Stochastic control and reinforcement learning

CIRM, Marseille

July, 7th to 10th 2026.

Schedule

Tuesday 7th.

- 9:00-9:45. Christensen Sören.
- 9:45-10:30. Pang Guodong.
- 10:30-11:00. Coffee break.
- 11:00-11:25. Graf Julius.
- 11:25-11:50. Sananes Simon.
- 14:00-14:45. Conforti Giovanni.
- 14:45-15:30. Siska David.
- 15:30-16:00. Coffee break.

Wednesday 8th.

- 9:00-9:45. Pham Huyen.
- 9:45-10:30. Oudjane Nadia.
- 10:30-11:00. Coffee break.
- 11:00-11:45. Hubert Emma.
- 11:45-12:10. Ruelland Justin.
- 14:00-14:45. Garcia Trillos Camilo Andres.
- 14:45-15:30. Geiss Hannah.
- 15:30-16:00. Coffee break.
- 16:00-16:25. Broux-Quémerais Guillaume.

Thursday 9th.

- 9:00-9:45. De Crescenzo Anna.
- 9:45-10:30. Li Xinyu.
- 10:30-11:00. Coffee break.
- 11:00-11:45. Ocello Antonio.
- 11:45-12:10. Bensaid Zakaria.
- 14:00-14:25. Sartor Enrico.
- 14:25-14:50. Nefzaoui Blanchard Cyril.
- 14:50-15:15. Yan Haoze.
- 15:15-15:45. Coffee break.
- 15:45-16:15. Yang Yanzhao.

Friday 10th.

- 9:00-9:45. Geiss Stefan.
- 9:45-10:30. Labart Céline.
- 10:30-11:00. Coffee break.
- 11:00-11:45. Kharroubi Idris.
- 11:45-12:10. Desbouis Kaplan.

Titles and abstracts

Zakaria Bensaid (Le Mans Université). Deep learning algorithms for FBSDEs with jumps.

In this paper, we introduce various machine learning solvers for forward-backward systems of stochastic differential equations (FBSDEs) driven by a Brownian motion and a Poisson random measure. In particular, we show the efficiency of the deep-learning algorithms to solve a coupled multi-dimensional FBSDE system driven by a time-inhomogeneous jump process with stochastic intensity, which describes the Nash equilibria for a specific mean-field game (MFG) problem for which we also provide the complete theoretical resolution. If time permits, theoretical convergence results will also be presented.

Guillaume Broux-Quémerais (Le Mans Université). Deep numerical schemes for systems of ergodic BSDEs with applications to regime-switching forward utilities.

In this paper, we develop two numerical methods for approximating forward utilities in a regime-switching stochastic factor model. Our approach exploits the characterization of these preferences through systems of ergodic Backward Stochastic Differential Equations (eBSDEs). We first establish a link between the solution of the ergodic BSDE system and that of an associated multidimensional BSDE with random terminal time, given by the return time of the positive recurrent stochastic factor. Building on this representation, we introduce a locally additive scheme obtained by minimizing aggregated local error terms. We then develop a DGM-inspired algorithm for approximating the associated system of ergodic partial differential equations. Numerical experiments demonstrate the performance of the proposed methods, with a particular focus on the impact of regime-switching on forward preferences in a two-state market model.

Christensen Sören (Kiel University). On Nonparametric Approaches to Data-Driven Stochastic Control.

While classical stochastic control theory typically assumes that the dynamics of the underlying diffusion process are fully specified, many contemporary applications require these dynamics to be inferred from data. We examine nonparametric learning frameworks for optimal stopping, singular control, and impulse control, in which data-driven strategies are constructed by estimating key characteristics of the underlying process. Depending on the control problem, the analysis ranges from pure exploration questions to sequential learning problems involving exploration–exploitation trade-offs. Particular emphasis is placed on how nonparametric concepts such as minimax rates and margin conditions determine the statistical complexity of these problems and yield refined convergence guarantees. The resulting perspective highlights connections between

stochastic control, statistical estimation, and learning-based approaches in continuous time.

Conforti Giovanni (Università degli Studi di Padova). Convergence bounds for diffusion flow matching and iterative Markovian fitting procedure.

Diffusion flow matching (DFM), or simply flow matching, is one of the most popular generative algorithms. In this talk, we discuss theoretical guarantees of convergence for DFM. Namely, we address the problem of quantifying the discrepancy between the algorithm' output and the underlying data distribution, taking into account both time-discretization and score-approximation errors. Iterating DFM yields the Iterative Markovian Fitting procedure (IMF) that may be regarded as a data-driven alternative to Sinkhorn algorithm for solving the entropic optimal transport problem. If time allows, we will show how to obtain convergence bounds in this more delicate setting.

De Crescenzo Anna (ETH Zürich). Mean-field control of heterogeneous systems.

We study the optimal control of mean-field systems with heterogeneous and asymmetric interactions. This leads to considering a family of controlled Brownian diffusion processes with dynamics depending on the whole collection of marginal probability laws. We prove the well-posedness of such systems and define the control problem together with its related value function. Leveraging tools tailored for this framework, such as derivatives along flows of measures and associated Itô calculus, we establish that the value function for this control problem satisfies a Bellman dynamic programming equation in a L^2-set of Wasserstein space-valued functions. To illustrate the applicability of our approach, we present a linear-quadratic graphon model with analytical solutions, and apply it to a systemic risk example involving heterogeneous banks.

