



# Maximum Diffusion Reinforcement Learning

## Mathematical Approach

Updated : 25.07.12

Kyungmin Kwon

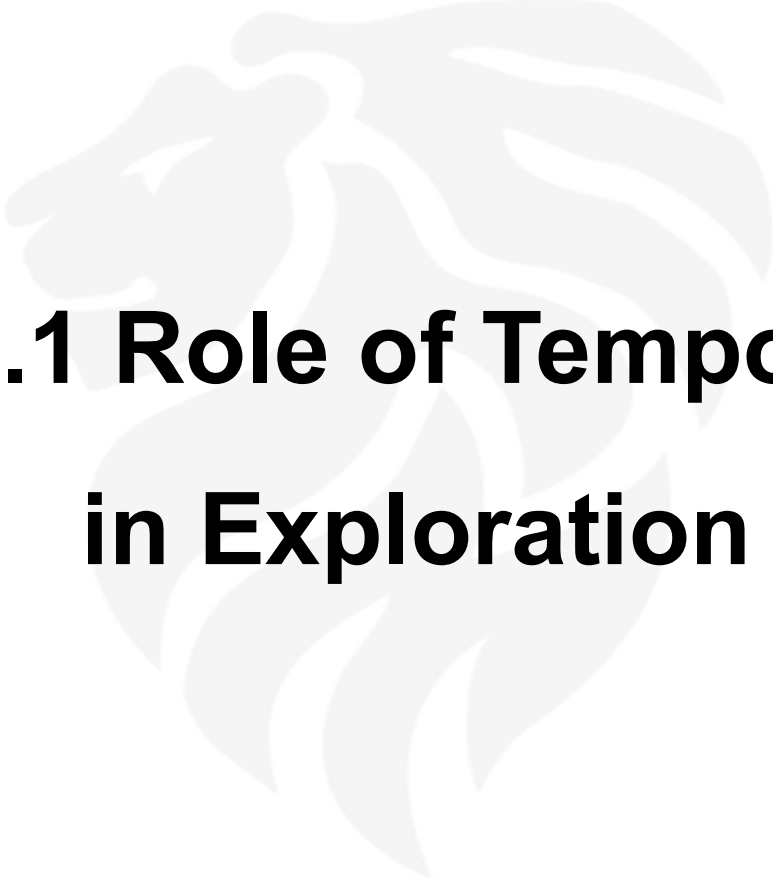


# Introduction

# Introduction to MaxDiff RL

---

- ***Based on principle of maximum caliber***
  - The principle of maximum caliber is a generalization of the principle of maximum entropy
  - MaxDiff RL is a generalization of MaxEnt RL.
- ***Underlying goal of RL algorithms***
  - Reaching desirable states
  - Realizing desirable trajectories ***What we want!***
- ***State trajectory as a mathematical abstraction***
  - Agent-environment state transition dynamics impact the performance of RL algorithms
  - Temporal correlations in the trajectories of RL agents ***How can we overcome?***



## **2.1 Role of Temporal Correlations in Exploration and Learning**

# Exploration, Temporal Correlation and Controllability

---

- **Exploration**

- A process by which agents become exposed to new experiences
- Broad importance to their learning performance.

- **Temporal correlation**

- The state transition dynamics of embodied learning agents can introduce temporal correlations
- It hinders their performance

- **Temporal correlation & Controllability**

**Definition 2.1.** 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  $u(t) : [t_0, t] \rightarrow \mathcal{U}$  that drives the system from state  $x^*$  at time  $t_0$  to  $x_1$  at time  $t$ .

- Controllability = diversity of state transition
- The lower controllability is, the higher temporal correlation is
- Cannot Exploration

# Role of Temporal Correlations in Exploration and Learning

---

- ***Consider randomized action exploration in linear time-varying (LTV) control system***

- In terms of continuous time deterministic trajectories

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad \cdots \quad (1)$$

- $A(t), B(t)$  appropriately dimensioned matrices
- $x(t), u(t)$  : state and control (action) vectors  $x(t) \in \mathcal{X} \subset \mathbb{R}^d$ ,  $u(t) \in \mathcal{U} \subset \mathbb{R}^m$
- $x(t_0) = x^*$  for  $\mathcal{T} = [t_0, t] \subset \mathbb{R}$

- ***The general form of solutions to this system with state-transition matrix  $\Psi(t, t_0)$***

$$x(t) = \Psi(t, t_0)x^* + \int_{t_0}^t \Psi(t, \tau)B(\tau)u(\tau)d\tau \quad \cdots \quad (2)$$

- The dynamics of any nonlinear system  $\sim$  locally LTV dynamics nearby trajectories

# Gramian Matrix

---

- **Define Controllability Gramian Matrix**

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

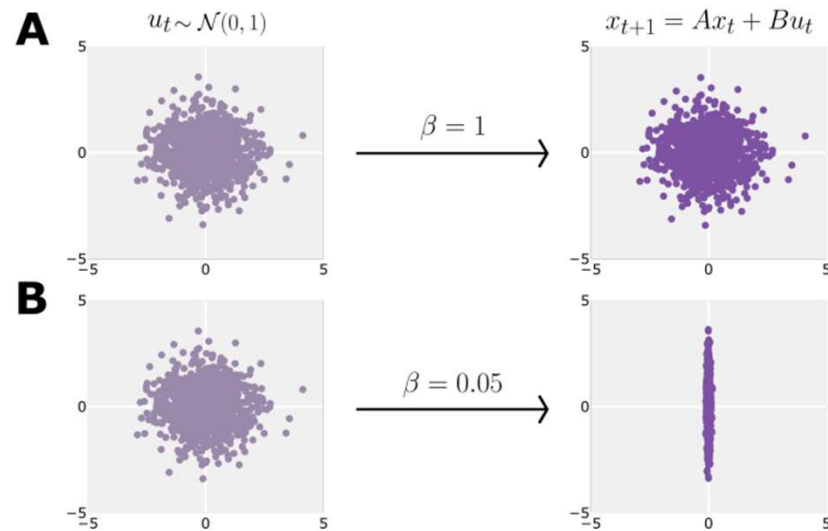
- A symmetric positive semidefinite matrix
- Depending on state-control matrix  $B(t)$  and state-transition matrix  $\Psi(t, t_0)$
- **Controllability metric** that quantifies the amount of energy required to actuate the different DOF of the system

# Characteristic of Gramian Matrix

## ▪ *Characteristic*

- When  $W(t_0, t)$  is full-rank, the system is **controllable** by Definition 2.1
- When  $W(t_0, t)$  is poorly conditioned, temporal correlations are introduced into the agent's state transitions

*Preventing effective exploration.. It's bad*





# System Trajectories on Random Variable

---

- **Connection btw. Naive random explorations, controllability, and temporal correlations**

- Recall **Eq.1** and design a controller that performs naive action randomization

$$u(t) \equiv \xi$$

- $\xi \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ ,  $\mathbf{0}$  is zero vector and  $\mathbb{I}$  is identity matrix
- The system trajectories are now **random variables**

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

- **Trajectory Statistics for characterizing the structure of temporal correlations by agent's dynamics**

- Mean trajectory from **Eq.2**

$$\mathbb{E}[x(t)] = \mathbb{E}\left[\Psi(t, t_0)x^* + \int_{t_0}^t \Psi(t, \tau)B(\tau)u(\tau)d\tau\right] = \Psi(t, t_0)x^* + \mathbb{E}\left[\int_{t_0}^t \Psi(t, \tau)B(\tau)u(\tau)d\tau\right] = \Psi(t, t_0)x^*$$

- The system follows an autonomous trajectory in the absence of control inputs

# System Trajectories on Random Variables (cont.)

---

- Define auto-covariance

$$C[x^*] = \mathbb{E}[(x(t) - \mathbb{E}[x(t)])(x(t) - \mathbb{E}[x(t)])^T | x(t_0) = x^*] = \Psi(t, t_0)x^* + \mathbb{E}\left[\int_{t_0}^t B(\tau)u(\tau)d\tau\right] = \Psi(t, t_0)x^*$$

- Trajectory auto-covariance

$$\begin{aligned} C[x^*] &= \mathbb{E}\left[\left(\Psi(t, t_0)x^* + \int_{t_0}^t \Psi(t, \tau)B(\tau)\xi d\tau - \mathbb{E}[x(t)]\right)\left(\Psi(t, t_0)x^* + \int_{t_0}^t \Psi(t, \tau)B(\tau)\xi d\tau - \mathbb{E}[x(t)]\right)^T\right] \\ &= \mathbb{E}\left[\left(\int_{t_0}^t \Psi(t, \tau)B(\tau)\xi d\tau\right)\left(\int_{t_0}^t \Psi(t, \tau)B(\tau)\xi d\tau\right)^T\right] = \mathbb{E}\left[\int_{t_0}^t d\tau \Psi(t, \tau)B(\tau)(\xi\xi^T)B(\tau)^T\Psi(t, \tau)^T\right] \\ &= \int_{t_0}^t d\tau \Psi(t, \tau)B(\tau)B(\tau)^T\Psi(t, \tau)^T = W(t_0, t) \end{aligned}$$

$$C[x^*] = W(t_0, t)$$

# Auto-covariance of Trajectories

---

$$C[x^*] = W(t_0, t)$$

- **Controllability is a measure of temporal correlations**

- Auto-covariance implicitly encodes temporal correlations of trajectories
- Auto-covariance and Gramian are not state( $x$ )-dependent properties in LTV systems

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

- **In Nonlinear system,**

- Linearizable nonlinear system

$$\dot{x}(t) = f(x, u) \Big|_{x=x^*, u=\xi} \approx f(x^*, \xi) + \frac{df}{dx} \Big|_{x=x^*} (x - x^*) + \frac{df}{du} \Big|_{u=\xi} (u - \xi) + \dots = A(t)x(t) + B(t)u(t)$$

- General nonlinear system

◆ So hard.. Proof by the Fokker-Planck Equation

*Even without a formal and explicit relationship, our analysis is valid in nonlinear system*

# In Nonlinear System

---

- ***State probability density function for describing the system's reachable states in nonlinear system***

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

- To highlight the role of controllability in the probability density of states reachable by the system through naive ***random exploration***
  - Equivalently a measure of temporal correlations of its state trajectories
- 
- ***Our assumption is different with other RL model***
    - We cannot assume that random inputs are capable of producing effective exploration of system states without an understanding of its controllability.
    - If  $W(t, t_0)$  is not full-rank, exploration is restricted to a linear subspace of an agent's exploration domain

**Remark 2.1.** *Temporal correlations and controllability can determine whether it is possible and how challenging it is to learn.*



