

Connecting RL with LLM and Diffusion Models, via Regularities of (Rough) Differential Equations

Xin Guo, *UC Berkeley*

SURE-AI Star Midsummer on Reinforcement Learning
(Day Three)
June 23-25, 2026

Outline

Transfer Learning (TL) and LLM

Integrating TL with RL in LQR

RL and Robustness of score-based generative diffusion models

TL for general continuous-time RL

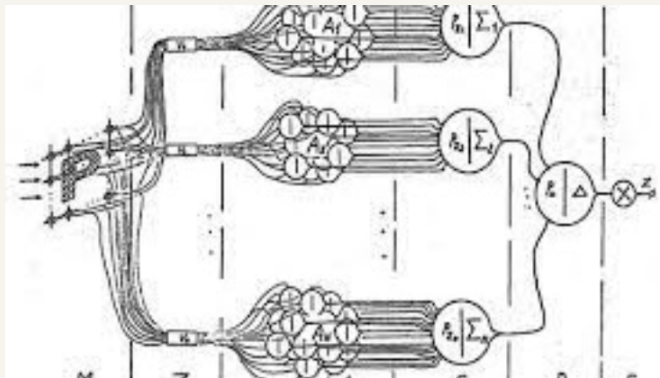
The lecture is based on

- ▶ G., Li, Xu (2025). Fast policy learning for linear quadratic control with entropy regularization. SIAM Control and Optimization, 2025. ArXiv:2311.14168.
- ▶ Cao, Gu., G., and Rosenbaum (2024). Transfer learning for portfolio optimization. Revised for *Mathematical Finance* arXiv:2307.13546
- ▶ G. and Lyu (2025). Policy transfer for continuous-time reinforcement learning: a (rough) differential equation approach, arxiv2510.15165

Transfer learning (TL)

- ▶ Natural language processing: BERT (Devlin et al. 2019)
- ▶ Popularity grows exponentially due to the intense interests in large language model, such as ChatGPT (2022)
- ▶ Well adopted in computer vision, natural language processing, medical and finance data applications

Transfer learning



From "Reminder of the First Paper on Transfer Learning in Neural Networks, 1976" by S. Bozinovski

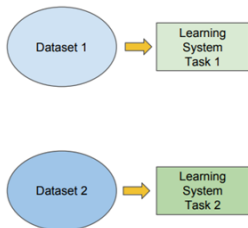
Informatica 44.3 (2020)

Transfer learning: an optimization/control view

- Exploiting knowledge learnt from previously learned tasks

Traditional ML

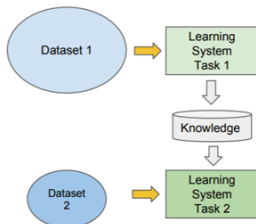
- Isolated, single task learning:
 - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks



vs

Transfer Learning

- Learning of a new tasks relies on the previous learned tasks:
 - Learning process can be faster, more accurate and/or need less training data



Transfer learning between Source and Target Tasks

- ▶ When there is **limited and relevant** data for a new task
- ▶ When there are **similarities** between the target task and the previously learned task
- ▶ Most works of transfer learning are in discrete-time and/or on image and language models
- ▶ Feature transfer and policy transfer are the most popular transfer learning techniques in large language models

Feature transfer

- ▶ The source task on the dataset $\{(x_S^i, y_S^i) \mid i = 1 \dots N_S\}$ with $y_S^i = f_S^*(x_S^i)$
- ▶ f_S^* is decomposed as: $f_S^* = g_S \circ h_S^*$
- ▶ The target task on $\{(x_T^i, y_T^i) \mid i = 1 \dots N_T\}$ with $y_T^i = f_T^*(x_T^i) = g_T \circ h_T^*$
- ▶ Feature transfer (during training): the feature extractors are shared such that

$$\hat{h}_S = \hat{h} = \hat{h}_T$$

- ▶ The goal is to learn and approximate g_T

Example: stock return prediction

(Cao, Gu, G., and Rosenbaum 2024)

- ▶ Raw data: path $z : [t - L + 1, t] \rightarrow \mathbb{R}^d$ with $z_\tau = (\tau, \log s_\tau, \log v_\tau)$, where $\{s_\tau\}_{\tau=t-L+1}^t$ and $\{v_\tau\}_{\tau=t-L+1}^t$ are historical daily close price and trading volume of given stocks over L days
- ▶ Data pre-processing: perform a (truncated) signature transform over z up to degree M : $x_t^{L,M}$
- ▶ Objective: predict the stock price at time $\tau = t + 1$, denoted by y_t , from $\{x_\tau^{L,M}\}_{\tau=t-L+1}^t$ by regression (via the universal nonlinearity property of signature transform)

