microparticle interactions [149], forming the basis for future collectives wherein the long-envisioned potential of complex inter-particle communications can be implemented [120]. We plan on extending our approach into studying larger collections of microparticles in search of general principles for the top-down design of active matter systems, where an understanding of system symmetries and environmental forcing may enhance their task-capability. Unifying perspectives from their respective fields, our work suggests that future microrobots and active matter systems may become more robust and task-capable when we design them to exploit the physics of the environments they inhabit.

This chapter exemplifies a *robot thermodynamics* approach to microsystem design. We used low-rattling selection—itself derived from path-continuity-constrained maximum caliber trajectory statistics (in Ch. 2.4)—as a guiding principle in the design of a collective of active microparticles. In doing so, we were able to optimize the properties of microparticle collectives and achieve novel task-capabilities. In the following chapter, we will explore a robot thermodynamics approach to reinforcement learning. By encouraging agents to satisfy maximally diffusive trajectory statistics, we will find that many of the limitations of reinforcement learning in robotics can be overcome.

CHAPTER 5

Overcoming Temporal Correlations in Robot Learning

Throughout this thesis, we have argued for the use of path distributions as mathematical primitives for inference and synthesis of goal-directed behaviors, proposing robot thermodynamics as a framework to achieve this. In Chs. 2.4 and 3, we explored the properties of systems with bounded velocity fluctuations. We found that the path distributions of such systems admit a "low-rattling selection principle" capable of predicting global steady-state statistics from a local measure of order, termed rattling. In Ch. 4, we explored the use of design parameters as a means of reshaping an agent's path distribution, opting to choose parameters that minimize a system's rattling as a means of inducing self-organization. In doing so, we were able to make a simple collection of active particles generate alternating currents to power state-of-the-art microrobot arms. While choosing design parameters provides us with means of making gross-level changes to an agent's path distribution, here we will make use of reinforcement learning and policy optimization to reshape path distributions—and, in doing so, fully realize the promise of robot thermodynamics as a learning and control framework. By centering our approach on maximally diffusive trajectory statistics (see Ch. 2.6), in this chapter we are able to develop a reinforcement learning framework that overcomes the temporal correlations that plague embodied learning agents. We note that most of the work in this chapter was previously published in [7], which this thesis contributed to in its entirety.

Robots and animals both experience the world through their bodies and senses. Their embodiment constrains their experiences, ensuring they unfold continuously in space and time. As a result, the experiences of embodied agents are intrinsically correlated. Correlations create fundamental challenges for machine learning, as most techniques rely on the assumption that data are independent and identically distributed. In reinforcement learning, where data are directly collected from an agent's sequential experiences, violations of this assumption are often unavoidable. Here, we derive a method that overcomes this issue by exploiting the statistical mechanics of ergodic processes, which we term maximum diffusion reinforcement learning. By decorrelating agent experiences, our approach provably enables single-shot learning in continuous deployments over the course of individual task attempts. Moreover, we prove our approach generalizes well-known maximum entropy techniques, and robustly exceeds state-of-the-art performance across popular benchmarks. Our results at the nexus of physics, learning, and control form a foundation for transparent and reliable decision-making in embodied reinforcement learning agents.

5.1. Introduction

Reinforcement learning (RL) is a flexible decision-making framework based on the experiences of artificial agents, whose potential for scalable real-world impact has been well-established through the power of deep learning architectures. From controlling nuclear fusion reactors [150] to besting curling champions [151], RL agents have achieved remarkable feats when they can exhaustively explore how their actions impact the state of their environment. Despite their impressive achievements, RL agents—especially deep RL agents—suffer from limitations preventing their widespread deployment in the real world:

Their performance varies across initializations, their sample inefficiency demands the use of simulators, and they struggle to learn outside of episodic problem structures [152–154]. At the heart of these shortcomings lies a violation of the assumption that data are independent and identically distributed (*i.i.d.*), which underlies most of machine learning. While learning typically requires *i.i.d.* data, the experiences of RL agents are unavoidably sequential and correlated across points in time. It is no wonder, then, that many of deep RL's most impactful advances have sought to overcome this roadblock [19, 155, 156].

Over the past decades, researchers have started to converge onto an understanding that destroying temporal correlations is essential to sample efficiency and agent performance, seeking to address them in two primary domains: During optimization and during sample generation. When we consider optimizing a policy from a database of sequential agent-environment interactions, sampling in random batches is known to reduce temporal correlations. For this reason, experience replay [157] and its many variants [158–160] have been successful in producing large performance and sample efficiency gains across tasks and algorithms [161–163]. This simple insight—merely sampling experiences out of order—was a key contributing factor to one of deep RL's landmark triumphs: Achieving superhuman performance in Atari video game benchmarks [164].

Nonetheless, temporal correlations also arise during data generation, where their impact cannot be alleviated through sampling alone. In turn, temporal correlations must be mitigated during data acquisition as well, which requires techniques to sufficiently randomize the sample generation process. In this regard, the maximum entropy (MaxEnt) RL framework has emerged as a key advance [18, 36, 41, 165–170]. These methods seek to generate randomness during optimization and data acquisition by maximizing the entropy

of an agent's policy, which decorrelates their action sequences. In doing so, MaxEnt RL techniques have been able to achieve better exploration and more robust performance [42]. However, does maximizing the entropy of an agent's policy actually decorrelate their experiences?

In this chapter, we will prove that this is generally not the case. To address this gap we introduce maximum diffusion (MaxDiff) RL, a framework that provably decorrelates agent experiences during sample generation, and realizes statistics indistinguishable from *i.i.d.* sampling by exploiting the statistical mechanics of ergodic processes. Our approach efficiently exceeds state-of-the-art performance by diversifying agent experiences and improving state exploration. By articulating the relationship between an agent's embodiment, diffusion, and learning, we prove that MaxDiff RL agents are capable of single-shot learning regardless of how they are initialized. We additionally prove that MaxDiff RL agents are robust to random seeds and environmental stochasticity, which enables consistent and reliable performance with low-variance across agent deployments and learning tasks. The work in this chapter sheds a light on foundational issues holding back the field, highlighting the impact that agent properties and data acquisition can play on downstream learning tasks, and paving the way towards more transparent and reliable decision-making in embodied RL agents.

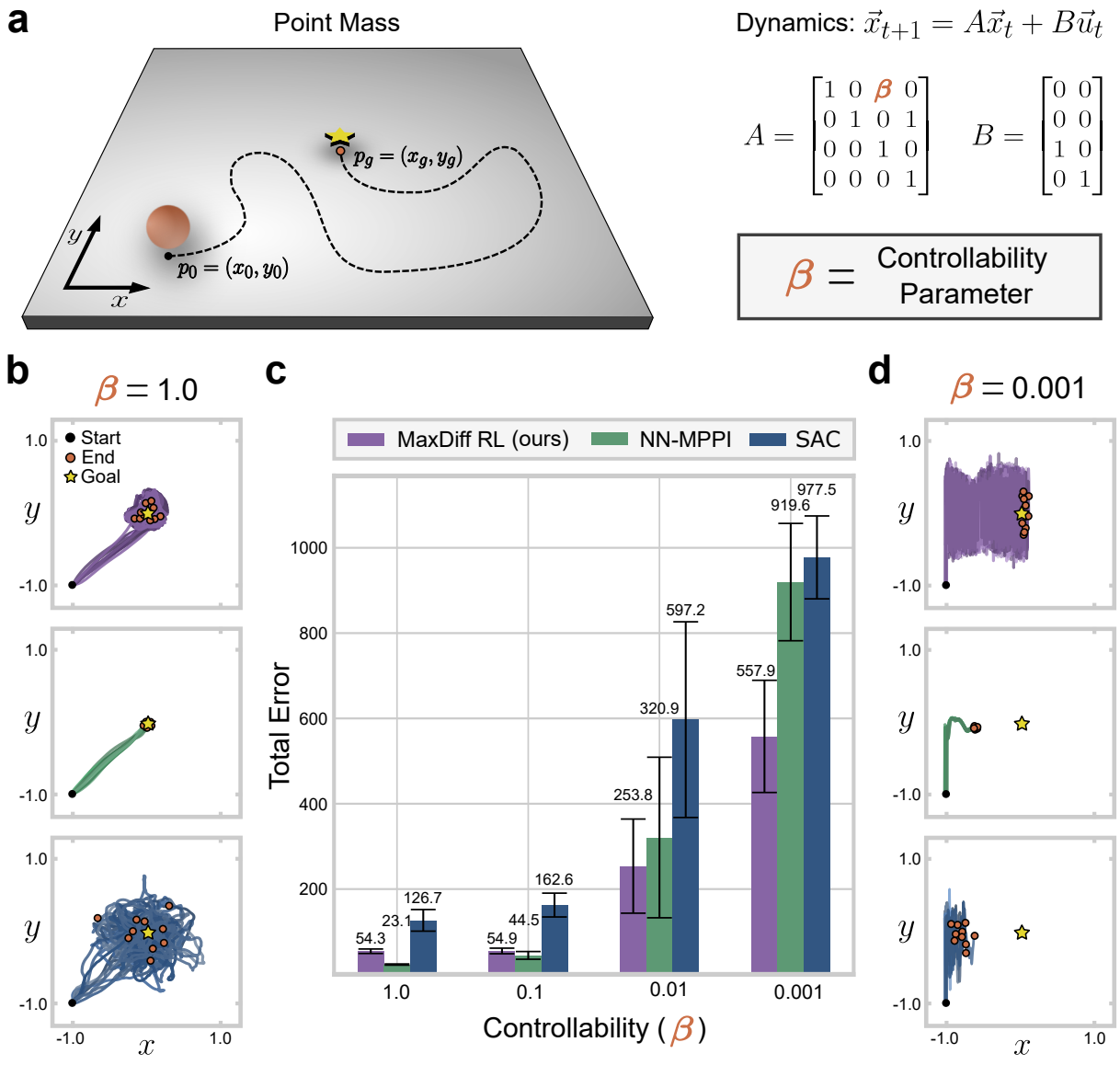
5.2. Results

5.2.1. Temporal Correlations Hinder Performance

Whether temporal correlations and their impact can be avoided depends on the properties of the underlying agent-environment dynamics. Completely destroying correlations between

agent experiences requires the ability to discontinuously jump from state to state without continuity of experience. For some RL agents, this poses no issue. In settings where agents are disembodied, there may be nothing preventing effective exploration through jumps between uncorrelated states. This is one of the reasons why deep RL recommender systems have been successful in a wide range of applications, such as YouTube video suggestions [171–173]. However, continuity of experience is essential to many RL problem domains. For instance, the smoothness of Newton's laws makes correlations unavoidable in the motions of most physical systems, even in simulation. This suggests that for systems like robots or self-driving cars overcoming the impact of temporal correlations on performance presents a major challenge.

To illustrate the impact this can have on learning performance, we devised a toy task to evaluate deep RL algorithms as a function of correlations intrinsic to the agent's state transitions. Our toy task and agent dynamics are shown in Fig. 5.1(A), corresponding to a double integrator system with parametrized momentum anisotropy. The task requires learning reward, dynamics, and policy representations from scratch in order to move a planar point mass from a fixed initial position to a goal location. The system's true linear dynamics are simple enough to explicitly write down, which allows us to rigorously study temporal correlations across state transitions by analyzing its controllability. Controllability is a formal property of control systems that describes their ability to reach arbitrary states in an environment [21, 22]. In linearizable systems, state transitions become degenerate and irreversibly correlated when they are uncontrollable. However, if the agent is controllable the impact of correlations can be overcome, at least in principle. While the relationship between controllability and temporal correlations has been studied for decades [174], it is



only recently that researchers have begun to study its impact on learning processes [175–177].

Figure 5.1 parametrically explores the relationship between our toy system's controllability properties and the learning performance of state-of-the-art deep RL algorithms. The point mass dynamics are parametrized by $\beta \in [0, 1]$, which determines the relative difficulty

Figure 5.1. **Temporal correlations break the state-of-the-art in RL.** For most systems, controllability properties determine temporal correlations between state transitions (see Ch. 2.3.2). (A), Planar point mass with dynamics simple enough to explicitly write down and whose policy admits a globally optimal analytical solution. The system's 4-dimensional state space is comprised of its planar positions and velocities. We parametrize its controllability through $\beta \in [0, 1]$, where $\beta = 0$ produces a formally uncontrollable system. The task is to translate the point mass from p_0 to p_g within a fixed number of steps at different values of β, and the reward is specified by the negative squared Euclidean distance between the agent's state and the goal. We compare state-of-the-art model-based and model-free algorithms, NN-MPPI and SAC respectively, to our proposed maximum diffusion (MaxDiff) RL framework. (B),(D), Representative snapshots of MaxDiff RL, NN-MPPI, and SAC agents (top to bottom) in well-conditioned ($\beta = 1$) and poorly-conditioned ($\beta = 0.001$) controllability settings. (C), Even in this simple system, poor controllability can break the performance of RL agents. As $\beta \to 0$ the system's ability to move in the x-direction diminishes, hindering the performance of NN-MPPI and SAC, while MaxDiff RL remains task-capable. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test.

of translating along the x-axis (Fig. 5.1(A)). When $\beta = 0$ the system is uncontrollable and can only translate along the y-axis, which illustrates the sense in which state transitions become irreversibly correlated. While the system is formally controllable for all non-zero β, as $\beta \to 0$ fewer lateral transitions become available for the same range of actions, introducing temporal correlations along the system's x-coordinate (see Fig. 2.4 also). We evaluated the performance of state-of-the-art model-based and model-free deep RL algorithms on our task—model-predictive path integral control (NN-MPPI) [46] and soft actor-critic (SAC) [19], respectively—at varying values of β, from 1 to 0.001. As expected, at $\beta = 1$ both NN-MPPI and SAC are able to accomplish the toy task (Fig. 5.1(B)).