## 2.2 Exploration as trajectory sampling

# Agent's experience is a Stochastic Process

---

- ***Framing exploration to overcome temporal correlation***

- Embodied agents must achieve exploration by changing the state of the environment through action
- Our goal is to achieve state exploration in an embodied system
- The agent's experience is modeled as a trajectory, which is a stochastic process

**Definition 2.2.** *A stochastic process is a family 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  $(\Omega, \mathcal{F}, \mathbb{P})$ . We take the sample space  $\Omega$  to be measurable,  $\mathcal{F}$  to be a Borel  $\sigma$ -algebra, and  $\mathbb{P}$  to be a probability measure. We note that the random variables assume values in a compact state space  $\mathcal{X} \subset \mathbb{R}^d$ , and that each sample path takes value in a measurable space  $\mathcal{X}^{\mathcal{T}}$  with Borel  $\sigma$ -algebra  $\mathcal{B}(\mathcal{X}^{\mathcal{T}})$ .*

# Definition for Modeling Stochastic Trajectories

---

- **Probability Space**  $(\Omega, \mathcal{F}, \mathbb{P})$

- sample space  $\Omega$  to be measurable
- $\mathcal{F}$  to be Borel  $\sigma$ -algebra

*The Borel sigma-algebra is the collection of events generated from open sets through operations, allowing probability to be defined on them*

- **Stochastic process**

- Families of random variables indexed according to some ‘time-like’ set  $\mathcal{T}$
- It can be possible for discrete, continuous time

- **Sample Path**  $x_{\mathcal{T}}(\omega)$

- For each  $\omega \in \Omega$ , sample path  $x_{\mathcal{T}}(\omega) = \{X(t, \omega)\}_{t \in \mathcal{T}}$

# Definition for Modeling Stochastic Trajectories (cont.)

---

- **Pushforward measure (function)**  $x_T: \Omega \rightarrow \mathcal{X}^T$ 
  - $P_F: \mathcal{B}(\mathcal{X}^T)$  is a probability distribution over trajectories, defined on the Borel sigma-algebra
  - $P_F[x_T \in A] = P(x_T^{-1}(A))$  for some  $A \subset \mathcal{X}^T$
- **Agent's experience (state trajectories)**
  - Individual trajectory of the stochastic process
  - $x(t) = x_T(\omega) \in \mathcal{X}^T$



# Path Distribution (Trajectory Distribution)

---

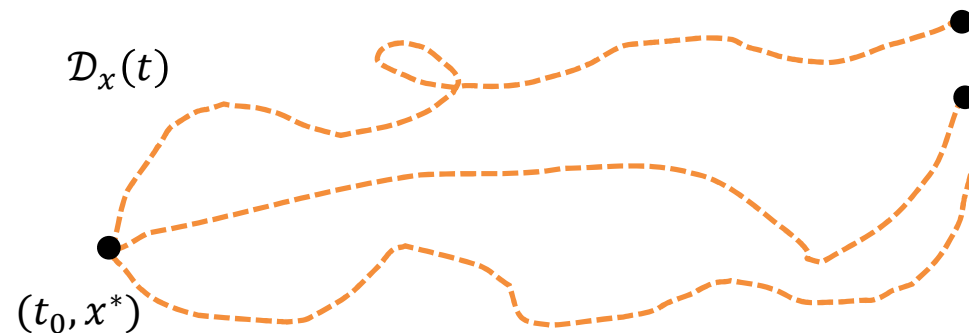
- **Probability density function on Feynman path integral formalism**

- To describe the likelihoods of individual state trajectories
- The probability density function is associated with the pushforward measure

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

*Path distribution, trajectory distribution*

- ◆  $P \in \mathcal{X}^{\mathcal{T}} \rightarrow [0, \infty)$  (not normalized)
- ◆  $\mathcal{D}_x(t)$  : integration over sample paths in the Feynman path integral formalism
- ◆  $P[x(t)]$  : probability density of a given state trajectory of the stochastic process  $x_{\mathcal{T}}(\omega) \in \mathcal{X}^{\mathcal{T}}$



# Temporal Correlation along Sample Paths

---

- **To quantify correlations along sample paths or state trajectories**

- A local measure of temporal correlations  $\mathcal{C}[x^*]$  over time intervals  $[t_i, t_i + \Delta t] \subset \mathcal{T}$
- Auto-covariance function  $K_{XX}(t_1, t_2)$  at any two points in time  $t_1, t_2 \in \mathcal{T}$  and  $\{X(t, \omega)\}_{t \in \mathcal{T}}$

$$K_{XX}(t_1, t_2) = \mathbb{E}[(X(t_1) - \mathbb{E}[X(t_1)])\mathbb{E}[(X(t_2) - \mathbb{E}[X(t_2)])]^T]$$

- Covariance @ two time points, Temporal correlation @ any time interval
- With an initial condition for each sample paths,  $X(t_i) = x^*$  for some  $x^* \in \mathcal{X}$

$$\mathcal{C}[x^*] = \mathbb{E}[(x(t) - \mathbb{E}[x(t)])(x(t) - \mathbb{E}[x(t)])^T | x(t_0) = x^*] = \int_{t_i}^{t_i + \Delta t} d\tau K_{XX}(t_i, \tau)$$

- Temporal correlation = Integrated autocovariance
- Practically, normalize the covariance  $\mathcal{C}[x^*]/\Delta t$

# Stochastic Control Process

---

- ***Probability densities over state trajectories***

- Probability densities are dependent on the dynamics that govern the agent-environment's time-evolution through state space
- Depending on the choice of controller, too

- Define a controller that produces an input to the system dynamics at every time point

$$u(t) : \mathcal{T} \rightarrow \mathcal{U}$$

- A choice of controller induces a different probability density over sample paths

**Definition 2.3.** *A stochastic control process is a stochastic process (Definition 2.2) 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 resulting density  $P_{u(t)} : \mathcal{X}^{\mathcal{T}} \rightarrow [0, \infty)$  is parametrized by a controller  $u(t) : \mathcal{T} \rightarrow \mathcal{U}$ .*

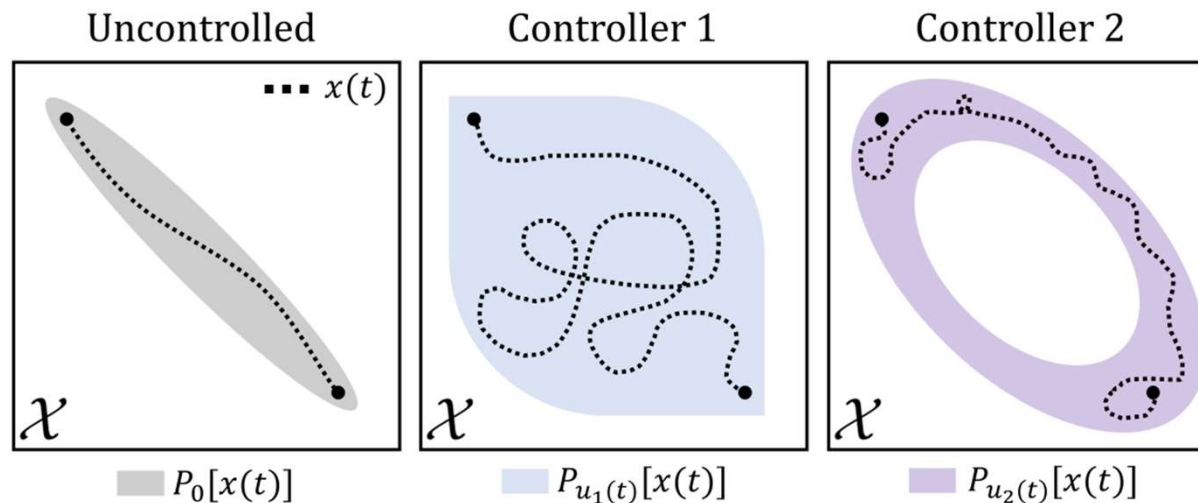
# Role of Stochastic Control Process

## ▪ Role of controller

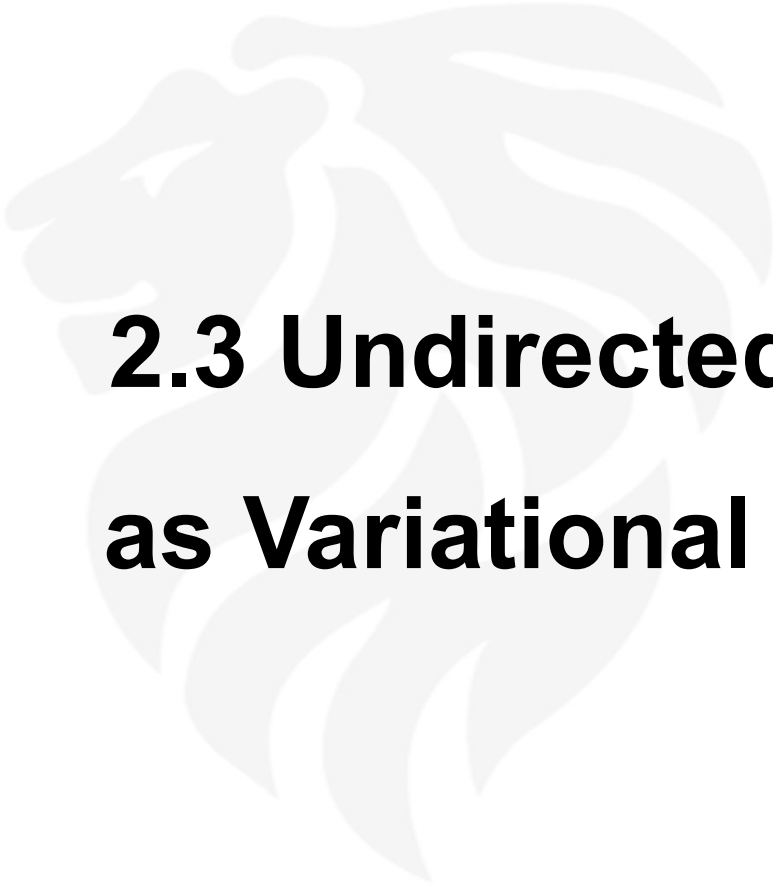
- A controller encourages the embedded agent to explore
- We should design a controller **that maximizes the regions of the exploration domain** from which we can sample trajectories

*To maximize the support of the agent's sample path distribution*

$$\text{supp}(P) = \{x \in \mathcal{X} | P > 0\}$$



- If the agent's sample paths are biased towards a given goal, then our agent's path distribution should reflect this.