Stock return prediction

Source task:

- ▶ Input and output data: $\{(x_{S,t}^{L,M}, y_{S,t})\}_{0 \leq t \leq H_S}$,
 $(x_{S,t}^{L,M}, y_{S,t}) \in \mathbb{R}^{d_{sig}} \times \mathbb{R}$, collected from source stock(s)
- ▶ Objective:

$$\hat{\theta}_S := \arg \min_{\theta} \frac{1}{H_S} \sum_{t=0}^{H_S-1} \left(x_{S,t}^{L,M} \cdot \theta - y_{S,t} \right)^2 + \lambda_S \|\theta\|_2^2$$

Target task:

- ▶ Input and output data: $\{(x_{T,t}^{L,M}, y_{T,t})\}_{0 \leq t \leq H_T}$,
 $(x_{T,t}^{L,M}, y_{T,t}) \in \mathbb{R}^{d_{sig}} \times \mathbb{R}$, from the target stock
- ▶ Objective

$$\hat{\theta}_T := \arg \min_{\theta} \frac{1}{H_T} \sum_{t=0}^{H_T-1} \left(x_{T,t}^{L,M} \cdot \theta - y_{T,t} \right)^2 + \lambda_S \|\theta - \hat{\theta}_S\|_2^2$$

Example: portfolio optimization

(Cao, Gu, G., and Rosenbaum 2024)

- ▶ Consider d assets, with returns $r = (r_1, \dots, r_d)^\top \sim \mathbb{P}$
- ▶ Portfolio allocation vector: $\phi = (\phi_1, \dots, \phi_d)^\top \in \mathbb{X}$
- ▶ Sharpe ratio: $\max_{\phi \in \mathbb{X}} \frac{\mathbb{E}^\mathbb{P}[\phi^\top r]}{\text{Var}(\phi^\top r)} = \max_{\phi \in \mathbb{X}} \frac{\mu_\mathbb{P}^\top \phi}{\phi^\top \Sigma_\mathbb{P} \phi}$

Source task:

$$\phi_S = \operatorname{argmax}_{\phi \in \mathbb{X}} \frac{\mu_S^\top \phi}{\sqrt{\phi^\top \Sigma_S \phi}}$$

Target task with transfer learning:

$$\phi_T = \operatorname{argmax}_{\phi \in \mathbb{X}} \frac{\mu_T^\top \phi}{\sqrt{\phi^\top \Sigma_T \phi}} - \lambda \|\phi_S - \phi\|_2^2$$

Policy transfer (or fine-tuning)

- ▶ The key idea is to use a good initialization for the target task $\hat{g}_T$:

$$\hat{f}_{T,0} = \hat{f}_S$$

- ▶ $\hat{f}_T$ is fine tuned iteratively using gradients on the target task based on loss functions

Integrating RL with transfer learning

Key idea: utilized simple (mostly LQ) models and algorithms for RL to more complex RL problems.

Questions

- ▶ the feasibility of policy transfer, and
- ▶ the benefit of policy transfer

for continuous-time reinforcement learning?

Keys: stability of optimal control

LQ-RL case

State $x_t \in \mathbb{R}^n$, action $u_t \in \mathbb{R}^k$:

$$dx_t = \left[A_t x_t + B_t \mathbf{E}(u_t | x_t) \right] dt + \sigma_t dW_t, \quad x_0 \sim \mathcal{D}_0, \quad (1)$$

$$\inf_{\pi \in \mathbf{A}} J_{\mathcal{D}_0}(\pi) := \mathbf{E} \left[\int_0^T x_t^\dagger Q_t x_t + u_t^\dagger R_t u_t + \tau \log h_t(u_t | x_t) dt + x_T^\dagger Q' x_T \right], \quad (2)$$

- ▶ W_t : d -dim BM, $\tau > 0$
- ▶ $(A_t, B_t, \sigma_t, Q_t, R_t, Q')$: appropriate deterministic matrices
- ▶ π : a stochastic policy
- ▶ $\mathbf{A}$ the set of admissible policies such that
 - the controlled SDE admits a unique strong solution;
 - the value of the cost function is finite