However, as $\beta \to 0$ the performance of NN-MPPI and SAC degrades parametrically (Fig. 5.1(C)), up until the point that neither algorithm can solve the task, as shown in Fig. 5.1(D). Hence, temporal correlations can completely hinder the learning performance of the state-of-the-art in deep RL even in toy problem settings such as this one, where a globally optimal policy can be analytically computed in closed form.

Failure to mitigate temporal correlations between state transitions can prevent effective exploration, severely impacting the performance of deep RL agents. As Fig. 5.1(D) illustrates, neither NN-MPPI nor SAC agents are able to sufficiently explore in the x-dimension of their state space as a result of their decreasing degree of controllability (see Ch. 2.3.2). This is the case despite the fact that NN-MPPI and SAC are both MaxEnt RL algorithms [19, 20], designed specifically to achieve improved exploration outcomes by decorrelating their agent's action sequences. In contrast, our proposed approach—MaxDiff RL—is able to consistently succeed at the task and is guaranteed to realize effective exploration by focusing instead on decorrelating agent experiences, i.e., their state sequences (see purple in Fig. 5.1(B-D)), as we discuss in the following subsection.

5.2.2. Maximum Diffusion Exploration and Learning

Most RL methods presuppose that taking random actions produces effective exploration [178, 179], and sophisticated techniques like MaxEnt RL are no different. However, as we previously illustrated, whether this is actually possible depends on the agent's controllability properties and the temporal correlations these spontaneously induce in their experiences (see Fig. 5.2(C) and Ch. 2.3.2). To overcome these limitations, we propose

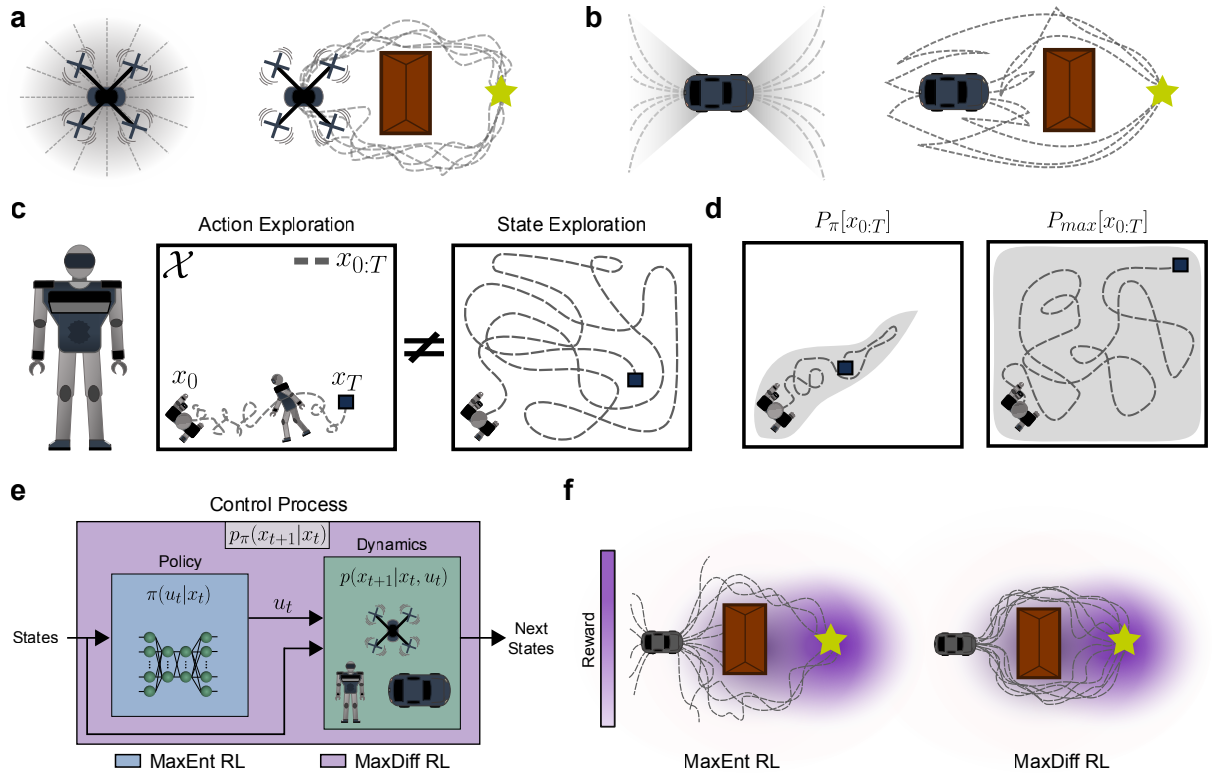


Figure 5.2. **Maximum diffusion RL mitigates temporal correlations to achieve effective exploration**. (A),(B), Systems with different planar controllability properties. (C), Whether action randomization leads to effective state exploration depends on the properties of the underlying state-transition dynamics (see Ch. 2.3.2), as in our illustration of a complex bipedal robot falling over and failing to explore. (D), While any given policy induces a path distribution (left), MaxDiff RL produces policies that maximize the path distribution's entropy (right). The projected support of the robot's path distribution is illustrated by the shaded gray region. We prove that maximizing the entropy of an agent's state transitions results in effective exploration (see Chs. 2.3.4 and 2.5.1). (E), Our approach generalizes the MaxEnt RL paradigm by provably optimizing trajectory entropy, as we show in this chapter. (F), This leads to distinct learning outcomes because agents reason about the impact of their actions on state transitions, rather than their actions alone.

decorrelating agent experiences as opposed to their action sequences, which forms the starting point to our derivation of the MaxDiff RL framework.

Prior to synthesizing policies or assessing their impact on learning outcomes, we require a formalization of agent experiences. Without considering policies, we see the agent-environment state transition dynamics as an autonomous stochastic process, whose sample paths $x(t)$ take value in a state space $\mathcal{X} \subset \mathbb{R}^d$ at each point in time within an interval $\mathcal{T} = [t_0, t]$. Then, we see agent experiences as collections of random variables parametrized by time, whose realizations $x(t)$ are the sample paths of the underlying agent-environment process. When $\mathcal{T} = \{1, \cdots, T\}$ is discrete, we use $x_{1:T}$ instead of $x(t)$. In this context, the probability density function over state trajectories, $P[x(t)]$ (or $P[x_{1:T}]$), completely characterizes an agent's experiences and their properties (see Ch. 2.1.2). We may now begin our derivation by asking: *What is the most decorrelated that agent experiences can be?*

To answer this question, we draw from the statistical physics literature on maximum caliber [6, 11, 16], which is a generalization of the variational principle of maximum entropy [180]. The goal of a maximum caliber variational optimization is to find the trajectory distribution $P_{max}[x(t)]$, which optimizes an entropy functional $S[P[x(t)]]$. The optimal distribution would then describe the paths of an agent with the least-correlated experiences, but its specific form and properties depend on how the variational optimization is constrained. Without constraints, agents could sample states discontinuously and uniformly in a way that is equivalent to *i.i.d.* sampling but is not consistent with the continuous experiences of embodied agents (Fig. 5.2(A,B)). Hence, we tailor our assumptions to agents with continuous experiences. Then, to ensure our optimization produces a distribution over continuous paths, we constrain the volume of states accessible within any finite time interval by bounding their fluctuations (see Ch. 2.3.3).

As we have already seen in Ch. 2.3.4, this constrained variational optimization surprisingly admits an analytical solution for the maximum entropy path distribution. The derived optimal path distribution is

$$P_{max}[x(t)] = \frac{1}{Z} \exp\Big[-\frac{1}{2} \int_{t_0}^{t} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau) d\tau \Big], \tag{5.1}$$

where Z is a normalization constant. At every point in space $x^* \in \mathcal{X}$, the matrix $\mathbf{C}[x^*]$ measures temporal correlations locally over an interval of duration Δt, such that

$$\mathbf{C}[x^*] = \int_{t_i}^{t_i+\Delta t} K_{XX}(t_i, \tau) d\tau, \tag{5.2}$$

where $K_{XX}(t_1, t_2)$ is the autocovariance of $x(t)$ at pairs of samples in time evaluated over a chosen interval, $[t_i, t_i + \Delta t] \subset \mathcal{T}$, with a given $x(t_i) = x^*$ (see Ch. 2.3.3). This distribution describes the trajectories of an optimal agent with minimally correlated paths, subject to the constraints imposed by continuity of experience. Moreover, Eq. 5.1 is equivalent to the path distribution of an anisotropic, spatially-inhomogeneous diffusion process. Thus, minimizing correlations among agent trajectories leads to diffusion-like exploration, whose properties can actually be analyzed using statistical mechanics (e.g., Fig. 2.5). This also means that the sample paths of the optimal agent are Markovian and ergodic (see Chs. 2.3.4 and 2.5.1 for associated theorems, corollaries, and their proofs). Unlike alternative RL frameworks, our approach does not assume the Markov property, but rather enforces it as a property intrinsic to the optimal agent's path distribution.

Satisfying ergodicity has profound implications for the properties of resulting agents. Ergodicity is a formal property of dynamical systems that guarantees that the statistics of individual trajectories are asymptotically equivalent to those of a large ensemble of

trajectories [30, 181]. Put in RL terms, while the sequential nature of RL agent experiences can make *i.i.d.* sampling technically impossible, the global statistics of an ergodic RL agent are indistinguishable from those of an *i.i.d.* sampling process. In this sense, ergodic Markov sampling is the best possible alternative to *i.i.d.* sampling in sequential decision-making processes. Beyond resolving the issue of generating *i.i.d.* samples in RL, ergodicity forms the basis of many of MaxDiff RL's theoretical guarantees, as we show in the following subsections.

When an agent's trajectories satisfy Eq. 5.1, we describe the agent as maximally diffusive. However, agents do not realize maximally diffusive trajectories spontaneously. Doing so requires finding a policy capable of satisfying maximally diffusive path statistics, which forms the core of what we term MaxDiff RL. While any given policy induces a path distribution, finding policies that realize maximally diffusive trajectories requires optimization and learning (Fig. 5.2(D)). To this end, we define:

$$
\begin{aligned}
P_\pi[x_{1:T}, u_{1:T}] &= \prod_{t=1}^{T-1} p(x_{t+1}|x_t, u_t)\pi(u_t|x_t) \\
P^r_{max}[x_{1:T}, u_{1:T}] &= \prod_{t=1}^{T-1} p_{max}(x_{t+1}|x_t)e^{r(x_t,u_t)},
\end{aligned}
\tag{5.3}
$$

where we discretized the distribution in Eq. 5.1 as $p_{max}(x_{t+1}|x_t)$, and analytically rederived the optimal path distribution under the influence of a reward landscape, $r(x_t, u_t)$ (see Ch. 2.5.1). Given the distributions in Eq. 5.3, the goal of MaxDiff RL can be framed as minimizing the Kullback-Leibler (KL) divergence between them—that is, between the agent's current path distribution and the maximally diffusive one.

To draw connections between our framework and the broader MaxEnt RL literature, we recast the KL-divergence formulation of MaxDiff RL as an equivalent stochastic optimal control (SOC) problem. In SOC, the goal is to find a policy that maximizes the expected cumulative rewards of an agent in an environment. In this way, we can express the MaxDiff RL objective as

$$\pi^*_{\text{MaxDiff}} = \underset{\pi}{\text{argmax}}\ E_{(x_{1:T},u_{1:T})\sim P_\pi}\left[\sum_{t=1}^{T-1}\gamma^t \hat{r}(x_t,u_t)\right], \tag{5.4}$$

with $\gamma \in [0,1)$ and modified rewards given by

$$\hat{r}(x_t,u_t) = r(x_t,u_t) - \alpha \log \frac{p(x_{t+1}|x_t,u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)}, \tag{5.5}$$

where $\alpha > 0$ is a temperature-like parameter we introduce to balance diffusive exploration and reward exploitation, as we discuss in the following section. With these results in hand, we may now state and prove one of our main theorems.

Theorem 5.1. *(MaxDiff RL generalizes MaxEnt RL) Let the state transition dynamics due to a policy π be $p_\pi(x_{t+1}|x_t) = E_{u_t\sim\pi}[p(x_{t+1}|x_t,u_t)]$. If the state transition dynamics are assumed to be decorrelated, then the optimum of Eq. 5.4 is reached when $D_{KL}(p_\pi||p_{max}) = 0$ and the MaxDiff RL objective reduces to the MaxEnt RL objective.*

Proof. Since controllability is central to this proof, we must first formalize and define a particular notion of controllability in the context of MDPs that was partially introduced in [182], implicit in the results of [167], and explicitly called out in [36].