## **2.3 Undirected Exploration as Variational Optimization**

# Under Undirected Exploration

---

- ***How to simultaneously control the spread of probability mass and the support of a probability distribution?***
  - Optimize its entropy
- ***Disembodied agents with unconstrained dynamics***
  - Maximum entropy → leading to complete asymptotic exploration (i.i.d. uniform sampling)
- ***Embodied agents with constrained dynamics***
  - Physical constraints prevent i.i.d. uniform sampling
  - Maximizing trajectory entropy will give minimally correlated experiences

***Find Analytical form of the maximum entropy path distribution under undirected exploration!***

# Maximization of Entropy

---

- **Maximum Caliber**

*trajectories*

- Maximum caliber is a generalization of the principle of maximum entropy to function spaces

- **Define expectation over sample paths**

- Real-valued function  $f(\cdot)$  of  $x_{\mathcal{T}}$

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

- **Maximizing the entropy of sample paths  $S[P[x(t)]]$  for uniform distribution**

- Start from the Shannon Entropy

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

$$\operatorname{argmax}_{P[x(t)]} - \int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \log P[x(t)] \mathcal{D}x(t)$$

$\int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \mathcal{D}x(t)$   
*unnormalized..*

# Constraint Condition

---

- ***Constraint the optimization problem***

- What sorts of principled constraints could be applied?
  - ◆ Inherently nonequilibrium systems – cannot be applied energy conservation
  - ◆ The system's ability to explore is closely tied to a measure of its temporal correlations
  - ◆ Choose to constrain the ***velocity fluctuations*** of our stochastic process

- ***Velocity fluctuation***

- *It is finite and consistent with the integrated autocovariance statistics of the process*
- *It is determined empirically because of unknown physics of embodied agent*
- *Embodied agents may be spatially inhomogeneous and difficult to know a priori*

***Through this constraint, we can ensure that agent sample paths are continuous in time!***



# Velocity Fluctuation

---

- **System's velocity fluctuation along sample paths  $x(t)$**

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

- Assume that the covariance matrix is full-rank
  - ◆ Velocity fluctuations are not degenerate anywhere in the state space
- It guarantees that our resulting path distribution is non-degenerate.

- **Zero-mean fluctuation for auto-covariance**

$$\langle \dot{x}(t)\dot{x}(t)^T \rangle_{x^*} = C[x^*]$$

- Assume that it is satisfied Lipschitz continuity so that their spatial variances are bounded in the exploration domain
- The linearizability of the underlying agent dynamics satisfies this property
- For valid probability density over trajectories,  $P[x(t)]$  **integrates to 1**

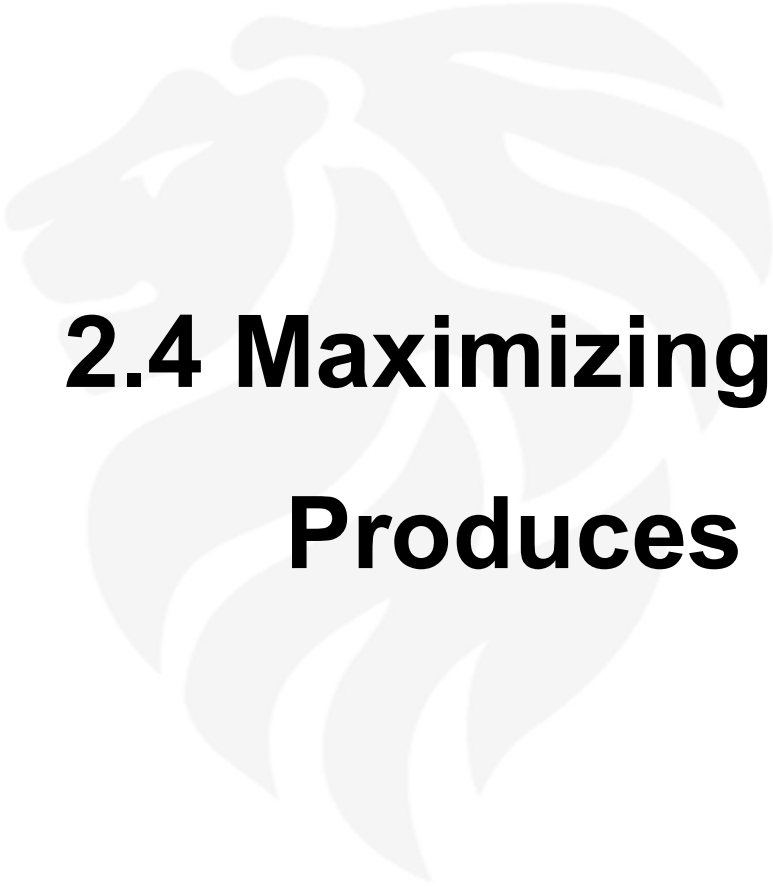
# Optimization problem using Lagrange Multiplier

---

- *Variational optimization problem with our constraints*

$$\operatorname{argmax}_{P[x(t)]} - \int_{\mathcal{X}^T} P[x(t)] \log P[x(t)] \mathcal{D}x(t) - \lambda_0 \left( \int_{\mathcal{X}^T} P[x(t)] \mathcal{D}x(t) - 1 \right) - \int_{\mathcal{X}} \operatorname{Tr}(\Lambda(x^*)^T (\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} - C[x^*])) dx_*$$

- $\lambda_0, \Lambda(x^*)$  : Lagrange multiplier
- Ensuring valid probability density & least-correlated sample path
- Providing the distribution with greatest support and the most uniformly spread probability mass



## **2.4 Maximizing Path Entropy**

### **Produces Diffusion**

# Stochastic Diffusion Process

---

**Theorem 2.1.** *The maximum caliber sample paths of a stochastic control process (Definition 2.3) with a maximum entropy exploration (in the sense of Eq. 17) are given by diffusion with spatially-varying coefficients.*

- Despite physical limitations, an agent's best exploration strategy looks like diffusion that adapts to its location

- **Maximum entropy probability density**

$$P_{max}[x(t)] = \frac{1}{Z} \exp \left[ - \int_{t_0}^t \dot{x}(\tau)^T D(x(\tau))^{-1} \dot{x}(\tau) d\tau \right] = \frac{1}{Z} \exp \left[ - \int_{t_0}^t \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) d\tau \right]$$

- Equivalent to the path probability of a diffusing particle with a spatially inhomogeneous diffusion tensor
- The least-correlated sample paths are statistically equivalent to diffusion
- We don't need to assume Markovian or ergodic dynamics
- These properties emerge automatically from maximizing path entropy

See Supplement

# Supple : Stochastic Diffusion Process

**Theorem 2.1.** *The maximum caliber sample paths of a stochastic control process (Definition 2.3) with a maximum entropy exploration (in the sense of Eq. 17) are given by diffusion with spatially-varying coefficients.*

- Proof

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

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

$$\log P_{\max}[x(t)] = -1 - \lambda_0 - \int_{t_0}^t \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) d\tau$$

$$P_{\max}[x(t)] = \exp[-1 - \lambda_0] \exp \left[ - \int_{t_0}^t \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) d\tau \right]$$

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

# Assumption of Stochastic Diffusion Process

---

- **Two Assumption**

- The diffusion tensor  $\Lambda^{-1}(\cdot)$  is full-rank and invertible everywhere in the state space
- $\Lambda^{-1}(\cdot)$  is Lipschitz and bounded everywhere on  $\mathcal{X}$

- **Our stochastic process is inherently Markovian and Ergodic**

**Corollary 2.1.1.** *The sample paths of a stochastic control process (Definition 2.3) with a maximum entropy exploration strategy (in the sense of Eq. 17) satisfy the **Markov property**.*

$$p_{max}(x_{t+\delta t}) \approx \frac{1}{Z_d} \exp \left[ -|x_{t+\delta t} - x_t|_{\Lambda(x_t)}^2 \right]$$

**Path distribution depends only on the current state**

**Corollary 2.1.2.** *A stochastic control process (Definition 2.3) in a compact and connected space  $\mathcal{X} \subset \mathbb{R}^d$  with a maximum entropy exploration strategy (in the sense of Eq. 17) is **ergodic**.*

**irreducible + aperiodicity in finite-time and space**

See Supplement

# Supple : Stochastic Diffusion Process is Markovian

---

- **Markovian**

Start from this equation

$$\operatorname{argmax}_{P[x(t)]} - \int_{\mathcal{X}^T} P[x(t)] \log P[x(t)] \mathcal{D}x(t) - \lambda_0 \left( \int_{\mathcal{X}^T} P[x(t)] \mathcal{D}x(t) - 1 \right) - \int_{\mathcal{X}} \operatorname{Tr}(\Lambda(x^*)^T (\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} - C[x^*])) dx_*$$

Let  $x_t$  be the initial condition

$$p_{max}(x_{t+\delta t}) = \frac{1}{Z} \exp \left[ - \int_t^{t+\delta t} \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) d\tau \right]$$

For very small  $\delta t \ll 1$

$$\dot{x}(\tau) \approx \frac{x_{t+\delta t} - x_t}{\delta t}$$

$$\begin{aligned} p_{max}(x_{t+\delta t}) &= \frac{1}{Z} \exp \left[ - \int_t^{t+\delta t} \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) d\tau \right] \approx \frac{1}{Z} \exp \left[ - \left( \frac{x_{t+\delta t} - x_t}{\delta t} \right)^T \Lambda(x_t) \frac{x_{t+\delta t} - x_t}{\delta t} \delta t \right] \\ &\approx \frac{1}{Z_d} \exp \left[ - |x_{t+\delta t} - x_t|_{\Lambda(x_t)}^2 \right] = \frac{1}{\sqrt{\det(2\pi\Lambda^{-1}(x^*))}} \exp \left[ - |x_{t+\delta t} - x_t|_{\Lambda(x_t)}^2 \right] \end{aligned}$$

$$|a|_{\Lambda}^2 = a^T \Lambda a$$

# Supple : Stochastic Diffusion Process is Ergodic

---

- **Ergodic (irreducible + aperiodicity in finite-state case)**

Discretizing our optimal stochastic control process in time and space

$$P_{max}[x_{1:N}] = \prod_{t=1}^{N-1} p_{max}(x_{t+1}|x_t)$$

$$p_{max}(x_{t+1}|x_t) > 0, \forall x_t, x_{t+1} \in \mathcal{X}, \forall t \in \mathcal{T}$$

$\mathcal{X}$  is compact, so we can discretize it into a finite number of representative states, yielding a finite-state approximation.