Stochastic policies and exploration for RL

- ▶ $\mathcal{P}(\mathcal{A})$: collection of probability density functions (pdfs) on $\mathcal{A}$
- ▶ $\pi : (t, x) \in [0, T] \times \mathbb{R}^d \rightarrow (\cdot|t, x) \in \mathcal{P}(\mathcal{A})$ a (stochastic) policy

Ingredient I: optimal stochastic policy is Gaussian

Lemma 1

The optimal stochastic policy of the LQR (1) – (2) is:

$$\pi_t^*(\cdot | x) = \mathcal{N} \left(-R_t^{-1} B_t^\dagger P_t^* x, \frac{\tau}{2} R_t^{-1} \right), \quad (3)$$

where P_t^* solves the following Riccati equation:

$$\frac{dP_t^*}{dt} + A_t^\dagger P_t^* + P_t^* A_t + Q_t - P_t^* B_t R_t^{-1} B_t^\dagger P_t^* = 0, \quad P_T^* = Q'. \quad (4)$$

(Wang, Zariphopoulou, and Zhou 2018; Giegrich, Reisinger, and Y. Zhang 2022)

Ingredient II: stability of Riccati equation

Key Lemma

Denote by $\mathcal{R}$ the solution map of the Riccati equation (4), i.e.,

$$\begin{aligned} \mathcal{R} : L^\infty([0, T], \mathbb{R}^{n \times n}) \times L^\infty([0, T], S_{\geq 0}^n) \times L^\infty([0, T], \mathbb{R}^{n \times k}) \\ \times L^\infty([0, T], S_{>0}^k) \times S_{\geq 0}^n \longrightarrow C([0, T], S_{\geq 0}^n) \\ (A_{t \in [0, T]}, Q_{t \in [0, T]}, B_{t \in [0, T]}, R_{t \in [0, T]}, Q') \longmapsto (P_t^*)_{t \in [0, T]} \end{aligned}$$

Then, $\mathcal{R}$ is continuous

Lemmas 1 and the key lemma yield:

Theorem (Transfer learning for LQRs (G. Lyu (2025)))

Given an LQR represented by model parameters θ , its optimal policy π^* , and an associated learning algorithm. Suppose there is another LQR with model parameters $\tilde{\theta}$. Then, there exists $\epsilon > 0$, such that with an appropriate initialization $\pi^{(0)}$, this learning algorithm has the same convergence rate and finds a near-optimal policy for the new LQR, provided that

$$d(\pi^{(0)}, \pi^*) + d(\tilde{\theta}, \theta) < \epsilon.$$

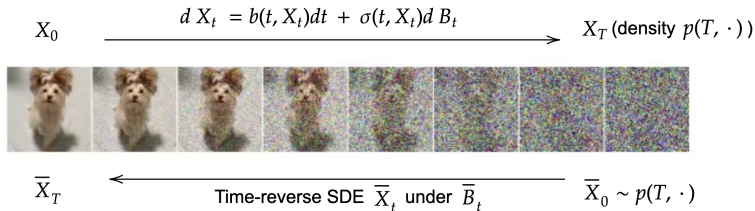
Here d denotes an appropriate distance on the associated metric space.

Mathematics behind diffusion models

Consider the (forward) dynamics :

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \text{ with } X_0 \sim p_{\text{data}}(\cdot)$$

where $p_{\text{data}}(\cdot)$ is the data distribution (access to limited samples)



1. Fix $T > 0$ (often large). Sample X_T (in some case we know the exact law of $(X_T|X_0)$)
2. Run the **time-reversed process** $\bar{X}$ starting at $\bar{X}_0 \sim p(T, \cdot)$, up to time T : $\bar{X}_T$ is a copy of $p_{\text{data}}(\cdot)$

Application to robustness of diffusion models

Question: What is $\bar{X}_t$? Is it simple?

Ingredient 1 (Anderson, '82; Haussmann and Pardoux, '86)

Let $p(t, \cdot)$ be the density of X_t , define

$$d\bar{X}_t = \bar{b}(t, \bar{X}_t)dt + \bar{\sigma}(t, \bar{X}_t)d\bar{B}_t, \text{ with } \bar{X}_0 \sim p(T, \cdot),$$

where $\bar{\sigma}(t, x) = \sigma(T - t, x)$ and

$$\bar{b}(t, x) = -b(T - t, x) + \sigma^2(T - t, x)\nabla \log p(T - t, x).$$

Then under mild conditions, (5) admits a strong solution and $\bar{X}_t$ has the same law as X_{T-t} .