Definition 5.1. *The state transition dynamics, $p(x_{t+1}|x_t, u_t)$, in an MDP, $(\mathcal{X}, \mathcal{U}, p, r, \gamma)$, are fully controllable when there exists a policy, $\pi : \mathcal{U} \times \mathcal{X} \to [0, \infty)$, such that:*

$$p_\pi(x_{t+1}|x_t) = E_{u_t \sim \pi(\cdot|x_t)}[p(x_{t+1}|x_t, u_t)] \tag{5.6}$$

and

$$D_{KL}\Big(p_\pi(x_{t+1}|x_t)\Big|\Big|\nu(x_{t+1}|x_t)\Big) = 0, \quad \forall t \in \mathbb{Z}^+ \tag{5.7}$$

for any arbitrary choice of state transition probabilities, $\nu : \mathcal{X} \times \mathcal{X} \to [0, \infty)$.

Thus, a system is *fully controllable* when it is simultaneously capable of reaching every state and controlling *how* each state is reached. In other words, a fully controllable agent can arbitrarily manipulate its state transition probabilities, $p_\pi(x_{t+1}|x_t)$, by using an optimized policy to match any desired transition probabilities, $\nu(x_{t+1}|x_t)$. Equipped with the definition of full controllability, we may now proceed with our proof.

We will now proceed from the undiscounted MaxDiff RL objective expressed in terms of a loss function without loss of generality,

$$\pi^*_{\text{MaxDiff}} = \underset{\pi}{\text{argmin}}\ E_{(x_{1:T}, u_{1:T}) \sim P_\pi}\left[\sum_{t=1}^{T-1} \hat{l}(x_t, u_t)\right],$$

where

$$\hat{l}(x_t, u_t) = l(x_t, u_t) + \alpha \log \frac{p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)},$$

and we may think of $l(x_t, u_t)$ as $-r(x_t, u_t)$. Taking this objective we may rearrange terms in the following way:

$$\begin{aligned}
E_{P_\pi}\left[\sum_{t=1}^{T-1}\hat{l}(x_t,u_t)\right] &= E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t) + \alpha\log\frac{p(x_{t+1}|x_t,u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)}\right] \\
&= E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t)\right] + \sum_{t=1}^{T-1} E_{(x_t,u_t)\sim p,\pi}\left[\alpha\log\frac{p(x_{t+1}|x_t,u_t)\pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t)}\right] \\
&= E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t) + \alpha\log\pi(u_t|x_t)\right] \\
&\qquad + \sum_{t=1}^{T-1} E_{(x_t,u_t)\sim p,\pi}\left[\alpha\log\frac{p(x_{t+1}|x_t,u_t)}{p_{max}(x_{t+1}|x_t)}\right].
\end{aligned}$$

Now, we proceed by applying Jensen's inequality to the last term of our expression above—bringing in the expectation over control actions into the logarithm, noting that $E_{u_t\sim\pi}[p_{max}(x_{t+1}|x_t)] = p_{max}(x_{t+1}|x_t)$, and doing more algebraic manipulations:

$$\begin{aligned}
&\leq E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t) + \alpha\log\pi(u_t|x_t)\right] + \sum_{t=1}^{T-1} E_{x_t\sim p}\left[\alpha\log\frac{E_{u_t\sim\pi}[p(x_{t+1}|x_t,u_t)]}{p_{max}(x_{t+1}|x_t)}\right] \\
&\leq E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t) + \alpha\log\pi(u_t|x_t)\right] + \sum_{t=1}^{T-1} E_{x_t\sim p}\left[\alpha\log\frac{p_\pi(x_{t+1}|x_t)}{p_{max}(x_{t+1}|x_t)}\right] \\
&\leq E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t,u_t) + \alpha\log\pi(u_t|x_t) + \alpha D_{KL}\big(p_\pi(x_{t+1}|x_t)\big|\big|p_{max}(x_{t+1}|x_t)\big)\right], \qquad (5.8)
\end{aligned}$$

where we also used the definition of $p_\pi(x_{t+1}|x_t)$ from Eq. 5.6.

To conclude our proof, we must show that the MaxEnt RL objective emerges from the MaxDiff RL objective under the assumption that an agent's state transitions are decorrelated. We can formalize what decorrelation requires of an agent in one of two contexts—that of agents with continuous experiences, or in general. Our derivations

throughout Chs. 2.3.4 and 2.5.1 achieve this in the context of agents with continuous experiences. Therein, we proved that the least-correlated continuous agent paths uniquely satisfy maximally diffusive trajectory statistics, which requires that $D_{KL}(p_\pi||p_{max}) = 0$ when there exists an optimizing policy π. Alternatively, completely decorrelating the state transitions of an agent in general requires being able to generate arbitrary jumps between states, which requires full controllability (see Definition 5.1). Given full controllability, the optimum of Eq. 5.8 is also reached when $D_{KL}(p_\pi||p_{max}) = 0$.

Applying the assumption of decorrelated state transitions in either of the two senses expressed above not only simplifies Eq. 5.8 by removing the KL divergence term but also by saturating Jensen's inequality, which recovers the equality between the left and right hand sides of our equations:

$$E_{P_\pi}\left[\sum_{t=1}^{T-1} \hat{l}_c(x_t, u_t)\right] = E_{P_\pi}\left[\sum_{t=1}^{T-1} l(x_t, u_t) + \alpha \log \pi(u_t|x_t)\right],$$

where we added the subscript c to indicate that this applies under the assumption of decorrelated state transitions—either in the context of agents with continuous paths (with maximum diffusivity as a necessary condition) or in general (with full controllability as a sufficient condition). Putting together our final results, we may now write down the simplified MaxDiff RL optimization objective with the added assumption of decorrelated state transitions:

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\ E_{(x_{1:T}, u_{1:T})\sim P_\pi}\left[\sum_{t=1}^{T-1} \hat{l}_c(x_t, u_t)\right], \tag{5.9}$$

with

$$\hat{l}_c(x_t, u_t) = l(x_t, u_t) + \alpha \log \pi(u_t|x_t), \tag{5.10}$$

or equivalently, we can write Eq. 5.9 as a maximization by replacing the cost with a reward function:

$$\hat{r}_c(x_t, u_t) = r(x_t, u_t) + \alpha \mathcal{H}(\pi(u_t|x_t)), \tag{5.11}$$

where we briefly changed our entropy notation, using $\mathcal{H}(\pi(u_t|x_t)) = S[\pi(u_t|x_t)]$, to highlight similarities with other results in the literature. Crucially, we recognize this objective as the MaxEnt RL objective [19, 42], which proves that MaxDiff RL is a strict generalization of MaxEnt RL to agents with temporally correlated experiences and concludes our proof. □

Crucially, we note that this result does not merely prove that MaxDiff RL is a generalization of the MaxEnt RL framework to agents with correlations in their state transitions. It also also proves that maximizing policy entropy cannot decorrelate agent experiences in general because maximizing policy entropy does not minimize $D_{KL}(p_\pi||p_{max})$ in Eq. 5.8. In contrast, MaxDiff RL actively enforces path decorrelation at all points in time. We can think of this intuitively by noting that MaxDiff RL simultaneously accounts for the effect of the policy and of the temporal correlations induced by agent-environment dynamics in its optimization (Fig. 5.2(E)). As such, MaxDiff RL typically produces distinct learning outcomes from MaxEnt RL (Fig. 5.2(F)). Our result also implies that all theoretical robustness guarantees of MaxEnt RL (e.g., [42]) should be interpreted as guarantees of MaxDiff RL when state transitions are decorrelated. Moreover, we suggest that many of

the gaps between MaxEnt RL's theoretical results and their practical performance may be explained by the impact of temporal correlations, as we saw in Fig. 5.1.

While these results seem to suggest that model-free implementations of MaxDiff RL are not feasible, we note that local estimates of the agent's path entropy can be learned from observations. This effectively reinterprets temporal correlations as a state-dependent property of the environment, which is consistent with the way we model temporal correlations. Similar entropy estimates have been used in model-free RL [183] and more broadly in the autoencoder literature [184]. For the results presented in this chapter, we also derived a model-agnostic objective using an analytical expression for the local path entropy,

$$\underset{\pi}{\operatorname{argmax}}\ E_{(x_{0:T},u_{0:T})\sim P_\pi}\left[\sum_{t=0}^{T-1} r(x_t,u_t) + \frac{\alpha}{2}\log\det \mathbf{C}[x_t]\right], \tag{5.12}$$

whose optimum realizes the same optimum as Eq. 5.4, and where we omitted γ. There are many ways to express the MaxDiff RL objective, each of which may have implementation-specific advantages (see Fig. 5.3(A) and Ch. 2.6.2). In this sense, MaxDiff RL is not a specific algorithm implementation but rather a general problem statement and solution framework, similar to MaxEnt RL. In this chapter, our MaxDiff RL implementation is exactly identical to NN-MPPI *except for the path entropy term* shown above. However, this simple modification can have a *drastic* effect on agent outcomes.

5.2.3. Robustness to Initializations in Ergodic Agents

The introduction of an entropy term in Eq. 5.12 means that MaxDiff RL agents must balance between two aims: Achieving the task and embodying diffusion (Fig. 5.3(A)).

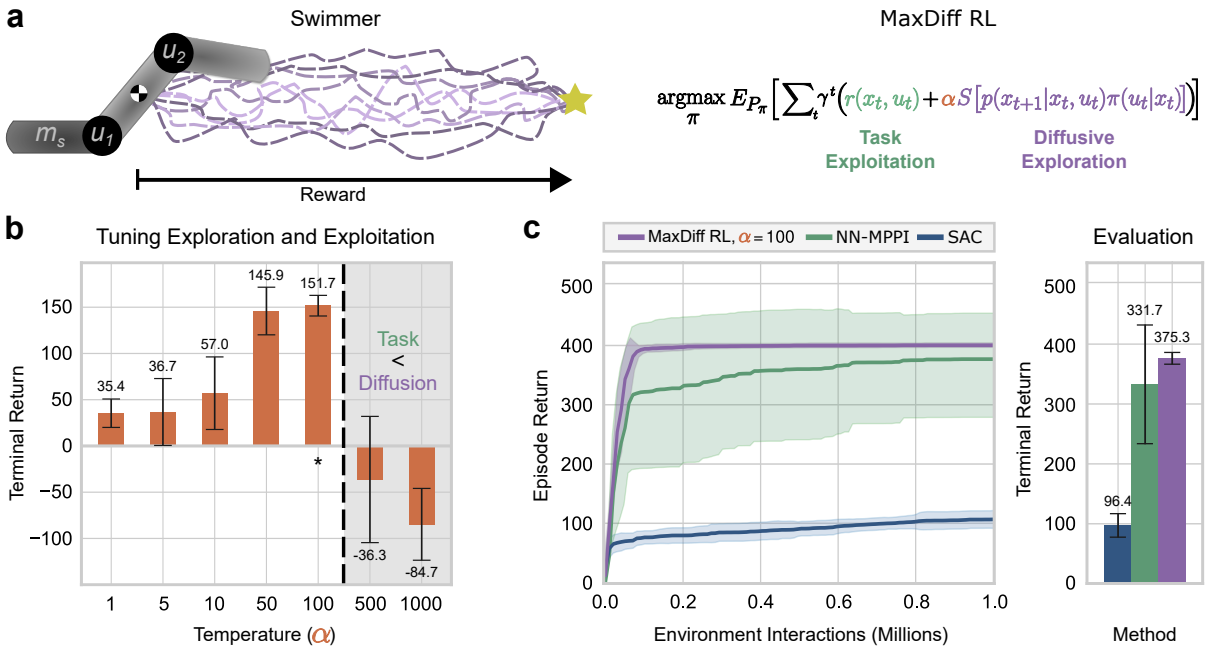


Figure 5.3. **Maximally diffusive RL agents are robust to random seeds and initializations**. (A), Illustration of MuJoCo swimmer environment (left panel). The swimmer has 2 degrees of actuation, u_1 and u_2, that rotate its limbs at the joints, with tail mass m_s and $m = 1$ for other limbs. MaxDiff RL synthesizes robust agent behavior by learning policies that balance task-capability and diffusive exploration (right panel). In practice this balance is tuned by a temperature-like parameter, α. (B), To explore the role that α plays in the performance of MaxDiff RL, we examine the terminal returns of swimmer agents (10 seeds each) across values of α with $m_s = 1$. Diffusive exploration leads to greater returns until a critical point (inset dotted line), after which the agent starts valuing diffusing more than accomplishing the task. (C), Using $\alpha = 100$, we compared MaxDiff RL against SAC and NN-MPPI with $m_s = 0.1$. We observe that MaxDiff RL outperforms comparisons on average with near-zero variability across random seeds, which is a formal property of MaxDiff RL agents. For all reward curves, the shaded regions correspond to the standard deviation from the mean across 10 seeds. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test.