$\mathcal{X}$  is connected, we can reach all the possible states when we start in any state : *irreducible*

Aperiodicity

$$p_{max}(x^*|x^*) > 0, \forall x^* \in \mathcal{X}$$

If a state can transition to itself in one step  $p_{max}(x^*|x^*)$ , its period is 1, making it *aperiodic*



# Full Expression of Lagrange Multiplier

- *In order to fully express our maximum entropy exploration*

- Determine matrix-valued Lagrange multiplier  $\Lambda(\cdot)$
- Discretizing time  $\{1, \dots, N\}$  and space; time index  $i, j$
- Feynman path integrals are discretized

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



$$\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} = \mathbf{C}^{-1}[x^*] = \Lambda(x^*)^{-1}$$

- Final Form of the maximum entropy exploration sample path distribution (discrete time, too)

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

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

System is described  
by *maximally diffusive*

See Supplement

# Supple : Full Expression of Lagrange Multiplier

---

- *In order to fully express our maximum entropy exploration*

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

- By Fubini-Tonelli theorem and divide two cases  $j-1 \rightarrow j \rightarrow j+1$

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

$$\begin{array}{ccc} \text{The others} & j-1 \rightarrow j & j \rightarrow j+1 \\ = \sum_{j=1}^{N-1} \prod_{i \neq j, i=1}^{N-1} \left[ \int_{\mathcal{X}} dx_{i+1} p_{\max}(x_{i+1} | x_i) \right] \times p_{\max}(x^* | x_{j-1}) \times \int_{\mathcal{X}} dx_{j+1} p_{\max}(x_{j+1} | x^*) (x_{j+1} - x^*)(x_{j+1} - x^*)^T \end{array}$$

- Define closed-form Gaussian integration

$$I = \int_{\mathcal{X}} dx_{j+1} p_{\max}(x_{j+1} | x^*) (x_{j+1} - x^*)(x_{j+1} - x^*)^T \approx \frac{1}{Z_D} \int_{\mathcal{X}} dx_{j+1} e^{-(x_{j+1} - x^*)^T \Lambda(x^*) (x_{j+1} - x^*)} (x_{j+1} - x^*)(x_{j+1} - x^*)^T$$

## Supple : Full Expression of Lagrange Multiplier

---

- Assume that the domain of exploration is large  $\|x_{j+1} - x_j\| \rightarrow \infty$   $\mathbf{1} = (1, 1, 1, 1 \dots 1)^T$

$$I \approx \frac{1}{Z_D} \Lambda(x^*)^{-1} \left[ \sqrt{\det(2\pi\Lambda^{-1}(x^*))} - (x_{j+1} - x^*)^T \mathbf{1} e^{-(x_{j+1} - x^*)^T \Lambda(x^*) (x_{j+1} - x^*)} \right]_{x_{j+1}=-\infty}^{x_{j+1}=\infty} = \frac{1}{Z_D} \Lambda(x^*)^{-1} \sqrt{\det(2\pi\Lambda^{-1}(x^*))}$$

- Velocity fluctuation is full-rank, so  $\Lambda^{-1}(x^*)$  must be full-rank

$$\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} = \sum_{j=1}^{N-1} \prod_{i \neq j, i=1}^{N-1} \left[ \int_{\mathcal{X}} dx_{i+1} p_{max}(x_{i+1} | x_i) \right] \times p_{max}(x^* | x_{j-1}) \times \Lambda(x^*)^{-1}$$

- Re-expanding by Dirac-delta function  $\delta(x_j - x^*)$

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

$$\langle \dot{x}(t) \dot{x}(t)^T \rangle_{x^*} = C^{-1}[x^*] = \Lambda(x^*)^{-1}$$

# Practical Guide : Maximum Entropy Undirected Exploration

---

- *When we are not interested in exploring directly on the **full state** space of our control system*

- $\mathcal{X}$  onto the desired exploration domain  $\mathcal{Y}$
- Consider some differentiable coordinate transformation  $y(t) = \psi(x(t))$
- Jacobian matrix  $J_\psi[\cdot]$  corresponding to the coordinate transformation  $\psi$

- Probability density in  $\mathcal{Y}$

$$P_{max}(y(t))$$

- Diffusion tensor  $\mathcal{C}[y^*]$  in  $\mathcal{Y}$

$$\mathcal{C}[y^*] = J_\psi[x^*]\mathcal{C}[x^*]J_\psi[x^*]^T$$



## **2.4 Directed Exploration as Variational Optimization**

# Directed Exploration

---

- ***Undirected exploration***

- A process that is blind to any notion of importance ascribed to state space (or exploration domain)

- ***Directed exploration***

- We already know a priori understanding of what regions of the exploration domain are important
- For directed exploration, we introduce ‘free energy’ minimization objective with a bounded potential  $V$
- Undirected exploration + free energy = directed exploration

# Diffusion under the Potential

---

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

- **Define potential over  $\mathcal{T} = [t_0, t]$**

- It captures the average cost over all possible system paths

$$\langle V[x(t)] \rangle_P = \int_{\mathcal{X}^{\mathcal{T}}} P[x(t)] \left( \int_{t_0}^t V[x(\tau)] d\tau \right) \mathcal{D}x(t)$$

- $\langle V[x(t)] \rangle_P$  must be bounded.

- **Free energy functional objective**

$$\operatorname{argmin}_{P[x(t)]} \langle V[x(t)] \rangle_P - S[P[x(t)]]$$

- Diffusion (exploration) + biasing (cost minimization)

# Diffusion under the Potential (cont.)

---

- *Minimum free energy path distribution*

$$\mathcal{F} \equiv \langle V[x(t)] \rangle_P - S[P[x(t)]]$$

$$P_{max}^V[x(t)] = \frac{1}{Z} \exp \left[ - \int_{t_0}^t d\tau (V[x(\tau)] + \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau)) \right]$$

- The optimal directed exploration strategy is to scale the strength of diffusion with respect to the desirability of the state
- Maximally diffusive with respect to the underlying potential

$$P_{max}^V[x(t)] = P_{max}[x(t)] e^{-\int_{t_0}^t d\tau V[x(\tau)]}$$

- Free space maximally diffusive in the absence of a potential

See Supplement



# Supple : Minimum Free Energy Path Distribution

**Theorem 2.1.** *The maximum caliber sample paths of a stochastic control process (Definition 2.3) with a maximum entropy exploration (in the sense of Eq. 17) are given by diffusion with spatially-varying coefficients.*

• Proof

$$\langle V[x(t)] \rangle_P = \int_{\mathcal{X}^T} P[x(t)] \left( \int_{t_0}^t V[x(\tau)] d\tau \right) \mathcal{D}x(t)$$

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

$$\log P_{max}[x(t)] = -1 - \lambda_0 - \int_{t_0}^t \left( V[x(\tau)] + \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) \right) d\tau$$

$$P_{max}[x(t)] = \exp[-1 - \lambda_0] \exp \left[ - \left( V[x(\tau)] + \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) \right) \right]$$

$$P_{max}[x(t)] = \frac{1}{Z} \exp \left[ - \int_{t_0}^t d\tau \left( V[x(\tau)] + \dot{x}(\tau)^T \Lambda(x(\tau)) \dot{x}(\tau) \right) \right]$$

# Properties of Directed Exploration

---

- ***Minimum free energy path distribution with Markov decision process***

- Discretized agent state trajectories  $l(\cdot) = V(\cdot)$

$$P_{max}^l[x_{1:N}] = \prod_{t=1}^{N-1} p_{max}[x_{t+1}|x_t] e^{-l(t)}$$

- A dependence on agent actions into the cost functions
- The lower potential is, the more frequently it appears

- ***Properties of directed exploration strategy***

- Markovian
  - ◆ If the potential is dependent on current (memoryless), Markovian isn't broken
- Ergodic
  - ◆ It can be satisfied under some mild assumptions

# Supple : Stochastic Diffusion Process under Potential is Ergodic

---

**Theorem 2.2.** *A stochastic control process (Definition 2.3) in a compact and connected space  $\mathcal{X} \subset \mathbb{R}^d$  with a maximum entropy exploration strategy in a potential (in the sense of Eq. 31) is ergodic.*

- Since the potential is well-bounded, the probability density is also greater than 0

$$V_{min} \leq V[x_t] \leq V_{max} < \infty$$
$$P_{max}^l[x_{1:N}] = \prod_{t=1}^{N-1} p_{max}[x_{t+1}|x_t] e^{-l(t)}$$

$$p_{max}^l[x_{1:N}] = p_{max}(x_{t+1}|x_t) e^{-l(t)} > 0, \forall x_t, x_{t+1} \in \mathcal{X}, \forall t \in \mathcal{T}$$

- So, it also satisfies **irreducible** and **aperiodicity** in discrete space and time, too
- (the rest condition is under the **proof of Theorem 2.2.**)

# Ergodicity of Directed Exploration

---

- ***Comments of ergodicity of directed exploration***
  - The net effect of the potential is to reshuffle probability mass in the stationary distribution of the agent's underlying Markov chain
  - Minimum free energy exploration leads to ergodic coverage of the exploration domain with respect to the potential
  - That is, encoded via the potential, this strategy provides an asymptotic guarantee on learning



## **2.6 Minimizing Path Free Energy**

### **Produces Diffusive Gradient Descent**

# Supple : Legendre's transformation (미완)

## THE CANONICAL EQUATIONS†

### §40. Hamilton's equations

THE formulation of the laws of mechanics in terms of the Lagrangian, and of Lagrange's equations derived from it, presupposes that the mechanical state of a system is described by specifying its generalised co-ordinates and velocities. This is not the only possible mode of description, however. A number of advantages, especially in the study of certain general problems of mechanics, attach to a description in terms of the generalised co-ordinates and momenta of the system. The question therefore arises of the form of the equations of motion corresponding to that formulation of mechanics.

The passage from one set of independent variables to another can be effected by means of what is called in mathematics *Legendre's transformation*. In the present case this transformation is as follows. The total differential of the Lagrangian as a function of co-ordinates and velocities is

$$dL = \sum_i \frac{\partial L}{\partial q_i} dq_i + \sum_i \frac{\partial L}{\partial \dot{q}_i} d\dot{q}_i.$$

This expression may be written

$$dL = \sum \dot{p}_i dq_i + \sum p_i d\dot{q}_i, \quad (40.1)$$

since the derivatives  $\partial L/\partial \dot{q}_i$  are, by definition, the generalised momenta, and  $\partial L/\partial q_i = \dot{p}_i$  by Lagrange's equations. Writing the second term in (40.1) as  $\sum p_i d\dot{q}_i = d(\sum p_i \dot{q}_i) - \sum \dot{q}_i dp_i$ , taking the differential  $d(\sum p_i \dot{q}_i)$  to the left-hand side, and reversing the signs, we obtain from (40.1)

$$d(\sum p_i \dot{q}_i - L) = - \sum \dot{p}_i dq_i + \sum \dot{q}_i dp_i.$$

The argument of the differential is the energy of the system (cf. §6); expressed in terms of co-ordinates and momenta, it is called the *Hamilton's function* or *Hamiltonian* of the system:

$$H(p, q, t) = \sum_i p_i \dot{q}_i - L. \quad (40.2)$$

† The reader may find useful the following table showing certain differences between the nomenclature used in this book and that which is generally used in the English literature.