Backward process $\bar{X}$

Note that b, σ are known (**part of the diffusion design**). The “headache” is $p(t, \cdot)$, or $\nabla \log p(t, \cdot)$ (**Stein's score**)!

Ingredient 2

Learn $\nabla \log p(t, \cdot)$ – score-matching (Ho, Jain, and Abbeel, '20; Song and Ermon, '19):

$$\min_{\theta} \mathbf{E}_{t, X_t} \|s_{\theta}(X_t, t) - \nabla \log p(X_t, t)\|_2^2$$

- ▶ $\mathbf{E}_{t, X_t}$ means $t \sim \text{Unif}[0, T]$, $X_0 \sim p_{\text{data}}$, and $X_t|X_0$ follows the forward SDE process
- ▶ Needs good estimation of $\nabla \log p$ **at each time t** to recover p_{data}

Score-based diffusion models

- Robustness of score-based diffusion models (SDMs):

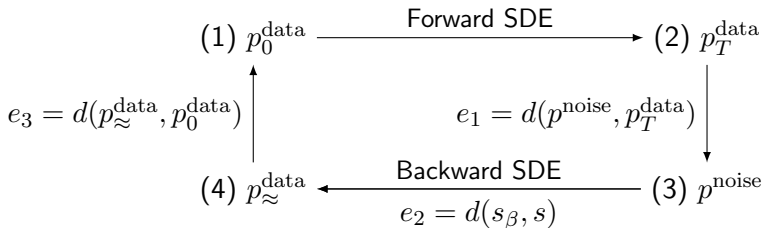


Figure: Approximation errors in score-based diffusion models.

- It is shown (Tang, Zhao (2024)) that under proper technical conditions,

$$e_3 \in O(e_1 + e_2)$$

Robustness of SDMs

- ▶ Robustness of SDMs $\iff$ Stability of control/policy
- ▶ Well-established connection between score-based diffusion models and stochastic control problems (B. J. Zhang and Katsoulakis 2023; B. J. Zhang, Liu, et al. 2024).

Connection with stochastic control problems (cont.)

Assumption 2

$$\mathrm{tr}(A_t) = -\frac{\tau}{4} \log \frac{|R_t|}{(\tau\pi)^k}, \quad B_t R_t^{-1} B_t^\dagger = \sigma_t \sigma_t^\dagger, \quad Q_t = 0$$

for any $t \in [0, T]$, and $Q' \succ 0$.

Lemma 4 (under Assumption 2)

The probability density function of the diffusion process on $[0, T]$

$$d\hat{X}_t = -A_{T-t}\hat{X}_t dt + \sigma_{T-t} dW_t, \quad \hat{X}_0 \sim \mathcal{N}(0, (Q')^{-1}) \quad (5)$$

can be expressed by:

$$\hat{p}^{Q'}(t, x) = (2\pi)^{-\frac{n}{2}} |Q'|^{\frac{1}{2}} \exp \left[-\frac{1}{2} J(T-t, x) \right],$$

- ▶ $J(t, x) = x^\dagger P_t^* x + r_t^*$ is the cost function of the LQR (1) – (2) associated with the optimal randomized policy
- ▶ the associated backward SDE is

$$\begin{aligned} d\hat{Y}_t &= \left[A_t \hat{Y}_t + \sigma_t \sigma_t^\dagger \nabla \log \hat{p}^{Q'}(T - t, \hat{Y}_t) \right] dt + \sigma_t dW_t, \\ \hat{Y}_0 &\sim \hat{p}^{Q'}(T, \cdot), \end{aligned}$$

where we use $\hat{p}^{Q'}$ to indicate the dependence of $\hat{p}$ on Q'

Stability of score-based diffusion models

Theorem (Error bound, G. and Lyu 2025)

Under proper regularity conditions, there exist constants $(C_1, C_2, C_3) > 0$, depending on the data of the LQR (1) – (2), such that for any $\epsilon > 0$, there exists $\delta_0 > 0$, such that $\|M - Q'\| < \delta_0$ implies

$$d_{\text{TV}}(Y_T, \hat{Y}_T) \leq d_{\text{TV}}(p^{\text{noise}}, \hat{p}^{Q'}(T, \cdot)) + C_1 \epsilon,$$

and

$$W_2(Y_T, \hat{Y}_T) \leq \sqrt{C_2 W_2^2(p^{\text{noise}}, \hat{p}^{Q'}(T, \cdot)) + C_3 \epsilon^2},$$

where Y_t satisfies backward SDE with score function $\nabla \log \hat{p}^M$, and $\hat{Y}_t$ satisfies the backward SDE with score function $\nabla \log \hat{p}^{Q'}$