While asymptotically there is no trade-off between maximally diffusive exploration and task exploitation, managing the relative balance between these two aims is important over finite time horizons, which we achieve with a temperature-like parameter, α. In practice, our entropy term plays a similar role as other exploration bonuses that reward agent curiosity or provide intrinsic motivation [185–187]. Unlike other bonuses, however, the role of path entropy can be interpreted through the lens of statistical mechanics. If α is set too high, the system's fluctuations can overpower the reward and break the agent's ergodicity in ways that resemble the physics of diffusion processes in potential fields [39]. Unfortunately, predicting where this critical α threshold lies is generally challenging due to its conceptual ties to the phenomenon of ergodicity-breaking in nonequilibrium processes [188].

Since ergodicity provides many of MaxDiff RL's desirable properties and guarantees, tuning the value of α is essential. In Fig. 5.3 and Movie S1[1], we explore the effect of tuning α on the learning performance of MaxDiff RL agents in MuJoCo's swimmer environment. The swimmer system is comprised of three rigid links of nominally equal mass, $m = 1$, with two degrees of actuation at the joints. The agent's objective is to swim as fast as possible within a fixed time interval, while being subjected to viscous drag forces (Fig. 5.3(A)). In Fig. 5.3(B), we vary α across multiple orders of magnitude and examine its impact on the terminal returns of MaxDiff RL swimmer agents. As we modulate the value of α from 1 to 100, we observe that diffusive exploration leads to greater returns. However, after $\alpha = 100$ we cross the critical threshold beyond which the strength of the system's diffusive exploration overpowers the reward (see inset dotted line in Fig 5.3(B)), thereby breaking

[1] `https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM2_ESM.mp4`

the ergodicity of our agents with respect to the underlying potential and performing poorly at the task—just as predicted by our theoretical framework.

Given a constant temperature of $\alpha = 100$ that preserves the swimmer's ergodicity, we compared the performance of MaxDiff RL to NN-MPPI and SAC across 10 seeds each. To ensure the task was solvable by all agents, we lowered the mass of the swimmer's third link (i.e., its tail) to $m_s = 0.1$. We find that while SAC struggles to succeed within a million environment interactions, NN-MPPI achieves good performance but with high variance across seeds. This is in stark contrast to MaxDiff RL, whose performance is near-identical and competitive across all random seeds (see Fig. 5.3(C) and Movie S2[2]). Hence, by decorrelating state transitions, our agent was able to exhibit robustness to seeds and environment randomization beyond what is typically possible in deep RL. Moreover, since our implementation of MaxDiff RL is identical to that of NN-MPPI, we can attribute any performance gains and added robustness to the properties of MaxDiff RL's theoretical framework.

Robustness to random seeds and environmental randomizations is a highly desirable feature of deep RL agents [153, 189, 190]. However, guaranteeing such robustness is challenging because it requires modeling the impact of neural representations on learning outcomes. Nonetheless, we can provide representation-agnostic guarantees through the probably approximately correct in Markov decision processes (PAC-MDP) learning framework [191, 192].

[2] https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM3_ESM.mp4

Definition 5.2. *An algorithm $\mathcal{A}$ is said to be PAC-MDP if, for any $\epsilon > 0$ and $\delta \in (0, 1)$, a policy π can be produced with poly$(|\mathcal{X}|, |\mathcal{U}|, 1/\epsilon, 1/\delta, 1/(1-\gamma))$ sample complexity that is at least ϵ-optimal with probability at least $1 - \delta$. In other words, if $\mathcal{A}$ satisfies*

$$\Pr\big(\mathcal{V}_{\pi^*}(x_1) - \mathcal{V}_{\pi}(x_1) \leq \epsilon\big) \geq 1 - \delta$$

with polynomial sample complexity for all $x_1 \in \mathcal{X}$, where

$$\mathcal{V}_{\pi}(x_t) = E_{p,\pi}\left[\sum_{n=1}^{\infty} \gamma^n r(x_{n+t}, u_{n+t}) \middle| x_t = x\right] \tag{5.13}$$

is the value function due to policy π under state-transition model p, and $\mathcal{V}_{\pi^*}(\cdot)$ is the optimal value function, then $\mathcal{A}$ is PAC-MDP.

Thus, an algorithm is PAC-MDP if it is capable of producing a policy that is at least ϵ-optimal at least $100 \times (1 - \delta)\%$ of the time for any valid choice of ϵ and δ. We note that this framework is representation-agnostic in the sense that, regardless of whether $\mathcal{A}$ involves any kind of neural network representation, any algorithm that satisfies Definition 5.2 is guaranteed to be at least ϵ-optimal. Under this framework, we can provide formal robustness guarantees.

Theorem 5.2. *(MaxDiff RL agents are robust to random seeds) If there exists a PAC-MDP algorithm $\mathcal{A}$ with policy π^{max} for the MaxDiff RL objective (Eq. 5.4), then the Markov chain induced by π^{max} is ergodic, and $\mathcal{A}$ will be asymptotically ϵ-optimal regardless of initialization.*

Proof. This theorem follows directly from the ergodicity of maximally diffusive trajectories (which we proved in Corollary 2.1.2 and Theorem 2.2), some basic facts

about MDPs [29], and the application of Birkhoff's ergodic theorem [193] onto our definition of PAC-MDP (Definition 5.2). First, since $\mathcal{A}$ is capable of producing an ϵ-optimal policy, π^{max}, we take $D_{KL}(p_{\pi^{max}}||p_{max}) \approx 0$ for some choice of ϵ, given that $p_{\pi^{max}}(x_{t+1}|x_t) = \int_{\mathcal{U}} p(x_{t+1}|x_t, u_t)\pi^{max}(u_t|x_t)du_t$. Then, it is well-known that any given policy in an MDP gives rise to a Markov chain on the state-space of the MDP [29]. Naturally, the properties of the policy-induced Markov chain depend on the properties of the resulting state transition kernel (e.g., $p_\pi(x_{t+1}|x_t)$).

Let $\{x_t\}_{t\in\mathbb{N}}$ be a Markov chain with state transition properties determined by $p_{\pi^{max}}(x_{t+1}|x_t)$. Because we know that $D_{KL}(p_{\pi^{max}}||p_{max}) \approx 0$, the Markov chain described by $p_{\pi^{max}}(x_{t+1}|x_t)$ is ergodic (per Theorem 2.2) with invariant measure ρ. To proceed further, we will now restate Birkhoff's well-known ergodic theorem [30, 193].

Theorem 5.3. *(Birkhoff's ergodic theorem) Let $\{x_t\}_{t\in\mathbb{N}}$ be an aperiodic and irreducible Markov process on a state space $\mathcal{X}$ with invariant measure ρ and let $f : \mathcal{X} \to \mathbb{R}$ be any measurable function with $E[|f(x)|] < \infty$. Then, one has*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(x_t) = E_{x_1\sim\rho}[f(x_1)] \tag{5.14}$$

almost surely.

In other words, Birkhoff's ergodic theorem states the the time-average of any function of an ergodic Markov chain is equal to its ensemble average.

Now, we return to the definition of PAC-MDP to slightly manipulate the expression:

$$\Pr\big(\mathcal{V}_{\pi^*}(x_1) - \mathcal{V}_{\pi^{max}}(x_1) \leq \epsilon\big) \geq 1 - \delta$$

$$E_{x_1 \sim \rho}\Big[\mathbf{1}\{\mathcal{V}_{\pi^*}(x_1) - \mathcal{V}_{\pi^{max}}(x_1) \leq \epsilon\}\Big] \geq 1 - \delta,$$

where $\mathbf{1}\{\cdot\}$ denotes an indicator function. In other words, to be PAC-MDP is equivalent to being at least ϵ-optimal on average at least $100 \times (1 - \delta)\%$ of episodes. To conclude our proof, note that

$$f(x_t) = \mathbf{1}\{\mathcal{V}_{\pi^*}(x_t) - \mathcal{V}_{\pi^{max}}(x_t) \leq \epsilon\}$$

is a bounded observable and, as a result, Birkhoff's theorem can be applied onto it.

Lastly, let $\{x_t\}_{t \in \mathbb{N}}$ and $\{x'_t\}_{t \in \mathbb{N}}$ both be ergodic Markov chains with identical transition kernels given by $p_{\pi^{max}}$, but with different initial conditions $x_1, x'_1 \in \mathcal{X}$. Then, since Birkhoff's ergodic theorem guarantees that the time-averages of observables from $\{x_t\}_{t \in \mathbb{N}}$ and $\{x'_t\}_{t \in \mathbb{N}}$ will converge to the same unique ensemble average over the invariant measure ρ (Theorem 5.3), the following is true:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} |f(x_t) - f(x'_t)| = 0$$

for any $x_1, x'_1 \in \mathcal{X}$ almost surely. This proves that any PAC-MDP algorithm is guaranteed to be robust to random seeds and environmental initializations if the underlying Markov chain induced by the policy is ergodic, which concludes our proof. □

Thus, since maximally diffusive agents are ergodic, any two arbitrary initializations will asymptotically achieve identical learning outcomes, which implies robustness to random

seeds and environmental stochasticity. Despite excluding neural representations from our analysis, Fig. 5.3(C) suggests that our guarantees hold empirically.

5.2.4. Zero-Shot Generalization Across Embodiments

When agents can find optimal policies, their dynamics become indistinguishable from an ergodic diffusion process. In doing so, the MaxDiff RL objective (see Eq. 5.5) reduces the influence of agent dynamics on performance. This suggests that successful MaxDiff RL policies may exhibit favorable generalization properties across agent embodiments. To explore this possibility, as well as the robustness of MaxDiff RL agents to variations in their neural representations, we devised a transfer experiment in the MuJoCo swimmer environment. We designed two variants of the swimmer: One with a heavy, less controllable tail of $m_s = 1$, and another with a light, more controllable tail of $m_s = 0.1$ (Fig. 5.4(A)). We trained two sets of representations for each algorithm. One set was trained with the light-tailed swimmer, and another set was trained with the heavy-tailed swimmer. Then, we deployed and evaluated each set of representations on both the swimmer variant that they observed during training, as well as its counterpart. Our experiment's outcomes are shown in Fig. 5.4(B,C), where the results are categorized as "baseline" if the trained and deployed swimmer variants match, or "transfer" if they were swapped. The baseline experiments validate other results shown throughout the chapter: All algorithms benefit from working with a more controllable system whose dynamics induce weaker temporal correlations (see Fig. 5.4(B) and Movie S2[3]). However, as MaxDiff RL is the only approach

[3] https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM3_ESM.mp4

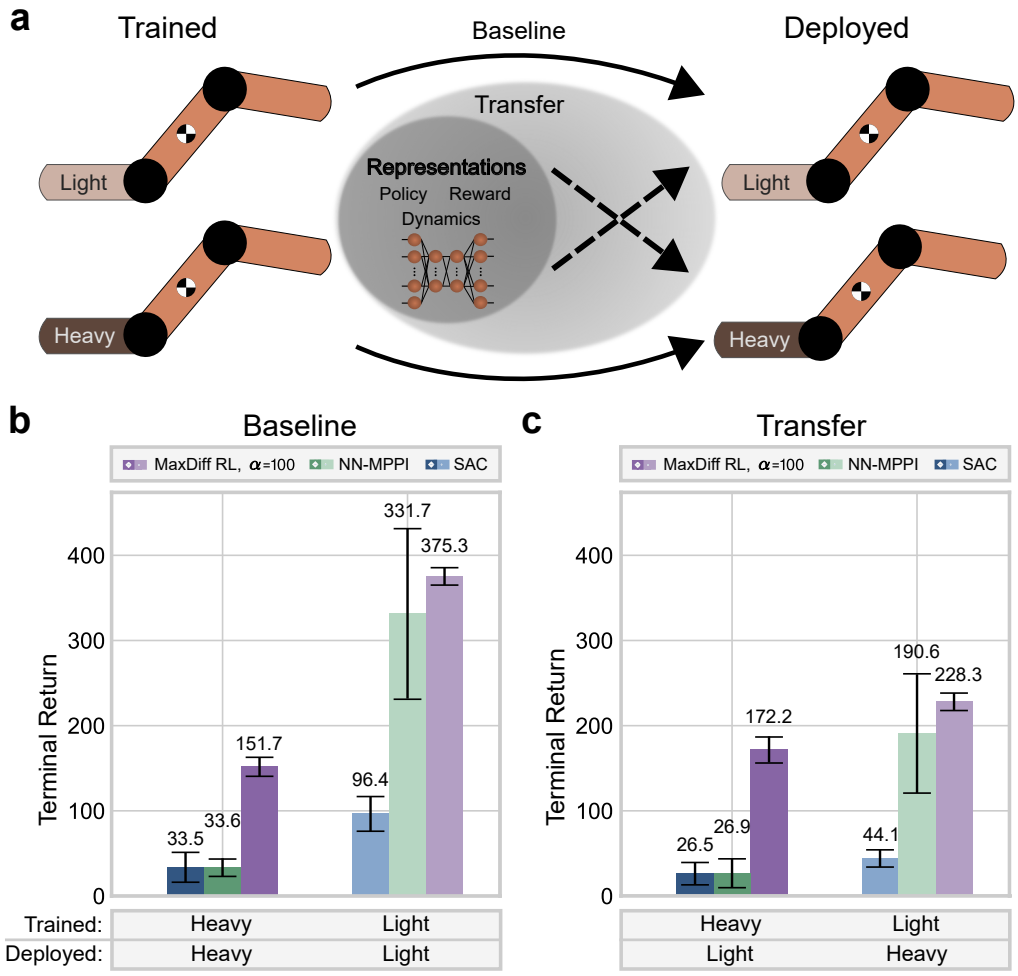


Figure 5.4. **Trained system embodiment determines deployed system performance**. (A), Two variants of the MuJoCo swimmer environment: One with $m_s = 1$ and one with $m_s = 0.1$. As a baseline, we deploy learned representations on the same swimmer variant trained on. Then, we carry out a transfer experiment where the trained and deployed swimmer variants are swapped. (B), Baseline experiments confirm previous results: All algorithms benefit from a more controllable swimmer. (C), Both NN-MPPI and SAC performance degrades when deployed on a more controllable system than was trained on, which is undesirable. In contrast, MaxDiff RL benefits from the "Heavy-to-Light" transfer and we also observe that MaxDiff RL performance further increases in the "Light-to-Heavy" transfer experiment. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean with $n = 1000$ (100 evaluations over 10 seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test.