Here	Elsewhere
Principle of least action	Hamilton's principle
Maupertuis' principle	Principle of least action
Action	Maupertuis' principle
Abbreviated action	Hamilton's principal function
—Translator.	Action

From the equation in differentials

$$dH = - \sum \dot{p}_i dq_i + \sum \dot{q}_i dp_i, \quad (40.3)$$

in which the independent variables are the co-ordinates and momenta, we have the equations

$$\dot{q}_i = \partial H / \partial p_i, \quad \dot{p}_i = - \partial H / \partial q_i. \quad (40.4)$$

These are the required equations of motion in the variables  $p$  and  $q$ , and are called *Hamilton's equations*. They form a set of  $2s$  first-order differential equations for the  $2s$  unknown functions  $p_i(t)$  and  $q_i(t)$ , replacing the  $s$  second-order equations in the Lagrangian treatment. Because of their simplicity and symmetry of form, they are also called *canonical equations*.

The total time derivative of the Hamiltonian is

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} + \sum \frac{\partial H}{\partial q_i} \dot{q}_i + \sum \frac{\partial H}{\partial p_i} \dot{p}_i.$$

Substitution of  $\dot{q}_i$  and  $\dot{p}_i$  from equations (40.4) shows that the last two terms cancel, and so

$$dH/dt = \partial H / \partial t. \quad (40.5)$$

In particular, if the Hamiltonian does not depend explicitly on time, then  $dH/dt = 0$ , and we have the law of conservation of energy.

As well as the dynamical variables  $q$ ,  $\dot{q}$  or  $q$ ,  $p$ , the Lagrangian and the Hamiltonian involve various parameters which relate to the properties of the mechanical system itself, or to the external forces on it. Let  $\lambda$  be one such parameter. Regarding it as a variable, we have instead of (40.1)

$$dL = \sum \dot{p}_i dq_i + \sum p_i d\dot{q}_i + (\partial L / \partial \lambda) d\lambda,$$

and (40.3) becomes

$$dH = - \sum \dot{p}_i dq_i + \sum \dot{q}_i dp_i - (\partial H / \partial \lambda) d\lambda.$$

Hence

$$(\partial H / \partial \lambda)_{p,q} = - (\partial L / \partial \lambda)_{\dot{q},q}, \quad (40.6)$$

which relates the derivatives of the Lagrangian and the Hamiltonian with respect to the parameter  $\lambda$ . The suffixes to the derivatives show the quantities which are to be kept constant in the differentiation.

This result can be put in another way. Let the Lagrangian be of the form  $L = L_0 + L'$ , where  $L'$  is a small correction to the function  $L_0$ . Then the corresponding addition  $H'$  in the Hamiltonian  $H = H_0 + H'$  is related to  $L'$  by

$$(H')_{p,q} = - (L')_{\dot{q},q}. \quad (40.7)$$

It may be noticed that, in transforming (40.1) into (40.3), we did not include a term in  $dt$  to take account of a possible explicit time-dependence

of the Lagrangian, since the time would there be only a parameter which would not be involved in the transformation. Analogously to formula (40.6), the partial time derivatives of  $L$  and  $H$  are related by

$$(\partial H / \partial t)_{p,q} = - (\partial L / \partial t)_{\dot{q},q}. \quad (40.8)$$

## PROBLEMS

PROBLEM 1. Find the Hamiltonian for a single particle in Cartesian, cylindrical and spherical co-ordinates.

SOLUTION. In Cartesian co-ordinates  $x, y, z$ ,

$$H = \frac{1}{2m} (p_x^2 + p_y^2 + p_z^2) + U(x, y, z);$$

in cylindrical co-ordinates  $r, \phi, z$ ,

$$H = \frac{1}{2m} \left( p_r^2 + \frac{p_\phi^2}{r^2} + p_z^2 \right) + U(r, \phi, z);$$

in spherical co-ordinates  $r, \theta, \phi$ ,

$$H = \frac{1}{2m} \left( p_r^2 + \frac{p_\theta^2}{r^2} + \frac{p_\phi^2}{r^2 \sin^2 \theta} \right) + U(r, \theta, \phi).$$

PROBLEM 2. Find the Hamiltonian for a particle in a uniformly rotating frame of reference.

SOLUTION. Expressing the velocity  $\mathbf{v}$  in the energy (39.11) in terms of the momentum  $\mathbf{p}$  by (39.10), we have  $H = p^2/2m - \boldsymbol{\Omega} \cdot \mathbf{r} \times \mathbf{p} + U$ .

PROBLEM 3. Find the Hamiltonian for a system comprising one particle of mass  $M$  and  $n$  particles each of mass  $m$ , excluding the motion of the centre of mass (see §13, Problem).

SOLUTION. The energy  $E$  is obtained from the Lagrangian found in §13, Problem, by changing the sign of  $U$ . The generalised momenta are

$$\begin{aligned} \mathbf{p}_a &= \partial L / \partial \mathbf{v}_a \\ &= m \mathbf{v}_a - (m^2/\mu) \sum_a \mathbf{v}_a. \end{aligned}$$

Hence

$$\begin{aligned} \sum \mathbf{p}_a &= m \sum \mathbf{v}_a - (nm^2/\mu) \sum \mathbf{v}_a \\ &= (mM/\mu) \sum \mathbf{v}_a, \\ \mathbf{v}_a &= \mathbf{p}_a/m + (1/M) \sum \mathbf{p}_c. \end{aligned}$$

Substitution in  $E$  gives

$$H = \frac{1}{2m} \sum_a \mathbf{p}_a^2 + \frac{1}{2M} \left( \sum_a \mathbf{p}_a \right)^2 + U.$$

# Supple : The principle of Least action (미완)

## §2. The principle of least action

The most general formulation of the law governing the motion of mechanical systems is the *principle of least action* or *Hamilton's principle*, according to which every mechanical system is characterised by a definite function  $L(q_1, q_2, \dots, q_s, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_s, t)$ , or briefly  $L(q, \dot{q}, t)$ , and the motion of the system is such that a certain condition is satisfied.

Let the system occupy, at the instants  $t_1$  and  $t_2$ , positions defined by two sets of values of the co-ordinates,  $q^{(1)}$  and  $q^{(2)}$ . Then the condition is that the system moves between these positions in such a way that the integral

$$S = \int_{t_1}^{t_2} L(q, \dot{q}, t) dt \quad (2.1)$$

takes the least possible value.† The function  $L$  is called the *Lagrangian* of the system concerned, and the integral (2.1) is called the *action*.

The fact that the Lagrangian contains only  $q$  and  $\dot{q}$ , but not the higher derivatives  $\ddot{q}$ ,  $\ddot{\ddot{q}}$ , etc., expresses the result already mentioned, that the mechanical state of the system is completely defined when the co-ordinates and velocities are given.

Let us now derive the differential equations which solve the problem of minimising the integral (2.1). For simplicity, we shall at first assume that the system has only one degree of freedom, so that only one function  $q(t)$  has to be determined.

Let  $q = q(t)$  be the function for which  $S$  is a minimum. This means that  $S$  is increased when  $q(t)$  is replaced by any function of the form

$$q(t) + \delta q(t), \quad (2.2)$$

where  $\delta q(t)$  is a function which is small everywhere in the interval of time from  $t_1$  to  $t_2$ ;  $\delta q(t)$  is called a *variation* of the function  $q(t)$ . Since, for  $t = t_1$  and for  $t = t_2$ , all the functions (2.2) must take the values  $q^{(1)}$  and  $q^{(2)}$  respectively, it follows that

$$\delta q(t_1) = \delta q(t_2) = 0. \quad (2.3)$$

† It should be mentioned that this formulation of the principle of least action is not always valid for the entire path of the system, but only for any sufficiently short segment of the path. The integral (2.1) for the entire path must have an extremum, but not necessarily a minimum. This fact, however, is of no importance as regards the derivation of the equations of motion, since only the extremum condition is used.

The change in  $S$  when  $q$  is replaced by  $q + \delta q$  is

$$\int_{t_1}^{t_2} L(q + \delta q, \dot{q} + \delta \dot{q}, t) dt - \int_{t_1}^{t_2} L(q, \dot{q}, t) dt.$$

When this difference is expanded in powers of  $\delta q$  and  $\delta \dot{q}$  in the integrand, the leading terms are of the first order. The necessary condition for  $S$  to have a minimum† is that these terms (called the *first variation*, or simply the *variation*, of the integral) should be zero. Thus the principle of least action may be written in the form

$$\delta S = \delta \int_{t_1}^{t_2} L(q, \dot{q}, t) dt = 0, \quad (2.4)$$

or, effecting the variation,

$$\int_{t_1}^{t_2} \left( \frac{\partial L}{\partial q} \delta q + \frac{\partial L}{\partial \dot{q}} \delta \dot{q} \right) dt = 0.$$

Since  $\delta \dot{q} = d\delta q/dt$ , we obtain, on integrating the second term by parts,

$$\delta S = \left[ \frac{\partial L}{\partial \dot{q}} \delta q \right]_{t_1}^{t_2} + \int_{t_1}^{t_2} \left( \frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right) \delta q dt = 0. \quad (2.5)$$

The conditions (2.3) show that the integrated term in (2.5) is zero. There remains an integral which must vanish for all values of  $\delta q$ . This can be so only if the integrand is zero identically. Thus we have

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}} \right) - \frac{\partial L}{\partial q} = 0.$$

When the system has more than one degree of freedom, the  $s$  different functions  $q_i(t)$  must be varied independently in the principle of least action. We then evidently obtain  $s$  equations of the form

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = 0 \quad (i = 1, 2, \dots, s). \quad (2.6)$$

These are the required differential equations, called in mechanics *Lagrange's equations*.‡ If the Lagrangian of a given mechanical system is known, the equations (2.6) give the relations between accelerations, velocities and co-ordinates, i.e. they are the equations of motion of the system.

† Or, in general, an extremum.

‡ In the calculus of variations they are Euler's equations for the formal problem of determining the extrema of an integral of the form (2.1).