Benefit of transfer learning for RL

Design of fast algorithm called IPO with global linear and local super-linear convergence (G., Li, Xu (2025))

- ▶ Focus on the Gaussian policy $\pi_{K,\Sigma}$ and the associated cost function $J^{K,\Sigma}(t, x)$
- ▶ Optimize the **one-step deviation** from the current policy in each iteration
 - The (optimal) value function $V^*(x) = \max_{K,\Sigma} J_{K,\Sigma}(x)$ satisfies

$$V^*(x) = \max_{(K,\Sigma)} \mathbb{E}_{u \sim \pi_{K,\Sigma}} \left[r(x, u) + \gamma V^*(Ax + Bu + w) \right]$$

- IPO updating rule (iteratively finding $V(x)$):

$$(K', \Sigma') = \arg \max_{\tilde{K}, \tilde{\Sigma}} \mathbb{E}_{u \sim \pi_{\tilde{K}, \tilde{\Sigma}}} \left[r(x, u) + \gamma \underbrace{V^{K,\Sigma}}_{\text{freezing } (K,\Sigma)} (Ax + Bu + w) \right]$$

Numerical experiments: transfer Learning using IPO

1. Parameters: $n = 400$, $k = 200$, A , B , L , R are generated randomly.
2. Performance measure: normalized error $= \frac{C(K, \Sigma) - C(K^*, \Sigma^*)}{C(K^*, \Sigma^*)}$.

Transfer Learning Setup. $\bar{A}$ and $\bar{B}$ are the state transition matrices which are generated by adding a perturbation to A and B :

$$\bar{A}_{i,j} = A_{i,j} + u_{i,j}, \quad \bar{B}_{i,j} = B_{i,j} + u'_{i,j}, \quad u_{i,j}, u'_{i,j} \sim \text{UNIF}[0, 10^{-2}].$$

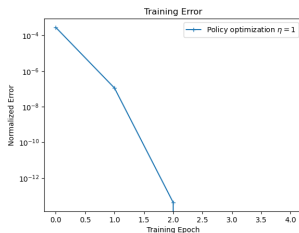


Figure: Transfer learning using IPO with $(\bar{K}^0, \bar{\Sigma}^0) = (K^*, \Sigma^*)$.

Policy transfer in general continuous-time RL

State $x_t \in \mathbb{R}^n$, feedback control $\alpha_t := \alpha(x_t) \in \mathbb{R}^k$, such that

$$\begin{aligned} dx_t &= \mu(t, x_t, \alpha_t)dt + \sigma(t, x_t)dW_t, \quad x_0 \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^n), \\ \inf_{\alpha \in \mathbf{A}} J &: \mathcal{P}(C_{[0,T]}) \longrightarrow \mathbb{R}, \end{aligned}$$

where $\mathcal{P}(C_{[0,T]})$ denotes the space of Borel probability measures on $C_{[0,T]}$, equipped with the weak convergence topology, and

$$C_{[0,T]} := C([0, T], \mathbb{R}^n; \|\cdot\|_{\infty; [0,T]}).$$

- ▶ $\mu \in \text{Lip}^1$, i.e. μ is bounded and Lipschitz continuous;
- ▶ $\sigma \in \text{Lip}^2$, i.e. σ is bounded, and has a Lip^1 first-order derivative;
- ▶ J is continuous;
- ▶ $\mathbf{A}$ is a compact subspace of $\text{Lip}(\mathbb{R}^n, \mathbb{R}^k; \|\cdot\|_{\infty})$.

Theorem (Policy transfer of general continuous-time RL)

Let $(\mu^i, \sigma^i, x_0^i) \in \text{Lip}^1 \times \text{Lip}^2 \times L^2$ ($i = 1, 2$) represent two general control problems. Then, the optimal control α^* of one of them is ϵ -optimal to the other as long as the distance between (μ^1, σ^1, x_0^1) and (μ^2, σ^2, x_0^2) is sufficiently small (i.e., under $\|\cdot\|_\infty \times \|\cdot\|_{\text{Lip}^1} \times \|\cdot\|_{L^2}$).