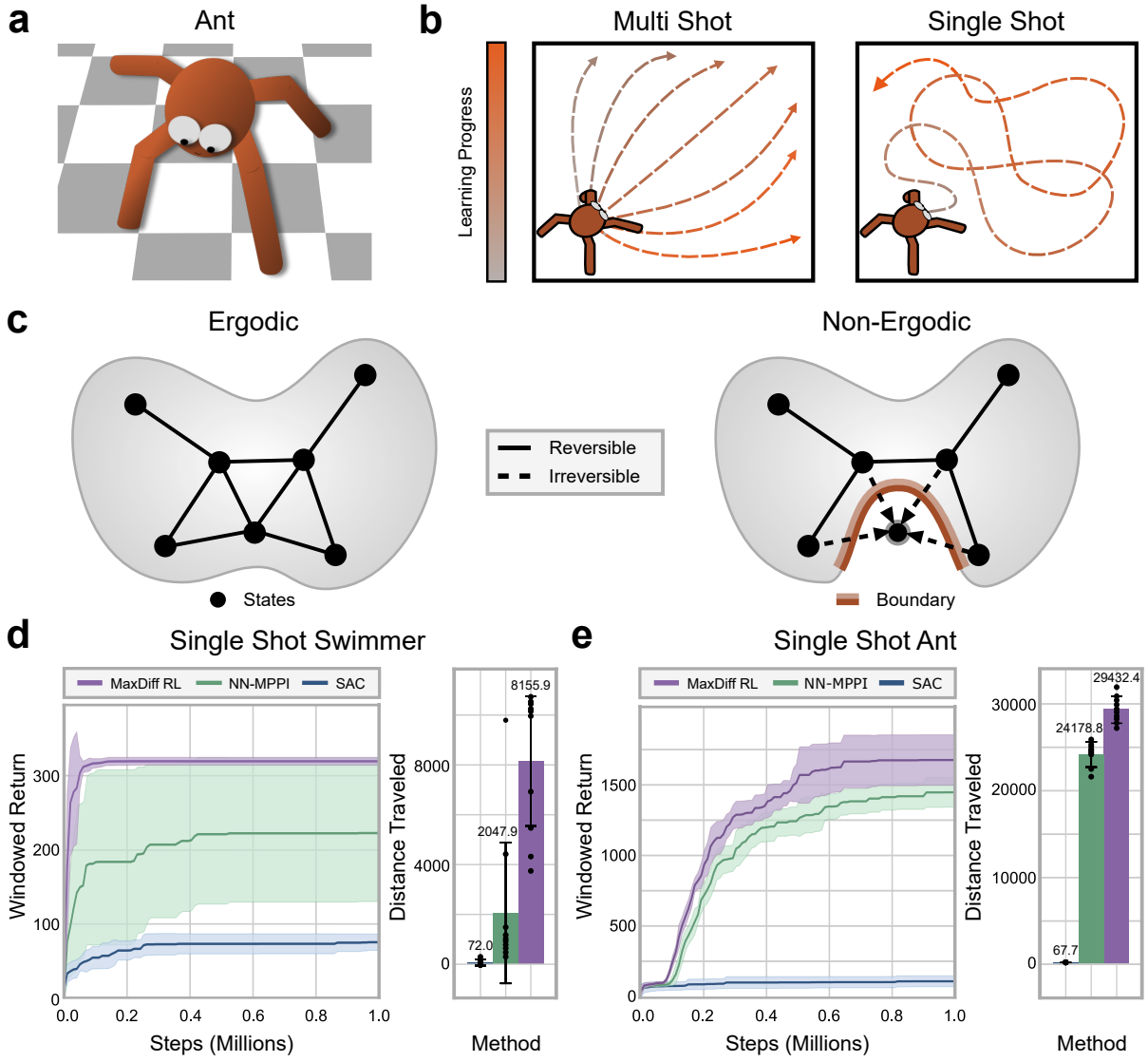
taking temporal correlations into account, it is the only method that remains task-capable with a heavy-tailed swimmer.

For the transfer experiments, all of the learned neural representations of the reward function, control policy, and agent dynamics were deployed on the swimmer variant that was not seen during training (Fig. 5.4(A)). First, we note that for both NN-MPPI and SAC representation transfer leads to degrading performance across the board. This is the case even when the swimmer variant they were deployed onto was more controllable, which is counterintuitive and undesirable behavior. In contrast, our MaxDiff RL agents can actually benefit and improve their performance when deployed on the more controllable swimmer variant, as desired (see "Heavy-to-Light" transfer in Fig. 5.4(C) and Movie S3[4]). In other words, as the task becomes easier in this way, we can expect the performance of MaxDiff RL agents to improve.

A more surprising result is the performance increase in MaxDiff RL agents between the baseline heavy-tailed swimmer and the "Light-to-Heavy" transfer swimmer (Fig. 5.4(c) and Movie S3[5]). We found that training with a more controllable swimmer increased the performance of agents when deployed on a heavy-tailed swimmer, showing that system controllability during training matters more to overall performance than the particular embodiment of the deployed system. This kind of zero-shot generalization [194] from an easier task to a more challenging task is reminiscent of results seen in RL agents trained via curriculum learning [195], as well as of the incremental learning dynamics of biological systems during motor skill acquisition [196]. However, here it emerges spontaneously from

[4] https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM4_ESM.mp4

[5] https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM4_ESM.mp4



the properties of MaxDiff RL agents. In part, this occurs because greater controllability leads to improved exploration, which increases the diversity of data observed during training.

Figure 5.5. **Maximally diffusive RL agents are capable of single-shot learning.** (A), Illustration of MuJoCo ant environment. (B), Typical algorithms learn across many different initializations and deployments of an agent, which is known as multi-shot learning. In contrast, single-shot learning insists on a single task attempt, which requires learning through continuous deployments. Here, we prove that MaxDiff RL agents are equivalently capable of single-shot and multi-shot learning in a broad variety of settings. (C), Single-shot learning depends on the ability to generate data samples ergodically, which MaxDiff RL guarantees when there are no irreversible state transitions in the environment. (D), Single-shot learning in the swimmer MuJoCo environment. We find that MaxDiff RL achieves robust performance comparable to its multi-shot counterpart. (E), In contrast to the swimmer, the MuJoCo ant environment contains irreversible state transitions (e.g., flipping upside down) preventing ergodic trajectories. Nonetheless, MaxDiff RL remains state-of-the-art in single-shot learning. Note that we report returns over a window of 1000 steps in analogy to our multi-shot results, where episodes consist of 1000 environment interactions. For all reward curves, the shaded regions correspond to the standard deviation from the mean across 10 seeds. For all bar charts, data are presented as mean values above each error bar, where each error bar represents the standard deviation from the mean and the data distribution is plotted directly ($n = 10$ seeds for each condition). All differences between MaxDiff RL and comparisons within this figure are statistically significant with $P < 0.001$ using an unpaired two-sided Welch's t-test.

5.2.5. Single-Shot Learning in Ergodic Agents

When agents are deployed in the real world, they face situations at test time that were never encountered during training. Since exhaustively accounting for every possible scenario is infeasible, agents capable of real-time adaptation and learning during individual deployments are desirable [154]. Most RL methods excel at episodic multi-shot learning over the course of several deployments (Fig. 5.5(B)), where randomized instantiations of a given task and environment passively provide a kind of variability that is essential to the learning process [197]. However, episodic problems of this kind are very rare in

real-world applications. For this reason, there is a need for methods that allow agents to perform a task successfully within a single trial—or, in other words, for methods that enable single-shot learning.

Single-shot learning concerns learning in non-episodic environments over the course of a single task attempt, similar to the "single-life" RL setting considered in [198]. Despite the challenges associated with studying the behavior of agents based on neural network representations, the ergodic properties of MaxDiff RL enables one to provide representation-agnostic guarantees on the feasibility of single-shot learning through the PAC-MDP learning framework.

Theorem 5.4. *(MaxDiff RL agents can learn in single-shot deployments) If there exists a PAC-MDP algorithm $\mathcal{A}$ with policy π^{max} for the MaxDiff RL objective (Eq. 5.4), then the Markov chain induced by π^{max} is ergodic, and any individual initialization of $\mathcal{A}$ will asymptotically satisfy the same ϵ-optimality as an ensemble of initializations.*

Proof. The proof of this theorem is simple given the proof to Theorem 5.2. Once again, let $\{x_t\}_{t\in\mathbb{N}}$ be an ergodic Markov chain with state transition statistics given by $p_{\pi^{max}}$ and let

$$f(x_t) = \mathbf{1}\{\mathcal{V}_{\pi^*}(x_t) - \mathcal{V}_{\pi^{max}}(x_t) \leq \epsilon\}$$

be an observable. Then, through a straightforward application of Birkhoff's theorem (Theorem 5.3) we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{\mathcal{V}_{\pi^*}(x_t) - \mathcal{V}_{\pi^{max}}(x_t) \leq \epsilon\} = E_{x_1\sim\rho}[\mathbf{1}\{\mathcal{V}_{\pi^*}(x_1) - \mathcal{V}_{\pi^{max}}(x_1) \leq \epsilon\}],$$

which proves that any individual initial condition will satisfy the ensemble average. In turn, we have

$$\Pr\big(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{max}}(x_0) \leq \epsilon\big) \geq 1-\delta \implies \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbf{1}\{\mathcal{V}_{\pi^*}(x_t) - \mathcal{V}_{\pi^{max}}(x_t) \leq \epsilon\} \geq 1-\delta$$

almost surely, which proves that an algorithm that is PAC-MDP during multi-shot (episodic) learning is guaranteed to be PAC-MDP during single-shot (non-episodic) learning if the underlying Markov chain induced by the policy is ergodic, which concludes our proof. □

Thus, any MaxDiff RL agent capable of solving a task in a multi-shot fashion (in the PAC-MDP sense) is capable of solving the same task in a single-shot fashion. Since any two MaxDiff RL agents will asymptotically achieve identical learning outcomes, any individual MaxDiff RL agent will also achieve identical learning outcomes as an ensemble. We note that this proof also clarifies why ergodic sampling along continuous Markovian trajectories is the best possible alternative to *i.i.d.* sampling—that is, because Birkhoff's theorem guarantees that observables computed from these correlated experiences will be (asymptotically) interchangeable from an *i.i.d.* computation of same observables. Since ergodicity is central to this proof, we expect its guarantees to fail when ergodicity is broken by either the agent or the environment.

Figure 5.5 demonstrates the single-shot learning capabilities of MaxDiff RL agents, and explores what happens when ergodicity is broken by the topological properties of the environment. Here, we examine both the MuJoCo swimmer and ant environments (Fig. 5.5(A)). The primary difference between these two environments is the existence of irreversible state transitions that can violate the ergodicity requirement of our single-shot learning guarantees topologically (Fig. 5.5(C)), which have been previously referred to as

"sink states" in the literature [197]. Unlike the swimmer, the ant is capable of transitioning into such states by flipping upside down, thereby breaking ergodicity. Irreversible state transitions are common in real-world applications because they can arise as a result of unsafe behavior, such as a robot breaking or malfunctioning during learning. While such transitions can be prevented in principle through the use of safety-preserving methods [47, 199, 200], we omit their implementation to illustrate our point. As expected, the MaxDiff RL single-shot swimmer is capable of learning in continuous deployments (see Fig. 5.5(D) and Movie S4[6]), retaining the same robustness of its multi-shot counterpart in Fig. 5.3(C), and achieving similar task performance. Despite ergodicity-breaking in the single-shot ant environment, MaxDiff RL still leads to improved outcomes over NN-MPPI and SAC, as in Fig. 5.5(E), where we plot the final distance traveled to ensure that no reward hacking took place. However, the loss of ergodicity leads to an increase in the variance of single-shot MaxDiff RL agent performance, which we expect as a result of our robustness guarantees no longer holding.