# Connection Between Likelihood and Lagrangian

---

- **Maximum Likelihood trajectory of minimum free energy path distribution**

- Negative log-likelihood and Hamiltonian density

$$-\log P_{max}^V[x(t)] = \int_{t_0}^t d\tau \left( V[x(\tau)] + \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau) \right) = \int_{t_0}^t \tilde{\mathcal{H}}(\tau, x(\tau), \dot{x}(\tau)) d\tau$$

$$\tilde{\mathcal{H}}(\tau, x(\tau), \dot{x}(\tau)) = V[x(\tau)] + \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau)$$

- Conjugate momentum and Legendre transform

$$p = \frac{\partial \tilde{\mathcal{H}}}{\partial \dot{x}} = \mathbf{C}^{-1}[x] \dot{x}$$

$$\tilde{\mathcal{L}}(\tau, x(\tau), \dot{x}(\tau)) = p^T \dot{x} - \tilde{\mathcal{H}}(\tau, x(\tau), \dot{x}(\tau)) = -V[x(\tau)] + \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau)$$

- Lagrangian's associated action (for maximization, redefine the Hamiltonian flipped the sign)

$$\mathcal{H} = -\tilde{\mathcal{H}}, \quad \mathcal{L} = -\tilde{\mathcal{L}} = V[x(\tau)] - \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau)$$

$$\mathcal{A} = \int_{t_0}^t \mathcal{L}(\tau, x(\tau), \dot{x}(\tau)) d\tau = \int_{t_0}^t V[x(\tau)] - \frac{1}{2} \dot{x}(\tau)^T \mathbf{C}^{-1}[x(\tau)] \dot{x}(\tau) d\tau$$



# Diffusive Gradient Descent

---

- **Maximum Likelihood trajectory of minimum free energy path distribution (FEPD)**

- Assume that the potential is differentiable, then by using the Euler-Lagrange equations

$$0 = \nabla_x \mathcal{L} - \frac{d}{dt} [\nabla_{\dot{x}} \mathcal{L}]$$

- The maximum likelihood dynamics of a system whose trajectories satisfy our minimum FEPD

$$\ddot{x}(t) = -\mathbf{C}[x(t)] \left[ \nabla_x V[x(t)] + \frac{1}{2} \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t) \right]$$

- ◆ In directions of descent for the potential (go to low potential)
- ◆ In directions that increase the system's autocovariance statistics (controllability)
- Two assumptions to simplify
  - ◆ Our measure of temporal correlations varies slowly over space  $\nabla_x \mathbf{C}[x(t)] \approx 0$
  - ◆ Our system dynamics are LTV

$$\ddot{x}(t) = -\mathbf{C}[x(t)] \nabla_x V[x(t)] = -W(t, t_0) \nabla_x V[x(t)]$$

See Supplement

# Supple : Connection Between Likelihood and Lagrangian

---

- **Maximum Likelihood trajectory of minimum free energy path distribution (FEPD)**

- Assume that the potential is differentiable, then by using the Euler-Lagrange equations

$$0 = \nabla_x \mathcal{L} - \frac{d}{dt} [\nabla_{\dot{x}} \mathcal{L}]$$

$$\begin{aligned} 0 &= \nabla_x V[x(t)] - \frac{1}{2} \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t) - [-\ddot{x}(t)^T \mathbf{C}^{-1}[x(t)] - \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t)] \\ &= \nabla_x V[x(t)] + \ddot{x}(t)^T \mathbf{C}^{-1}[x(t)] + \frac{1}{2} \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t) \end{aligned}$$

- The maximum likelihood dynamics of a system whose trajectories satisfy our minimum FEPD

$$\begin{aligned} \ddot{x}(t)^T \mathbf{C}^{-1}[x(t)] &= \nabla_x V[x(t)] + \frac{1}{2} \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t) \\ \ddot{x}(t) &= -\mathbf{C}[x(t)] \left[ \nabla_x V[x(t)] + \frac{1}{2} \dot{x}(t)^T \nabla_x \mathbf{C}^{-1}[x(t)] \dot{x}(t) \right] \end{aligned}$$

$$\nabla_x \mathbf{C}^{-1}[x(t)] = -\mathbf{C}^{-1}[x(t)] [\nabla_x \mathbf{C}^{-1}[x(t)]] \mathbf{C}^{-1}[x(t)]$$

# Controllable Diffusive State Exploration

---

- *Analysis of our dynamics* ( $\gamma = 0$ )

$$\ddot{x}(t) + \gamma \dot{x}(t) + \mathbf{C}[x(t)] \nabla_x V[x(t)] = 0$$

- The absence of damping term
  - ◆ Any physical system approximately satisfying maximally diffusive trajectory statistics will experience **dissipation** (effectively damping).
- The descent directions that optimize the potential is affected by its controllability
  - ◆ Controllable agents can minimize arbitrary potentials merely through diffusive state exploration



### **3. Synthesizing Maximally Diffusive Trajectories**

# Directed Exploration

---

- ***Theoretical Agent***
  - Experiences spontaneously satisfy the path statistics of a maximally diffusive stochastic control process
- ***Real Agent***
  - The autonomous dynamics of control systems doesn't satisfy statistics
  - An approach from which to synthesize controllers that generate maximally diffusive trajectories



# **3.1 Maximally Diffusive Trajectories via KL control**

# KL divergence for Policies and Controllers

---

- **Kullback-Leibler (KL) Control**

- Define a path probability density for an arbitrary stochastic control process  $P_{u(t)}[x(t)]$
- Policies and controllers that minimize the KL divergence

$$\operatorname{argmin}_{u(t)} D_{KL}(P_{u(t)}[x(t)] \| P_{max}^V[x(t)])$$

◆ Support of  $P_{max}^V[x(t)]$  is infinite and assumed that  $\mathcal{X}$  is a compact domain

- Control synthesis problem from KL divergence

$$\operatorname{argmin}_{u(t)} \langle V[x(t)] \rangle_{P_{u(t)}} + D_{KL}(P_{u(t)}[x(t)] \| P_{max}[x(t)])$$

- From potential to cost function (**running cost**  $l$ ,  $L[x(t), u(t)] = \int_{\mathcal{T}} l(x(t), u(t)) dt$ )

$$\operatorname{argmin}_{u(t)} \mathbb{E}_{P_{u(t)}}[L[x(t), u(t)]] + D_{KL}(P_{u(t)}[x(t)] \| P_{max}[x(t)])$$

See Supplement

# Supple : Connection Between Likelihood and Lagrangian

---

$$\operatorname{argmin}_{u(t)} D_{KL}(P_{u(t)}[x(t)] \| P_{max}^V[x(t)])$$

$$P_{max}^V[x(t)] = P_{max}[x(t)] e^{-\int_{t_0}^t d\tau V[x(\tau)]}$$

$$\begin{aligned} D_{KL}(P_{u(t)}[x(t)] \| P_{max}^V[x(t)]) &= \int_{\mathcal{X}} P_{u(t)}[x(t)] \log \frac{P_{u(t)}[x(t)]}{P_{max}^V[x(t)]} \mathcal{D}x(t) \\ &= \int_{\mathcal{X}} P_{u(t)}[x(t)] [\log P_{u(t)}[x(t)] - \log P_{max}^V[x(t)]] \mathcal{D}x(t) \\ &= \int_{\mathcal{X}} P_{u(t)}[x(t)] \left[ \log P_{u(t)}[x(t)] - \log P_{max}[x(t)] + \int_{t_0}^t d\tau V[x(\tau)] \right] \mathcal{D}x(t) \\ &= \langle V[x(t)] \rangle_{P_{u(t)}} + D_{KL}(P_{u(t)}[x(t)] \| P_{max}[x(t)]) \end{aligned}$$



# Temperature-like Parameter

---

- ***Introducing Temperature-like parameter  $\alpha > 0$***

$$\operatorname{argmin}_{u(t)} \langle V[x(t)] \rangle_{P_{u(t)}} + \alpha D_{KL}(P_{u(t)}[x(t)] \| P_{max}[x(t)])$$

- Which optimizes task performance
  - Which optimizes the statistics of the system's state space diffusion
  - Solving the task with thorough exploration of the cost landscape
- 
- ***Exploration and Exploitation***
    - Theoretically, no formal trade-off between exploration and exploitation asymptotically
    - Practical use in balancing between exploration and exploitation

# Asymptotically I.I.D sampling

---

- **When  $P_{u(t)}[x(t)]$  ( $\approx P_{max}[x(t)]$ ) is maximally diffusive statistics,**

- Bias-minimizing estimator of the cost function (like unbiased estimator)

$$\mathbb{E}_{P_{u(t)}}[L[x(t), u(t)]] \approx \mathbb{E}_{P_{max}}[L[x(t), u(t)]]$$

- Thanks to ergodicity of  $P_{max}[x(t)]$ , estimator is equivalent to i.i.d. sampling of state-action cost

◆ We will prove next section

- Useful for known cost (or reward) function



## **3.2 Maximally Diffusive Trajectories via stochastic optimal control**

# KL Control Problem is SOC problem

---

- ***SOC problem : discrete time formulation for the SOC objective***

- Framed as Markov decision processes (MDPs) where the objective is to find **a policy**

$$\pi^* = \operatorname{argmin}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_{\pi}} \left[ \sum_{t=1}^N \gamma^t l(x_t, u_t) \right]$$

- MDP is a 5-tuple  $(\mathcal{X}, \mathcal{U}, p, r, \gamma)$  with state space  $\mathcal{X}$  and action space  $\mathcal{U}$
- The transition probability density  $p : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \rightarrow [0, \infty)$  with  $x_t, x_{t+1} \in \mathcal{X}$  and  $u_t \in \mathcal{U}$
- Discretized running cost :  $l$
- The environment emits a bounded loss  $l : \mathcal{X} \times \mathcal{U} \rightarrow [l_{min}, l_{min}]$  discounted at a rate  $\gamma \in [0, 1)$

- ***Redefining path distribution to include actions***

- Framed as Markov decision processes (MDPs) where the objective is to find **a policy**

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

- Action dependence because the maximally diffusive path distribution is action-independent

# KL Control Problem under the Pontential

---

- *Synthesized controller under the potential*

$$P_{max}^l[x_{1:N}, u_{1:N}] = \prod_{t=1}^N p_{max}(x_{t+1}|x_t, u_t) e^{-l(x_t, u_t)}$$
$$\operatorname{argmin}_{\pi} D_{KL}(P_{\pi}[x_{1:N}, u_{1:N}] || P_{max}^l[x_{1:N}, u_{1:N}])$$