Key ingredient of analysis

Key: the continuity of the map

$$\begin{aligned} F : \mathcal{A} &\longrightarrow \mathbb{R}, \\ \alpha &\longmapsto J(\mathcal{L}(X_{[0,T]}^\alpha), \alpha), \end{aligned}$$

with $\mathcal{L}(X_{[0,T]}^\alpha)$ the law of the controlled state process, derived from

- ▶ equivalent rough SDE representation, and
- ▶ stability of rough SDE

(Lyons, Caruana, and Lévy 2007; Friz and Victoir 2010; Friz and Hairer 2014).

Lemma (equivalent RDE)

Let $\mu \in \text{Lip}^1(\mathbb{R}^n, \mathbb{R}^n)$, $\sigma \in \text{Lip}^2(\mathbb{R}^n, \mathbb{R}^{n \times d})$, and $y_0 \in \mathbb{R}^n$. Then, the unique strong solution of the Stratonovich diffusion SDE

$$dY_t = \mu(Y_t)dt + \sigma(Y_t) \circ dW_t, \quad Y_0 = y_0$$

is indistinguishable to the (almost surely well-defined and unique) solution of the RDE

$$dY_t = \mu(Y_t)dt + \sigma(Y_t)d\mathbf{W}_t, \quad Y_0 = y_0, \quad (6)$$

where $\mathbf{W}_t$ is the enhanced BM in the sense of Stratonovich

Lemma (stability of RDE translation)

Let $Y_0 \in L^2(\Omega, \mathbb{F}, \mathbb{P}; \mathbb{R}^n)$ be independent of $W_{[0,T]}$, $\mathcal{R}$ be the solution map of the Stratonovich SDE:

$$dY_t = \mu(Y_t)dt + \sigma(Y_t) \circ dW_t,$$

$$\begin{aligned} \mathcal{R} : \text{Lip}^1(\mathbb{R}^n, \mathbb{R}^n) \times \text{Lip}^2(\mathbb{R}^n, \mathbb{R}^{n \times d}) \times L^2(\Omega, \mathbb{F}, \mathbb{P}; \mathbb{R}^n) &\longrightarrow \mathcal{P}(C([0, T], \mathbb{R}^n)) \\ (\mu, \sigma, Y_0) &\longmapsto \mathcal{L}(Y_{[0,T]}) \end{aligned}$$

Then, $\mathcal{R}$ is continuous.

Here Lip^1 (Lip^2) is equipped with the $\|\cdot\|_{\infty; \text{compact}}$ ($\|\cdot\|_{\text{Lip}^1; \text{compact}}$) metric, and $\mathcal{P}(C([0, T], \mathbb{R}^n))$ equipped with the weak topology.

Remark on RDE approach

RDE approach allows for

- ▶ general payoff functions which can be path-dependent
- ▶ general processes beyond Brownian motion, for instance general Markov or Gaussian driven processes, and fractional Brownian motion

Open questions: General robustness studies of RDE?

Thank you!
Questions?

Additional Reference I

Friz, Peter K and Martin Hairer (2014). A course on rough paths.
Springer.

Friz, Peter K and Nicolas B Victoir (2010).
Multidimensional stochastic processes as rough paths: theory and app
Vol. 120. Cambridge University Press.

Giegrich, Michael, Christoph Reisinger, and Yufei Zhang (2022).
“Convergence of policy gradient methods for finite-horizon
stochastic linear-quadratic control problems”. In:
arXiv preprint arXiv:2211.00617.

Lyons, Terry J, Michael Caruana, and Thierry Lévy (2007).
Differential equations driven by rough paths: Ecole d'Eté de Probabilités
Springer.

Additional Reference II

Wang, Haoran, Thaleia Zariphopoulou, and Xunyu Zhou (2018). “Exploration versus exploitation in reinforcement learning: A stochastic control approach”. In: [arXiv preprint arXiv:1812.01552](#).

Zhang, Benjamin J and Markos A Katsoulakis (2023). “A mean-field games laboratory for generative modeling”. In: [arXiv preprint arXiv:2304.13534](#).

Zhang, Benjamin J, Siting Liu, et al. (2024). “Wasserstein proximal operators describe score-based generative models and resolve memorization”. In: [arXiv preprint arXiv:2402.06162](#).