5.3. Discussion

Throughout this chapter, we have highlighted the ways in which RL is fragile to temporal correlations intrinsic to many sequential decision-making processes. We introduced a framework based on the statistical mechanics of ergodic processes to overcome these limitations, which we term MaxDiff RL. Our framework offers a generalization of the current state-of-the-art in RL and addresses many foundational issues holding back the field: The ergodicity of MaxDiff RL agents enables data acquisition that is indistinguishable from

[6] https://static-content.springer.com/esm/art%3A10.1038%2Fs42256-024-00829-3/MediaObjects/42256_2024_829_MOESM5_ESM.mp4

i.i.d. sampling, performance that is robust to seeds, and single-shot learning. Through its roots in statistical physics, our work forms a starting point for a more scientific study of embodied RL—one in which falsifiable predictions can be made about agent properties and their performance.

However, much more work at the nexus of physics, learning, and control remains to be done in pursuit of this goal. For one, approaches grounded in statistical physics for tuning or annealing temperature-like parameters during learning will be necessary to achieve effective exploration without sacrificing agent performance [201]. Additionally, control techniques capable of enforcing ergodicity in the face of environmental irreversibility are needed to guarantee desirable agent properties like robustness to random seeds in complex problem settings [181]. Beyond RL, our work also has the potential to open new lines of interdisciplinary inquiry in areas such as biological learning and animal behavior. For example, the importance of ergodicity to animal behaviors like foraging and tracking has been extensively studied [202]. As such, our work presents an avenue for studying these behaviors within an RL framework that is sensitive to physical embodiment. For biological motor learning, our findings also suggest that controllability may be a promising frame of reference for studying motor skill acquisition [203]. More broadly, our work is particularly well-suited to applications in soft matter systems where the impact of correlations may in fact be impossible to avoid entirely [7]. Taken together, our results present a major advance towards transparently understanding and reliably synthesizing complex behavior in embodied decision-making agents, which will be crucial to the long-term viability of deep RL as a field.

The culmination of this chapter concludes our motivation of *robot thermodynamics* across many different areas of the field of robotics. By taking into account the path continuity of RL agents, we were able to develop a framework that overcomes violations of the *i.i.d.* property, provides robustness guarantees, and makes learning in single-shot deployments possible—all while achieving state-of-the-art performance at benchmarks. In the final chapter of this thesis, we will present an outlook towards potential future directions for this body of work, and remark upon other dimensions of this work being currently explored.

CHAPTER 6

Conclusions

This dissertation has presented a novel framework for the design, analysis, and control of embodied autonomy, drawing inspiration from the principles of statistical mechanics and thermodynamics. By embracing uncertainty and nondeterminism in the modelling and control of complex systems, we developed a flexible set of tools that enable us to reason about agent behavior in terms of path distributions. Path distributions provided us with a means of parsimoniously reasoning about agent embodiment and decision-making as two sides of the same coin—as elements that shape the structure of an agent's path distribution. As illustrated throughout the chapters in this manuscript, this approach, which we term "robot thermodynamics," produced significant advances to the state-of-the-art in robotics across many areas of the field.

In Ch. 2, we laid the mathematical foundations of robot thermodynamics, introducing the concept of path distributions and demonstrating how they can be inferred and manipulated using the principle of maximum caliber. We derived a novel, unpublished result on the steady-state occupancy statistics of a broad class of stochastic processes, recovering the low-rattling selection principle. Additionally, we presented an original derivation of Pontryagin's maximum principle from the principle of maximum caliber, establishing connections between KL-control and stochastic optimal control in the process. In Ch. 3, we explored the application of our framework to the prediction of self-organization in active and robotic matter, introducing and experimentally validating a Boltzmann-like

principle for predicting the steady-state behavior of complex systems. Chapter 4 delved into the design of emergent behaviors in robotic microsystems, analyzing the complex dynamics of active colloidal microparticles and demonstrating a novel thermodynamic mechanism for asymmetry-induced order. Finally, in Ch. 5, we introduced maximum diffusion reinforcement learning (MaxDiff RL), a framework derived from the principles of robot thermodynamics that provably decorrelates agent experiences and enables single-shot learning in continuous deployments.

The results presented in this dissertation represent a significant step forward in the quest for robust, adaptable, and life-like robotic autonomy. In order to do so, we stepped away from the deterministic ideals of precision engineering and embraced the uncertain and unpredictable nature of real world. Path distributions and statistical physics present a promising avenue for rigorously examining the way in which agent properties and their decision-making can influence their behavior and environments. We believe the principles outlined in this thesis present a promising path towards an future in which an understanding of embodiment leads to the development of more safe, reliable, and adaptable autonomous agents.

References

1. Harsch, V. Otto von Guericke (1602–1686) and His Pioneering Vacuum Experiments. *Aviation, Space, and Environmental Medicine* **78,** 1075–1077. ISSN: 0095-6562 (2007).
2. Carnot, S. *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance* in *Annales scientifiques de l'École normale Supérieure* **1** (1872), 393–457.
3. Clausius, R. *Die Mechanische Wärmetheorie* (Vieweg, 1887).
4. Wrigley, E. A. Energy and the English industrial revolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371,** 20110568 (2013).
5. Gasparetto, A., Scalera, L., *et al.* A brief history of industrial robotics in the 20th century. *Advances in Historical Studies* **8,** 24–35 (2019).
6. Chvykov, P. *et al.* Low rattling: A predictive principle for self-organization in active collectives. *Science* **371,** 90–95 (2021).
7. Berrueta, T. A., Pinosky, A. & Murphey, T. D. Maximum diffusion reinforcement learning. *Nature Machine Intelligence* **6,** 504–514. ISSN: 2522-5839 (2024).
8. Savoie, W. *Effect of shape and particle coordination on collective dynamics of granular matter* PhD thesis (Georgia Institute of Technology, 2019).
9. Yang, J. F. *et al.* Emergent microrobotic oscillators via asymmetry-induced order. *Nature Communications* **13,** 5734 (2022).

10. Prigogine, I. & Stengers, I. *The End of Certainty* ISBN: 9780684837055 (Free Press, 1997).

11. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106,** 620–630 (1957).

12. Gallager, R. G. *Stochastic Processes: Theory for Applications* ISBN: 9781107039759 (Cambridge University Press, 2013).

13. Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications* ISBN: 9783642143946 (Springer Berlin Heidelberg, 2010).

14. Feynman, R. P., Hibbs, A. R. & Styer, D. F. *Quantum Mechanics and Path Integrals* (Dover Publications, 2010).

15. Jaynes, E. T. The Minimum Entropy Production Principle. *Annual Review of Physical Chemistry* **31,** 579–601. ISSN: 1545-1593 (1980).

16. Dixit, P. D. *et al.* Perspective: Maximum caliber is a general variational principle for dynamical systems. *The Journal of Chemical Physics* **148,** 010901 (2018).

17. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 379–423. ISSN: 0005-8580 (1948).

18. Haarnoja, T., Tang, H., Abbeel, P. & Levine, S. Reinforcement Learning with Deep Energy-Based Policies. *Proceedings of the International Conference on Machine Learning (ICML)* **70,** 1352–1361 (2017).

19. Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *Proceedings of the International Conference on Machine Learning (ICML)* **80,** 1861–1870 (2018).

20. So, O., Wang, Z. & Theodorou, E. A. *Maximum Entropy Differential Dynamic Programming* in *2022 IEEE International Conference on Robotics and Automation (ICRA)* (2022), 3422–3428.

21. Sontag, E. D. *Mathematical Control Theory: Deterministic Finite Dimensional Systems* ISBN: 9781461205777 (Springer, 2013).

22. Hespanha, J. P. *Linear Systems Theory: Second Edition* ISBN: 9780691179575 (Princeton University Press, 2018).

23. Sontag, E. D. in *Mathematical System Theory: The Influence of R. E. Kalman* 453–462 (Springer, 1991). ISBN: 9783662085462.

24. Cortesi, F. L., Summers, T. H. & Lygeros, J. *Submodularity of energy related controllability metrics* in *2014 IEEE Conference on Decision and Control (CDC)* (2014), 2883–2888.

25. Summers, T. H., Cortesi, F. L. & Lygeros, J. On Submodularity and Controllability in Complex Dynamical Networks. *IEEE Transactions on Control of Network Systems* **3,** 91–101 (2016).

26. Kardar, M. *Statistical Physics of Fields* (Cambridge University Press, 2007).

27. Kashima, K. Noise Response Data Reveal Novel Controllability Gramian for Nonlinear Network Dynamics. *Scientific Reports* **6,** 27300 (2016).

28. Risken, H. in *The Fokker-Planck Equation* 63–95 (Springer, 1996). ISBN: 978-3-642-96807-5.

29. Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* ISBN: 9781118625873 (Wiley, 2014).

30. Moore, C. C. Ergodic theorem, ergodic theory, and statistical mechanics. *Proceedings of the National Academy of Sciences* **112,** 1907–1911 (2015).

31. Michalet, X. & Berglund, A. J. Optimal diffusion coefficient estimation in single-particle tracking. *Physical Review E* **85,** 061916 (2012).

32. Boyer, D., Dean, D. S., Mejía-Monasterio, C. & Oshanin, G. Optimal estimates of the diffusion coefficient of a single Brownian trajectory. *Physical Review E* **85,** 031136 (2012).

33. Loper, J. Uniform Ergodicity for Brownian Motion in a Bounded Convex Set. *Journal of Theoretical Probability* **33,** 22–35 (2020).

34. Prigogine, I. Time, Structure, and Fluctuations. *Science* **201,** 777–785 (1978).

35. England, J. L. Dissipative adaptation in driven self-assembly. *Nature Nanotechnology* **10,** 919–923. ISSN: 1748-3395 (2015).

36. Rawlik, K., Toussaint, M. & Vijayakumar, S. On Stochastic Optimal Control and Reinforcement Learning by Approximate Inference. *Proceedings of Robotics: Science and Systems (RSS),* 353–361 (2012).

37. Miller, L. M., Silverman, Y., MacIver, M. A. & Murphey, T. D. Ergodic Exploration of Distributed Information. *IEEE Transactions on Robotics* **32,** 36–52 (2016).

38. Mavrommati, A., Tzorakoleftherakis, E., Abraham, I. & Murphey, T. D. Real-Time Area Coverage and Target Localization Using Receding-Horizon Ergodic Exploration. *IEEE Transactions on Robotics* **34,** 62–80 (2018).

39. Wang, X., Deng, W. & Chen, Y. Ergodic properties of heterogeneous diffusion processes in a potential well. *The Journal of Chemical Physics* **150,** 164121 (2019).

40. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* ISBN: 0387310738 (Springer, 2006).

41. Ziebart, B. D., Maas, A. L., Bagnell, J. A. & Dey, A. K. Maximum entropy inverse reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* **8,** 1433–1438 (2008).

42. Eysenbach, B. & Levine, S. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. *Proceedings of the International Conference on Learning Representations (ICLR)* (2022).

43. Scholz, C., Jahanshahi, S., Ldov, A. & Löwen, H. Inertial delay of self-propelled particles. *Nature Communications* **9,** 5156 (2018).

44. Srinivasan, M. & Ruina, A. Computer optimization of a minimal biped model discovers walking and running. *Nature* **439,** 72–75 (2006).

45. Ansari, A. R. & Murphey, T. D. Sequential Action Control: Closed-Form Optimal Control for Nonlinear and Nonsmooth Systems. *IEEE Transactions on Robotics* **32,** 1196–1214 (2016).

46. Williams, G. *et al.* Information theoretic MPC for model-based reinforcement learning. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA),* 1714–1721 (2017).

47. Ames, A., Grizzle, J. & Tabuada, P. *Control Barrier Function based Quadratic Programs with Application to Adaptive Cruise Control* in *2014 IEEE Conference on Decision and Control (CDC)* (2014).

48. Kliemann, W. Recurrence and invariant measures for degenerate diffusions. *Annals of Probability* **15,** 690–707 (1987).

49. Bou-Rabee, N. & Owhadi, H. Ergodicity of Langevin Processes with Degenerate Diffusion in Momentums. *International Journal of Pure and Applied Mathematics* **45,** 475–490 (2008).

50. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302. ISSN: 1476-4687 (2006).

51. Kardar, M. *Statistical Physics of Particles* (Cambridge University Press, 2007).

52. Corte, L., Chaikin, P., Gollub, J. P. & Pine, D. Random organization in periodically driven systems. *Nature Physics* **4,** 420–424 (2008).

53. Grosberg, A. & Joanny, J.-F. Nonequilibrium statistical mechanics of mixtures of particles in contact with different thermostats. *Physical Review E* **92,** 032118 (2015).

54. Sumino, Y. *et al.* Large-scale vortex lattice emerging from collectively moving microtubules. *Nature* **483,** 448–452. ISSN: 1476-4687 (2012).

55. Ramaswamy, S. The mechanics and statistics of active matter. *Annu. Rev. Condens. Matter Phys.* **1,** 323–345 (2010).