- KL control = SOC MaxDiff trajectory synthesis problem

$$\pi_{MaxDiff}^* = \operatorname{argmin}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_{\pi}} \left[ \sum_{t=1}^N \gamma^t \hat{l}(x_t, u_t) \right]$$
$$\hat{l}(x_t, u_t) = l(x_t, u_t) + \alpha \log \frac{p(x_{t+1}|x_t, u_t) \pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t, u_t)}$$

$r(x_t, u_t) = -l(x_t, u_t)$ $\hat{r}(x_t, u_t)$
--

See Supplement

# Supple : KL divergence on controller under the potential

---

$$\operatorname{argmin}_{\pi} D_{KL}(P_{\pi}[x_{1:N}, u_{1:N}] \| P_{max}^l[x_{1:N}, u_{1:N}])$$

$$P_{max}^V[x(t)] = P_{max}[x(t)] e^{-\int_{t_0}^t d\tau V[x(\tau)]}$$

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

- Introducing temperature-like parameter  $\alpha > 0$

$$D_{KL}(P_{\pi}[x_{1:N}, u_{1:N}] \| P_{max}^l[x_{1:N}, u_{1:N}]) = \mathbb{E}_{P_{\pi}} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \frac{p(x_{t+1}|x_t, u_t) \pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t, u_t)} \right]$$

- Temperature-like parameter  $\alpha > 0$  : scaling costs or rewards by  $1/\alpha$



## **3.3 Maximum Diffusion Reinforcement Learning**

# Fully Controllability

- **Fully controllability**

**Definition 3.1.** 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,  $\pi : \mathcal{U} \times \mathcal{X} \rightarrow [0, \infty)$ , such that:

$$p_\pi(x_{t+1}|x_t) = E_{u_t \sim \pi(\cdot|x_t)}[p(x_{t+1}|x_t, u_t)] \quad (49)$$

and

$$D_{KL}(p_\pi(x_{t+1}|x_t) \parallel \nu(x_{t+1}|x_t)) = 0, \quad \forall t \in \mathbb{Z}^+ \quad (50)$$

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

- **Full controllability** implies that any desired state transition probability can be realized through a policy.
- **Agent's state transition dynamics due to a policy**
  - Stochastic case

$$p_\pi(x_{t+1}|x_t) = \int_{\mathcal{U}} p(x_{t+1}|x_t) \pi(u_t|x_t) du_t$$

- Deterministic case

$$p_\pi(x_{t+1}|x_t) = \int_{\mathcal{U}} p(x_{t+1}|x_t) \pi(u_t|x_t) \delta(u_t - \tau_\pi(x_t)) du_t = p_\pi(x_{t+1}|x_t, \tau_\pi(x_t))$$



# The relationship between MaxDiff RL and MaxEnt RL

- **MaxDiff RL and MaxEnt RL with full controllability**

**Theorem 3.1.** (Theorem 1 of Main Text) Let the state transition dynamics due to a policy  $\pi$  be  $p_\pi(x_{t+1}|x_t)$ . If the state transition dynamics are assumed to be decorrelated, then the optimum of Eq. 48 is reached when  $D_{KL}(p_\pi||p_{max}) = 0$  and the MaxDiff RL objective reduces to the MaxEnt RL objective.

$$\pi_{MaxDiff}^* = \operatorname{argmin}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_\pi} \left[ \sum_{t=1}^N \gamma^t \hat{l}(x_t, u_t) \right] \leq \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \pi(u_t|x_t) + \alpha D_{KL}(p_\pi(x_{t+1}|x_t, u_t) || p_{max}(x_{t+1}|x_t, u_t)) \right]$$

- Full controllability allows an agent to fully decorrelate transitions by matching the maximally diffusive trajectory

$$D_{KL}(p_\pi(x_{t+1}|x_t, u_t) || p_{max}(x_{t+1}|x_t, u_t)) \sim 0 : \text{MaxDiff} \rightarrow \text{MaxEnt}$$

- MaxDiff RL is a strict generalization of MaxEnt RL to agents with temporally correlated experiences

See Supplement

## Supple : Theorem 3.1

**Theorem 3.1.** (Theorem 1 of Main Text) Let the state transition dynamics due to a policy  $\pi$  be  $p_\pi(x_{t+1}|x_t)$ . If the state transition dynamics are assumed to be decorrelated, then the optimum of Eq. 48 is reached when  $D_{KL}(p_\pi||p_{max}) = 0$  and the MaxDiff RL objective reduces to the MaxEnt RL objective.

- Neglecting the discounting factor  $\gamma$

$$\begin{aligned}
 \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N \hat{l}(x_t, u_t) \right] &= \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \frac{p(x_{t+1}|x_t, u_t) \pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t, u_t)} \right] \\
 &= \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) \right] + \sum_{t=1}^N \mathbb{E}_{(x_t, u_t) \sim p, \pi} \left[ \alpha \log \frac{p(x_{t+1}|x_t, u_t) \pi(u_t|x_t)}{p_{max}(x_{t+1}|x_t, u_t)} \right] \quad p_\pi(x_{t+1}|x_t) = \int_u p(x_{t+1}|x_t) \pi(u_t|x_t) du_t \\
 &= \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \pi(u_t|x_t) \right] + \sum_{t=1}^N \mathbb{E}_{(x_t, u_t) \sim p, \pi} \left[ \alpha \log \frac{p(x_{t+1}|x_t, u_t)}{p_{max}(x_{t+1}|x_t, u_t)} \right]
 \end{aligned}$$

- Jensen's inequality (It is good to imagine convex function)

$$\mathbb{E}_{(x_t, u_t) \sim p, \pi} \left[ \log \frac{p(x_{t+1}|x_t, u_t)}{p_{max}(x_{t+1}|x_t, u_t)} \right] \leq \mathbb{E}_{x_t \sim p} \left[ \log \frac{\mathbb{E}_{u_t \sim \pi} [p(x_{t+1}|x_t, u_t)]}{p_{max}(x_{t+1}|x_t, u_t)} \right]$$

- As the synthesized controller goes maximum diffusive trajectory,  $\mathbb{E}_{u_t \sim \pi} [p_{max}(x_{t+1}|x_t, u_t)] = p_{max}(x_{t+1}|x_t, u_t)$  for unbiased estimator

## Supple : Theorem 3.1 (cont.)

---

$$\mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N \hat{l}(x_t, u_t) \right] \leq \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \pi(u_t | x_t) \right] + \sum_{t=1}^N \mathbb{E}_{x_t \sim p} \left[ \alpha \log \frac{\mathbb{E}_{u_t \sim \pi} [p(x_{t+1} | x_t, u_t)]}{p_{\max}(x_{t+1} | x_t, u_t)} \right]$$

$$p_\pi(x_{t+1} | x_t) = E_{u_t \sim \pi(\cdot | x_t)} [p(x_{t+1} | x_t, u_t)] \quad (49)$$

$$\begin{aligned} &= \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \pi(u_t | x_t) \right] + \sum_{t=1}^N \mathbb{E}_{x_t \sim p} \left[ \alpha \log \frac{p_\pi(x_{t+1} | x_t, u_t)}{p_{\max}(x_{t+1} | x_t, u_t)} \right] \\ &= \mathbb{E}_{P_\pi} \left[ \sum_{t=1}^N l(x_t, u_t) + \alpha \log \pi(u_t | x_t) + \alpha D_{KL}(p_\pi(x_{t+1} | x_t, u_t) \| p_{\max}(x_{t+1} | x_t, u_t)) \right] \end{aligned}$$

# Comments of Relation between Entropy and Decorrelation

---

- Under the assumption of decorrelated state, Jensen's inequality goes equality because of removing KL divergence

$$\pi^* = \operatorname{argmin}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_{\pi}} \left[ \sum_{t=1}^N \gamma^t \hat{l}_c(x_t, u_t) \right], \quad \hat{l}_c(x_t, u_t) = l(x_t, u_t) + \alpha \log \pi(u_t | x_t)$$

## ▪ *Problem of MaxEnt*

- Maximization by replacing the cost with a reward function

$$\hat{r}_c(x_t, u_t) = r(x_t, u_t) + \alpha \mathcal{H}(\pi(u_t | x_t)), \quad \mathcal{H}(\pi(u_t | x_t)) = -\pi(u_t | x_t) \log \pi(u_t | x_t)$$

- **Maximizing policy entropy doesn't decorrelate state transitions in general because maximizing policy entropy does not minimize KL divergence**

$$p_{\pi}(x_{t+1} | x_t) = \int_{\mathcal{U}} \mathbf{p}(x_{t+1} | x_t) \pi(u_t | x_t) du_t \quad \text{By environment's dynamics, } p(x_{t+1} | x_t) \text{ can have peak}$$

- MaxDiff RL objective continues to prioritize effective exploration by decorrelating state transitions and encouraging the system to realize maximally diffusive trajectories

# PAC-MDP and Single-shot

- **Probably Approximately Correct in Markov Decision Process (PAC-MDP) framework**

**Definition 3.2.** An algorithm  $\mathcal{A}$  is said to be PAC-MDP (Probably Approximately Correct in Markov Decision Processes) if, for any  $\epsilon > 0$  and  $\delta \in (0, 1)$ , a policy  $\pi$  can be produced with  $\text{poly}(|\mathcal{X}|, |\mathcal{U}|, 1/\epsilon, 1/\delta, 1/(1 - \gamma))$  sample complexity that is at least  $\epsilon$ -optimal with probability at least  $1 - \delta$ . In other words, if  $\mathcal{A}$  satisfies

$$\Pr(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi}(x_0) \leq \epsilon) \geq 1 - \delta \quad (57)$$

with polynomial sample complexity for all  $x_0 \in \mathcal{X}$ , where

$$\mathcal{V}_{\pi}(x_t) = E_{p, \pi} \left[ \sum_{n=0}^{\infty} \gamma^n r(x_{n+t}, u_{n+t}) \middle| x_t = x \right] \quad (58)$$

is the value function and  $\mathcal{V}_{\pi^*}(x_t)$  is the optimal value function, then  $\mathcal{A}$  is PAC-MDP.

- It is capable of learning a policy with polynomial sample complexity that can get arbitrarily close to the optimal policy with arbitrarily high probability.
- **Multi-shot and single-shot**
  - Multi-shot : agents are reset with randomized initial conditions after each episode  
*learning outside of episodic environments is crucial to real-world applications*
  - **Single-shot** : In non-episodic tasks, agents learn continuously without resets

For any algorithm  $\mathcal{A}$

# Formal properties of MaxDiff RL agents

**Theorem 3.2.** (Theorem 3 of Main Text) If there exists a PAC-MDP algorithm  $\mathcal{A}$  with policy  $\pi^{max}$  for the MaxDiff RL objective (Eq. 48), then the Markov chain induced by  $\pi^{max}$  is ergodic, and any individual initialization of  $\mathcal{A}$  will asymptotically satisfy the same  $\epsilon$ -optimality as an ensemble of initializations.

$$D_{KL}(p_{\pi^{max}} \| p_{max}) \approx 0$$

**For a maximally diffusive trajectory policy, the induced state-transition Markov chain is ergodic**

**Theorem 3.3.** (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  $\rho$  and let  $f : \mathcal{X} \rightarrow \mathbb{R}$  be any measurable function with  $E[|f(x)|] < \infty$ . Then, one has

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T f(x_t) = E_{x_0 \sim \rho}[f(x_0)] \quad (59)$$

almost surely.

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

**Ergodic Markov chain : PAC-MDP during multi-shot  $\rightarrow$  PAC-MDP during single-shot, too**

**Best possible alternative to i.i.d. sampling by ergodicity of Markovian trajectories**

**Theorem 3.4.** (Theorem 2 of Main Text) If there exists a PAC-MDP algorithm  $\mathcal{A}$  with policy  $\pi^{max}$  for the MaxDiff RL objective (Eq. 48), then the Markov chain induced by  $\pi^{max}$  is ergodic, and  $\mathcal{A}$  will be asymptotically  $\epsilon$ -optimal regardless of initialization.

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

**Robust to random seeds and environmental initializations  
Under Ergodic Markov chain**

See Supplements

## Supple : Theorem 3.2

---

**Theorem 3.2.** *(Theorem 3 of Main Text) If there exists a PAC-MDP algorithm  $\mathcal{A}$  with policy  $\pi^{max}$  for the MaxDiff RL objective (Eq. 48), then the Markov chain induced by  $\pi^{max}$  is ergodic, and any individual initialization of  $\mathcal{A}$  will asymptotically satisfy the same  $\epsilon$ -optimality as an ensemble of initializations.*

For path distribution when the policy is optimal  $p_{\pi^{max}}(x_{t+1}|x_t) = \int_{\mathcal{U}} p(x_{t+1}|x_t) \pi^{max}(u_t|x_t)$

- Since  $\mathcal{A}$  is capable of producing an  $\epsilon$  – optimal policy  $\pi^{max}$ , so  $D_{KL}(p_{\pi^{max}} \| p_{max}) \approx 0$

**Theorem 2.2.** *A stochastic control process (Definition 2.3) in a compact and connected space  $\mathcal{X} \subset \mathbb{R}^d$  with a maximum entropy exploration strategy in a potential (in the sense of Eq. 31) is ergodic.*

- Since  $\mathcal{A}$  is MDP, it is naturally Markovian.
- By Theorem 2.2,  $p_{\pi^{max}}$  is ergodic.



## Supple : Theorem 3.3

**Theorem 3.3.** (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  $\rho$  and let  $f : \mathcal{X} \rightarrow \mathbb{R}$  be any measurable function with  $E[|f(x)|] < \infty$ . Then, one has

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T f(x_t) = E_{x_0 \sim \rho}[f(x_0)] \quad (59)$$

almost surely.

- About an ergodic Markov chain, the time average of any function is equal to its ensemble average
- From definition of PAC-MDP

$$\Pr(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{\max}}(x_0) \leq \epsilon) \geq 1 - \delta$$

$$\mathbb{E}_{x_0 \sim \rho}[\mathbf{1}(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{\max}}(x_0) \leq \epsilon)] \geq 1 - \delta$$

$\rho$  : stationary distribution

$\mathbf{1}$  : indicator function

- At least  $\epsilon$ -optimal on average at least  $100 \times (1 - \delta)\%$  of episodes

$$f(x_t) = \mathbf{1}\{(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{\max}}(x_0) \leq \epsilon)\}$$

- Birkhoff's theorem

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T \mathbf{1}\{(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{\max}}(x_0) \leq \epsilon)\} = E_{x_0 \sim \rho}[\mathbf{1}\{(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{\max}}(x_0) \leq \epsilon)\}]$$

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



## Supple : Theorem 3.4

**Theorem 3.4.** (Theorem 2 of Main Text) *If there exists a PAC-MDP algorithm  $\mathcal{A}$  with policy  $\pi^{max}$  for the MaxDiff RL objective (Eq. 48), then the Markov chain induced by  $\pi^{max}$  is ergodic, and  $\mathcal{A}$  will be asymptotically  $\epsilon$ -optimal regardless of initialization.*

- Let  $f(x_t) = \mathbf{1}\{(\mathcal{V}_{\pi^*}(x_0) - \mathcal{V}_{\pi^{max}}(x_0) \leq \epsilon)\}$
- Let  $\{x_t\}_{t \in \mathbb{N}}, \{x'_t\}_{t \in \mathbb{N}}$  be ergodic Markov chain with identical transition kernels given by  $p_{\pi^{max}}$
- Different initial condition,  $x_0, x'_0 \in \mathcal{X}$
- By Birkhoff's ergodic theorem, the same unique ensemble average over the invariant measure  $\rho$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T f(x_t) = E_{x_0 \sim \rho}[f(x_t)], \quad \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T f(x'_t) = E_{x_0 \sim \rho}[f(x'_t)]$$

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

# Model-based of MaxDiff RL agents

---

- *Thanks to Maximally diffusive trajectories (Markov chain) and ergodicity*
  - Single-shot (non-episode)
  - Robustness to initial conditions and random seeds
  - Model-based RL : dependence on the system's state transition dynamics
- *Model-free MaxDiff RL*
  - It can be possible to extend our results to model-free algorithms by reformulating the objective function



## **3.4 Maximum Diffusion Reinforcement Learning**

# Model-based of MaxDiff RL agents

---

- **Approach : Maximize the path entropy of trajectories induced by policy, not policy entropy**

$$\forall S[P_{max}[x(t)]] \geq S[P_{u(t)}[x(t)]] \text{ with equality if and only if } P_{max}[x(t)] = P_{u(t)}[x(t)]$$

$$\operatorname{argmax}_{u(t)} S(P_{u(t)}[x(t)])$$

- Path based and Policy based MaxDiff RL

We can think of a controller as a policy given by a Dirac delta distribution centered at  $u_t$ .

$$\begin{aligned} \max_{u(t)} S[P_{u(t)}[x(t)]], \quad & \max_{u_{1:N-1}} S\left[\prod_{t=1}^N p(x_{t+1}|x_t, u_t)\right] \\ \max_{\pi} S[P_{\pi}[x(t), u(t)]], \quad & \max_{\pi} S\left[\prod_{t=1}^N p(x_{t+1}|x_t, u_t)\pi(x_t, u_t)\right] \end{aligned}$$

- KL control = SOC MaxDiff

$$\begin{aligned} \operatorname{argmin}_{u(t)} \mathbb{E}_{P_{u(t)}} [L[x(t), u(t)] - \alpha S[P_{u(t)}[x(t)]]], \quad & \operatorname{argmin}_{\pi} \mathbb{E}_{P_{\pi}} [L[x(t), u(t)] - \alpha S[P_{\pi}[x(t), u(t)]]] \\ \operatorname{argmin}_{u_{1:N-1}} \mathbb{E}_{P_{u_{1:N}}} [l(x_t, u_t) - \alpha S[p(x_{t+1}|x_t, u_t)]], \quad & \operatorname{argmin}_{\pi} \mathbb{E}_{P_{\pi}} [l(x_t, u_t) - \alpha S[p(x_{t+1}|x_t, u_t)\pi(u_t|x_t)]] \end{aligned}$$



## **3.5 Simplified Synthesis via Local Entropy Maximization**

# Local Entropy Maximization

---

- **Assuming it is locally close to the optimal statistics**

$$\forall S[P_{max}[x(t)]] \geq S[P_{u(t)}^L[x(t)]] \text{ with equality if and only if } P_{max}[x(t)] = P_{u(t)}^L[x(t)]$$

- Path probability densities

$$P_{u(t)}^L[x(t)] = \frac{1}{Z} \exp \left[ -\frac{1}{2} \int_{t_0}^t \dot{x}(\tau)^T \mathbf{C}^{-1}(x(\tau)) \dot{x}(\tau) d\tau \right]$$

- Chain rule on the path entropy along a finite path (analytic form of the entropy of a Gaussian distribution)

$$S[P[x_{1:N}]] = \sum_{t=1}^N S[p(x_{t+1}|x_{1:t})] \propto \frac{1}{2} \sum_{t=1}^N \log \det \mathbf{C}[x_t]$$

- So, **MaxDiff RL objective satisfying local maximum with reward function  $r(x_t, u_t)$**

$$\operatorname{argmax}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_{\pi}} \left[ \sum_{t=1}^N r(x_t, u_t) + \frac{\alpha}{2} \log \det \mathbf{C}[x_t] \right]$$

# Local Entropy Maximization in Practical (1)

---

- $C[x_t]$  *may be not full-rank in real*

$$\det C[x_t] = 0 \rightarrow \log \det C[x_t] \sim \infty$$

- Create numerical stability issues
- **Solution 1 : Leading  $M$  eigenvalues**

$$\sum_{i=1}^M \log \lambda_i$$

- Good : no occurring numerical stability issues
- Bad : Restricting the exploration to an  $M$  dimensional subspace of  $\mathcal{X}$
- **Solution 2 : Logarithm of the trace**

$$\log \text{Tr}(C[x_t])$$

- Good : drastically reducing the computational complexity of the determinant in high dimensional systems.
- Bad : can only use when system states vary *independently*

# Local Entropy Maximization in Practical (2)

---

- **Model-free MaxDiff RL**

$$\operatorname{argmax}_{\pi} \mathbb{E}_{(x_{1:N}, u_{1:N}) \sim P_{\pi}} \left[ \sum_{t=1}^N r(x_t, u_t) + \alpha \log \det \mathcal{C}[x_t] \right]$$

- No explicitly depend on  $p(x_{t+1}|x_t, u_t)$  : model-free
- Environment property :  $\frac{1}{2} \log \det \mathcal{C}[x_t]$
- State-dependent property of the environment :  $S[x^*] = \frac{1}{2} \log \det \mathcal{C}[x_t]$

*Not require direct access to the state-transition dynamics*

- **Function approximation  $\hat{S}_{\theta}[x^*] \approx S[x^*]$**

- An estimate model-free implementation are possible by augmenting the value function with  $\hat{S}_{\theta}[x_t]$
- Data-driven estimates of  $\mathcal{C}[x^*]$  during its optimization
  - ◆ Empirically estimated from data,

$$\mathcal{C}[x^*] = \int_{t_i - \Delta t}^{t_i} K_{XX}(\tau, t_i) d\tau$$

*Guaranteed only under  
a stationary process*