56. Bertini, L., De Sole, A., Gabrielli, D., Jona-Lasinio, G. & Landim, C. Macroscopic fluctuation theory. *Reviews of Modern Physics* **87,** 593 (2015).

57. Paoluzzi, M., Maggi, C., Marini Bettolo Marconi, U. & Gnan, N. Critical phenomena in active matter. *Phys. Rev. E* **94,** 052602 (5 2016).

58. Speck, T. Stochastic thermodynamics for active matter. *EPL (Europhysics Letters)* **114,** 30006 (2016).

59. Chvykov, P. & England, J. Least-rattling feedback from strong time-scale separation. *Phys. Rev. E* **97,** 032115 (3 2018).

60. Cates, M. E. & Tailleur, J. Motility-Induced Phase Separation. *Annual Review of Condensed Matter Physics* **6,** 219–244 (2015).

61. Aguilar, J. *et al.* Collective clog control: Optimizing traffic flow in confined biological and robophysical excavation. *Science* **361,** 672–677. ISSN: 0036-8075 (2018).

62. Rubenstein, M., Cornejo, A. & Nagpal, R. Programmable self-assembly in a thousand-robot swarm. *Science* **345,** 795–799 (2014).

63. Werfel, J., Petersen, K. & Nagpal, R. Designing Collective Behavior in a Termite-Inspired Robot Construction Team. *Science* **343,** 754–758. ISSN: 0036-8075 (2014).

64. Li, S. *et al.* Particle robotics based on statistical mechanics of loosely coupled components. *Nature* **567,** 361–365. ISSN: 1476-4687 (2019).

65. Vásárhelyi, G. *et al.* Optimized flocking of autonomous drones in confined environments. *Science Robotics* **3** (2018).

66. Mayya, S., Notomista, G., Shell, D., Hutchinson, S. & Egerstedt, M. Non-Uniform Robot Densities in Vibration Driven Swarms Using Phase Separation Theory. *IEEE International Conference on Intelligent Robots and Systems (IROS)* (2019).

67. Duhr, S. & Braun, D. Why molecules move along a temperature gradient. *Proceedings of the National Academy of Sciences* **103,** 19678–19682. ISSN: 0027-8424 (2006).

68. Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (Elsevier, 1992).

69. Landauer, R. Statistical physics of machinery: forgotten middle-ground. *Physica A: Statistical Mechanics and its Applications* **194,** 551–562. ISSN: 0378-4371 (1993).

70. Landauer, R. Inadequacy of entropy and entropy derivatives in characterizing the steady state. *Phys. Rev. A* **12,** 636–638 (2 1975).

71. Chvykov, P. *On typicality and adaptation in driven dynamical systems* PhD thesis (Massachusetts Institute of Technology, 2019).

72. Calvert, J. & Randall, D. *A local-global principle for nonequilibrium steady states* 2024. arXiv: 2311.10957 [math.PR]. https://arxiv.org/abs/2311.10957.

73. Redner, G. S., Hagan, M. F. & Baskaran, A. Structure and dynamics of a phase-separating active colloidal fluid. *Phys. Rev. Lett.* **110,** 055701 (2013).

74. Palacci, J., Sacanna, S., Steinberg, A. P., Pine, D. J. & Chaikin, P. M. Living Crystals of Light-Activated Colloidal Surfers. *Science* **339,** 936–940. ISSN: 0036-8075 (2013).

75. Savoie, W. *et al.* A robot made of robots: Emergent transport and control of a smarticle ensemble. *Science Robotics* **4** (2019).

76. Kedia, H., Pan, D., Slotine, J.-J. & England, J. L. Drive-specific selection in multistable mechanical networks. *The Journal of Chemical Physics* **159,** 214106. ISSN: 0021-9606 (2023).

77. Epstein, T. & Fineberg, J. Control of Spatiotemporal Disorder in Parametrically Excited Surface Waves. *Phys. Rev. Lett.* **92,** 244502 (24 2004).

78. Karani, H., Pradillo, G. E. & Vlahovska, P. M. Tuning the Random Walk of Active Colloids: From Individual Run-and-Tumble to Dynamic Clustering. *Phys. Rev. Lett.* **123,** 208002 (20 2019).

79. Goldman, D. I., Shattuck, M., Moon, S. J., Swift, J. & Swinney, H. L. Lattice dynamics and melting of a nonequilibrium pattern. *Phys. Rev. Lett.* **90,** 104302 (2003).

80. Sigmund, O. *Systematic design of metamaterials by topology optimization* in *IUTAM Symposium on Modelling Nanomaterials and Nanosystems* (2009), 151–159.

81. Buzsáki, G. & Draguhn, A. Neuronal Oscillations in Cortical Networks. *Science* **304,** 1926–1929 (2004).

82. Kruse, K. & Jülicher, F. Oscillations in Cell Biology. *Current Opinion in Cell Biology* **17,** 20–26. ISSN: 0955-0674 (2005).

83. Katz, P. S. Evolution of central pattern generators and rhythmic behaviours. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371,** 20150057 (2016).

84. Minguet-Parramona, C. *et al.* An optimal frequency in Ca^{2+} oscillations for stomatal closure is an emergent property of ion transport in guard cells. *Plant Physiology* **170,** 33–42. ISSN: 15322548 (2016).

85. Hoyt, D. F. & Taylor, C. R. Gait and the energetics of locomotion in horses. *Nature* **292,** 239–240. ISSN: 1476-4687 (1981).

86. Yao, X. S. & Maleki, L. Optoelectronic microwave oscillator. *Journal of the Optical Society of America* **13,** 1725–1735 (1996).

87. White, T. J. *et al.* A high frequency photodriven polymer oscillator. *Soft Matter* **4,** 1796–1798 (9 2008).

88. Lagzi, I., Kowalczyk, B., Wang, D. & Grzybowski, B. A. Nanoparticle oscillations and fronts. *Angewandte Chemie* **49,** 8616–8619 (2010).

89. Gardi, G., Ceron, S., Wang, W., Petersen, K. & Sitti, M. Microrobot collectives with reconfigurable morphologies, behaviors, and functions. *Nature Communications* **13,** 2239. ISSN: 2041-1723 (2022).

90. Jenkins, A. Self-oscillation. *Physics Reports* **525,** 167–222. ISSN: 0370-1573 (2013).

91. Hua, M. *et al.* Swaying gel: Chemo-mechanical self-oscillation based on dynamic buckling. *Matter* **4,** 1029–1041. ISSN: 2590-2385 (2021).

92. He, X. *et al.* Synthetic homeostatic materials with chemo-mechano-chemical self-regulation. *Nature* **487,** 214–218 (2012).

93. Grzybowski, B. A. & Huck, W. T. The nanotechnology of life-inspired systems. *Nature Nanotechnology* **11,** 585–592. ISSN: 17483395 (2016).

94. Akbar, F. *et al.* Self-sufficient self-oscillating microsystem driven by low power at low Reynolds numbers. *Science Advances* **7,** eabj0767 (2021).

95. Shen, B. & Kang, S. H. Designing self-oscillating matter. *Matter* **4,** 766–769. ISSN: 2590-2385 (2021).

96. Hudson, J. & Mankin, J. Chaos in the Belousov–Zhabotinskii reaction. *The Journal of Chemical Physics* **74,** 6171–6177 (1981).

97. Maeda, S., Hara, Y., Sakai, T., Yoshida, R. & Hashimoto, S. Self-Walking Gel. *Advanced Materials* **19,** 3480–3484 (2007).

98. Altemose, A. *et al.* Chemically controlled spatiotemporal oscillations of colloidal assemblies. *Angewandte Chemie International Edition* **56,** 7817–7821 (2017).

99. Zhou, C. *et al.* Coordinating an Ensemble of Chemical Micromotors via Spontaneous Synchronization. *ACS Nano* **14,** 5360–5370. ISSN: 1936–0851 (2020).

100. Yoshida, R. Self-Oscillating Gels Driven by the Belousov–Zhabotinsky Reaction as Novel Smart Materials. *Advanced Materials* **22,** 3463–3483 (2010).

101. Onoda, M., Ueki, T., Tamate, R., Shibayama, M. & Yoshida, R. Amoeba-like self-oscillating polymeric fluids with autonomous sol-gel transition. *Nature Communications* **8,** 15862. ISSN: 2041-1723 (2017).

102. Zhao, Y. *et al.* Soft phototactic swimmer based on self-sustained hydrogel oscillator. *Science Robotics* **4,** eafax7112 (2019).

103. Horváth, J., Szalai, I., Boissonade, J. & De Kepper, P. Oscillatory dynamics induced in a responsive gel by a non-oscillatory chemical reaction: experimental evidence. *Soft Matter* **7,** 8462–8472 (18 2011).

104. Shin, B. *et al.* Hygrobot: A self-locomotive ratcheted actuator powered by environmental humidity. *Science Robotics* **3,** eaar2629 (2018).

105. Funaki, T. *et al. Miniaturized 3D Functional Interposer Using Bumpless Chip-on-Wafer (COW) Integration with Capacitors* in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (2021), 185–190.

106. Molnar, A. C. *et al. Nanoliter-scale autonomous electronics: Advances, challenges, and opportunities* in *2021 IEEE Custom Integrated Circuits Conference (CICC)* (IEEE, 2021), 1–6. ISBN: 978-1-7281-7581-2.

107. Funke, D. A. *et al.* Ultra low-power,-area and-frequency CMOS thyristor based oscillator for autonomous microsystems. *Analog Integrated Circuits and Signal Processing* **89,** 347–356 (2016).

108. Hwang, C., Bibyk, S., Ismail, M. & Lohiser, B. A very low frequency, micropower, low voltage CMOS oscillator for noncardiac pacemakers. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **42,** 962–966 (1995).

109. Galea, F., Casha, O., Grech, I., Gatt, E. & Micallef, J. *Ultra Low Frequency Low Power CMOS Oscillators for MPPT and Switch Mode Power Supplies* in *IEEE Conference on Ph.D. Research in Microelectronics and Electronics* (2018), 121–124.

110. Wang, Y. *et al.* Bipolar Electrochemical Mechanism for the Propulsion of Catalytic Nanomotors in Hydrogen Peroxide Solutions. *Langmuir* **22,** 10451–10456 (2006).

111. Paxton, W. F. *et al.* Catalytically Induced Electrokinetics for Motors and Micropumps. *Journal of the American Chemical Society* **128,** 14881–14888 (2006).

112. Brooks, A. M. *et al.* Shape-directed rotation of homogeneous micromotors via catalytic self-electrophoresis. *Nature Communications* **10,** 495 (2019).

113. Bandari, V. K. *et al.* A flexible microsystem capable of controlled motion and actuation by wireless power transfer. *Nature Electronics* **3,** 172–180 (2020).

114. Wehner, M. *et al.* An integrated design and fabrication strategy for entirely soft, autonomous robots. *nature* **536,** 451–455 (2016).

115. Vella, D. & Mahadevan, L. The "Cheerios effect". *American Journal of Physics* **73,** 817–825 (2005).

116. Xie, G. *et al.* Continuous, autonomous subsurface cargo shuttling by nature-inspired meniscus-climbing systems. *Nature Chemistry* **14,** 208–215. ISSN: 1755-4349 (2022).

117. Mei, Y. *et al.* Versatile Approach for Integrative and Functionalized Tubes by Strain Engineering of Nanomembranes on Polymers. *Advanced Materials* **20,** 4085–4090 (2008).

118. Solovev, A. A., Mei, Y., Bermúdez Ureña, E., Huang, G. & Schmidt, O. G. Catalytic Microtubular Jet Engines Self-Propelled by Accumulated Gas Bubbles. *Small* **5,** 1688–1692 (2009).

119. Solovev, A. A., Mei, Y. & Schmidt, O. G. Catalytic Microstrider at the Air–Liquid Interface. *Advanced Materials* **22,** 4340–4344 (2010).

120. Solovev, A. A., Sanchez, S. & Schmidt, O. G. Collective behaviour of self-propelled catalytic micromotors. *Nanoscale* **5,** 1284–1293 (4 2013).

121. Eckmann, J.-P., Kamphorst, S. O. & Ruelle, D. Recurrence Plots of Dynamical Systems. *Europhysics Letters* **4,** 973–977 (1987).

122. Marwan, N., Romano, M. C., Thiel, M. & Kurths, J. Recurrence plots for the analysis of complex systems. *Physics Reports* **438,** 237–329 (2007).

123. Lin, S.-S. & Gurol, M. D. Catalytic Decomposition of Hydrogen Peroxide on Iron Oxide: Kinetics, Mechanism, and Implications. *Environmental Science & Technology* **32,** 1417–1423. ISSN: 0013-936X (1998).

124. Plauck, A., Stangland, E. E., Dumesic, J. A. & Mavrikakis, M. Active sites and mechanisms for H_2O_2 decomposition over Pd catalysts. *Proceedings of the National Academy of Sciences* **113,** E1973–E1982 (2016).

125. Gallager, R. G. *Stochastic Processes: Theory for Applications* ISBN: 9781107039759 (Cambridge University Press, 2013).

126. Medeiros, E. S., Feudel, U. & Zakharova, A. Asymmetry-induced order in multilayer networks. *Phys. Rev. E* **104,** 024302 (2 2021).

127. Zhang, Y., Ocampo-Espindola, J. L., Kiss, I. Z. & Motter, A. E. Random heterogeneity outperforms design in network synchronization. *Proceedings of the National Academy of Sciences* **118.** ISSN: 0027-8424 (2021).

128. Zhang, Y., Nishikawa, T. & Motter, A. E. Asymmetry-induced synchronization in oscillator networks. *Phys. Rev. E* **95,** 062215 (6 2017).

129. Nicolaou, Z. G., Case, D. J., Wee, E. B. v. d., Driscoll, M. M. & Motter, A. E. Heterogeneity-stabilized homogeneous states in driven media. *Nature Communications* **12,** 4486. ISSN: 2041-1723 (2021).

130. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I. & Shochet, O. Novel Type of Phase Transition in a System of Self-Driven Particles. *Phys. Rev. Lett.* **75,** 1226–1229 (6 1995).

131. Parisi, G. Order parameter for spin-glasses. *Physical Review Letters* **50,** 1946 (1983).

132. Ferré, G., Maillet, J.-B. & Stoltz, G. Permutation-invariant distance between atomic configurations. *The Journal of Chemical Physics* **143,** 104114 (2015).

133. Purcell, E. M. Life at low Reynolds number. *American Journal of Physics* **45,** 3–11 (1977).

134. Fenton, L. The sum of log-normal probability distributions in scatter transmission systems. *IEEE Transactions on Communication Systems* **8,** 57–67 (1960).

135. Moore, E. F. & Shannon, C. E. Reliable circuits using less reliable relays. *Journal of the Franklin Institute* **262,** 191–208. ISSN: 0016-0032 (1956).

136. Chen, S.-l., Lin, C.-t., Pan, C., Chieng, C.-c. & Tseng, F.-g. Growth and detachment of chemical reaction generated micro-bubbles on micro-textured catalyst. *Microfluidics and Nanofluidics* **7,** 807. ISSN: 1613-4990 (2009).

137. Moreno Soto, A., Maddalena, T., Fraters, A., van der Meer, D. & Lohse, D. Coalescence of diffusively growing gas bubbles. *Journal of Fluid Mechanics* **846,** 143–165 (2018).

138. Lv, P. *et al.* Self-Propelled Detachment upon Coalescence of Surface Bubbles. *Phys. Rev. Lett.* **127,** 235501 (23 2021).

139. Weon, B. M. & Je, J. H. Coalescence Preference Depends on Size Inequality. *Phys. Rev. Lett.* **108,** 224501 (22 2012).

140. Chen, R., Yu, H. W., Zhu, L., Patil, R. M. & Lee, T. Spatial and temporal scaling of unequal microbubble coalescence. *AIChE Journal* **63,** 1441–1450 (2017).

141. Wang, W., Chiang, T.-y., Velegol, D. & Mallouk, T. E. Understanding the efficiency of autonomous nano- and microscale motors. *Journal of the American Chemical Society* **135,** 10557–10565. ISSN: 0002-7863 (2013).

142. Wang, W., Duan, W., Sen, A. & Mallouk, T. E. Catalytically powered dynamic assembly of rod-shaped nanomotors and passive tracer particles. *Proceedings of the National Academy of Sciences* **110,** 17744–17749 (2013).

143. Lee, T.-C. *et al.* Self-Propelling Nanomotors in the Presence of Strong Brownian Forces. *Nano Letters* **14,** 2407–2412 (2014).

144. Zhang, Y. & Hess, H. Chemically-powered swimming and diffusion in the microscopic world. *Nature Reviews Chemistry* **5,** 500–510 (2021).

145. Wang, X.-Q. *et al.* In-built thermo-mechanical cooperative feedback mechanism for self-propelled multimodal locomotion and electricity generation. *Nature Communications* **9,** 3438. ISSN: 2041-1723 (2018).

146. Aubin, C. A. *et al.* Towards enduring autonomous robots via embodied energy. *Nature* **602,** 393–402. ISSN: 1476-4687 (2022).

147. Miskin, M. Z. *et al.* Electronically integrated, mass-manufactured, microscopic robots. *Nature* **584,** 557–561. ISSN: 1476-4687 (2020).

148. Brooks, A. M. & Strano, M. S. A conceptual advance that gives microrobots legs. *Nature* **584,** 530–531. ISSN: 1476-4687 (2020).

149. Yang, J. F. *et al.* Memristor Circuits for Colloidal Robotics: Temporal Access to Memory, Sensing, and Actuation. *Advanced Intelligent Systems,* 2100205 (2021).

150. Degrave, J. *et al.* Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602,** 414–419 (2022).

151. Won, D.-O., Müller, K.-R. & Lee, S.-W. An adaptive deep reinforcement learning framework enables curling robots with human-like performance in real-world conditions. *Science Robotics* **5** (2020).

152. Irpan, A. *Deep reinforcement learning doesn't work yet* `https://www.alexirpan.com/2018/02/14/rl-hard.html`. 2018.

153. Henderson, P. *et al.* Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* **32** (2018).

154. Ibarz, J. *et al.* How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research* **40,** 698–721 (2021).

155. Lillicrap, T. P. *et al.* Continuous control with deep reinforcement learning. *Proceedings of the International Conference on Learning Representations (ICLR)* (2016).

156. Plappert, M. *et al.* Parameter space noise for exploration. *Proceedings of the International Conference on Learning Representations (ICLR)* (2018).

157. Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning* **8,** 293–321. ISSN: 1573-0565. `https://doi.org/10.1007/BF00992699` (1992).

158. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. *Proceedings of the International Conference on Learning Representations (ICLR)* (2016).

159. Andrychowicz, M. *et al.* Hindsight experience replay. *Advances in Neural Information Processing Systems (NeurIPS)* **30** (2017).

160. Zhang, S. & Sutton, R. S. A deeper look at experience replay. *NeurIPS Deep Reinforcement Learning Symposium* (2017).

161. Wang, Z. *et al.* Sample Efficient Actor-Critic with Experience Replay. *Proceedings of the International Conference on Learning Representations (ICLR)* (2017).

162. Hessel, M. *et al.* Rainbow: Combining improvements in deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* **32** (2018).

163. Fedus, W. *et al.* Revisiting fundamentals of experience replay. *Proceedings of the International Conference on Machine Learning (ICML),* 3061–3071 (2020).

164. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518,** 529–533 (2015).

165. Ziebart, B. D., Bagnell, J. A. & Dey, A. K. Modeling Interaction via the Principle of Maximum Causal Entropy. *Proceedings of the 27th International Conference on Machine Learning (ICML),* 1255–1262 (2010).

166. Ziebart, B. D. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy* (Carnegie Mellon University, 2010).

167. Todorov, E. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences* **106,** 11478–11483 (2009).

168. Toussaint, M. Robot Trajectory Optimization Using Approximate Inference. *Proceedings of the 26th International Conference on Machine Learning (ICML),* 1049–1056 (2009).

169. Levine, S. & Koltun, V. Guided Policy Search. *Proceedings of the 30th International Conference on Machine Learning (ICML)* **28,** 1–9 (2013).

170. Haarnoja, T. *et al.* Learning to walk via deep reinforcement learning. *Proceedings of Robotics: Science and Systems (RSS)* (2018).

171. Chen, M. *et al.* Top-K Off-Policy Correction for a REINFORCE Recommender System. *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM),* 456–464 (2019).

172. Afsar, M. M., Crump, T. & Far, B. Reinforcement learning based recommender systems: A Survey. *ACM Computing Surveys* **55** (2022).

173. Chen, X., Yao, L., McAuley, J., Zhou, G. & Wang, X. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems* **264,** 110335 (2023).

174. Mitra, D. W matrix and the geometry of model equivalence and reduction. *Proceedings of the Institution of Electrical Engineers* **116,** 1101–1106 (1969).

175. Dean, S., Mania, H., Matni, N., Recht, B. & Tu, S. On the Sample Complexity of the Linear Quadratic Regulator. *Foundations of Computational Mathematics* **20,** 633–679 (2020).

176. Tsiamis, A. & Pappas, G. J. Linear Systems can be Hard to Learn. *2021 60th IEEE Conference on Decision and Control (CDC),* 2903–2910 (2021).

177. Tsiamis, A., Ziemann, I. M., Morari, M., Matni, N. & Pappas, G. J. *Learning to Control Linear Systems can be Hard* in *Proceedings of 35th Conference on Learning Theory (COLT)* **178** (2022), 3820–3857.

178. Thrun, S. B. *Efficient Exploration in Reinforcement Learning* tech. rep. (Carnegie Mellon University, 1992).

179. Amin, S., Gomrokchi, M., Satija, H., van Hoof, H. & Precup, D. A Survey of Exploration Methods in Reinforcement Learning. *arXiv preprint arXiv:2109.00157* (2021).

180. Kapur, J. N. *Maximum Entropy Models in Science and Engineering* ISBN: 9788122402162 (Wiley, 1989).

181. Taylor, A. T., Berrueta, T. A. & Murphey, T. D. Active learning in robotics: A review of control principles. *Mechatronics* **77,** 102576 (2021).

182. Todorov, E. *Linearly-solvable Markov decision problems* in *Advances in Neural Information Processing Systems (NeurIPS)* **19** (MIT Press, 2007), 1369–1376.

183. Seo, Y. *et al. State entropy maximization with random encoders for efficient exploration* in *Proceedings of the 38th International Conference on Machine Learning (ICML)* (2021), 9443–9454.

184. Prabhakar, A. & Murphey, T. Mechanical intelligence for learning embodied sensor-object relationships. *Nature Communications* **13,** 4108 (2022).

185. Chentanez, N., Barto, A. & Singh, S. *Intrinsically Motivated Reinforcement Learning* in *Advances in Neural Information Processing Systems (NeurIPS)* **17** (MIT Press, 2004).

186. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. *Curiosity-driven Exploration by Self-supervised Prediction* in *Proceedings of the International Conference on Machine Learning (ICML)* (2017), 2778–2787.

187. Taiga, A. A., Fedus, W., Machado, M. C., Courville, A. & Bellemare, M. G. *On Bonus Based Exploration Methods In The Arcade Learning Environment* in *Proceedings of the International Conference on Learning Representations (ICLR)* (2020).

188. Palmer, R. G. Broken ergodicity. *Advances in Physics* **31,** 669–735 (1982).

189. Islam, R., Henderson, P., Gomrokchi, M. & Precup, D. Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control. *arXiv preprint arXiv:1708.04133* (2017).

190. Moos, J. *et al.* Robust Reinforcement Learning: A Review of Foundations and Recent Advances. *Machine Learning and Knowledge Extraction* **4,** 276–315. ISSN: 2504-4990 (2022).

191. Strehl, A. L., Li, L., Wiewiora, E., Langford, J. & Littman, M. L. PAC model-free reinforcement learning. *Proceedings of the International Conference on Machine learning (ICML),* 881–888 (2006).

192. Strehl, A. L., Li, L. & Littman, M. L. Reinforcement Learning in Finite MDPs: PAC Analysis. *Journal of Machine Learning Research* **10** (2009).

193. Hairer, M. *Lecture notes on Ergodic Properties of Markov Chains* July 2018.

194. Kirk, R., Zhang, A., Grefenstette, E. & Rocktäschel, T. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research* **76,** 201–264 (2023).

195. Oh, J., Singh, S., Lee, H. & Kohli, P. *Zero-shot task generalization with multi-task deep reinforcement learning* in *Proceedings of the International Conference on Machine Learning (ICML)* (2017), 2661–2670.

196. Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L. & Haith, A. M. in *Comprehensive Physiology* 613–663 (John Wiley & Sons, Ltd, 2019). ISBN: 9780470650714.

197. Lu, K., Grover, A., Abbeel, P. & Mordatch, I. *Reset-Free Lifelong Learning with Skill-Space Planning* in *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

198. Chen, A., Sharma, A., Levine, S. & Finn, C. *You Only Live Once: Single-Life Reinforcement Learning* in *Advances in Neural Information Processing Systems (NeurIPS)* **35** (2022), 14784–14797.

199. Taylor, A., Singletary, A., Yue, Y. & Ames, A. *Learning for Safety-Critical Control with Control Barrier Functions* in *Proceedings of the 2nd Conference on Learning for Dynamics and Control (L4DC)* **120** (2020), 708–717.

200. Xiao, W. *et al.* BarrierNet: Differentiable Control Barrier Functions for Learning of Safe Robot Control. *IEEE Transactions on Robotics,* 1–19 (2023).

201. Seung, H. S., Sompolinsky, H. & Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A* **45,** 6056–6091 (8 1992).

202. Chen, C., Murphey, T. D. & MacIver, M. A. Tuning movement for sensing in an uncertain world. *eLife* **9,** e52371. ISSN: 2050-084X (2020).

203. Song, S. *et al.* Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *Journal of NeuroEngineering and Rehabilitation* **18,** 126. ISSN: 1743-0003 (2021).