Desbouis Kaplan (Institut de Mathématiques de Bordeaux). Probabilistic approach to mean-field ergodic control problems.

In this talk, we will be insterested in MF ergodic control problems, in strong (and weak if time allows) formulation, and their links with Ergodic FBSDE. After giving an existence and uniqueness result, we will talk about the long time behaviour of the finite horizon control problem towards the ergodic one.

Garcia Trillos Camilo Andres (University College London). Deep Learning Beyond Markov: FBSVIEs and Time-Inconsistent Control.

Many problems in modern stochastic control involve memory and time inconsistency, where classical Markovian methods break down. Forward-backward stochastic Volterra integral equations (FBSVIEs) offer a powerful framework for capturing these effects in many contexts, with important applications in finance. In this talk, we discuss FBSVIEs and their connections to BSDEs, and explain how they arise naturally in time-inconsistent control. We then present neural network-based methods for solving these equations, leveraging ideas from simulation, optimisation, and reinforcement learning. This perspective yields flexible and scalable algorithms for tackling high-dimensional, memory-dependent systems beyond the reach of traditional techniques. Based on joint work with K. Andersson and A. Gnoatto.

Geiss Stefan (University of Jyväskylä). On the approximation of a class of stochastic integrals with anticipating integrands.

We study the quantitative approximation of a class of stochastic integrals, where we use discrete time approximations under an initial enlargement of the filtration. When the underlying driving process is the geometric Brownian motion, then this approximation problem models discrete time-hedging under insider information. Three different approximation schemes are compared by computing L_2-scaling limits when the number of discretization steps goes to infinity. As the integrands of the stochastic integrals are anticipating, Skorohod integration is used. The talk is based on joined work with H. Geiss and O. Hinkkanen (Approximation of certain stochastic integrals with anticipating integrands. ArXiv:2606.08776, 2026).

Geiss Hannah (University of Jyväskylä). Convergence rate of random walk approximations of mean field BSDEs.

We study the rate of convergence w.r.t. a Wasserstein type distance for random walk approximations of (quadratic) mean field BSDEs. Our method uses a freezing technique. We extend results by Briand et al. about the rate of convergence of a Donsker-type theorem for BSDEs from the classical setting to the mean field setting. In this connection the mean field setting leads to new phenomena: A Hölder continuous terminal condition causes a singularity in time in the generator when the mean field terms are frozen. To handle this singularity we introduce a concept of modified Hölder continuity which enables to achieve, up to a logarithmic term, the same polynomial approximation rates as in the classical non-mean field setting. In the quadratic case, if the terminal condition Lipschitz is continuous , we achieve the same approximation rates using a result of Richou. This is joint work with Boualem Djehiche, Stefan Geiss, Céline Labart, and Jani Nykänen.

Graf Julius (University of California, Berkeley). Learning Market Making with Closing Auctions.

We investigate the market-making problem on a trading session in which a continuous phase on a limit order book is followed by a closing auction. Whereas standard optimal market-making models typically rely on terminal inventory penalties to manage end-of-day risk, ignoring the significant liquidity events available in closing auctions, we propose a Deep Q-Learning framework that explicitly incorporates this mechanism. We introduce a market-making framework designed to explicitly anticipate the closing auction, continuously refining the projected clearing price as the trading session evolves. We develop a generative stochastic market model to simulate the trading session and to emulate the market. Our theoretical model and Deep Q-Learning method is applied on the generator in two settings: (1) when the mid price follows a rough Heston model with generative data from this stochastic model; and (2) when the mid price corresponds to historical data of assets from the S&P 500 index and the performance of our algorithm is compared with classical benchmarks from optimal market making.

Hubert Emma (Université Paris Dauphine & PSL). Revisiting contract theory with volatility control.

In this talk, we revisit the resolution of continuous-time principal-agent problems with drift and volatility control, originally addressed by Cvitanić, Possamaï, and Touzi (2018) through dynamic programming and second-order backward stochastic differential equations(2BSDEs), and develop new results in this framework. We begin by introducing an alternative problem in which the principal is allowed to directly control the quadratic variation of the output process. On the one hand, the resolution of this contractible-volatility problem follows the classical methodology of Sannikov (2008), thus relying on standard (first-order) BSDEs only. On the other hand, we introduce a new form of contracts allowing the principal to achieve her contractible-volatility value, thereby ensuring both the optimality of this contract form and the equivalence between the original and the alternative problems. At the same time, this alternative approach reveals that the optimality of the contract form introduced by Cvitanić, Possamaï, and Touzi (2018) implicitly relies on an additional duality assumption, which was not identified before. This observation motivates the construction of new families of contracts that remain optimal even when the duality assumption fails. Altogether, this line of work both simplifies and strengthens the existing theory of continuous-time principal-agent problems with volatility control and opens new directions for further extensions and applications in economics and finance. Talk based on joint works with Alessandro Chiusolo, Dylan Possamaï, and Nizar Touzi.

Kharroubi Idris (Sorbonne University, LPSM). Optimal stopping of branching diffusion processes.

We explore an optimal stopping problem for branching diffusion processes. It consists in looking for optimal stopping lines, a type of stopping time that maintains the branching structure of the processes under analysis. By using a dynamic programming approach, we characterize the value function for a multiplicative cost, which may depend on the particleâs label. We reduce the problemâs dimensionality by setting a branching property and defining the problem in a finite-dimensional context. Within this framework, we focus on the value function, establishing uniform continuity and boundedness properties, together with an innovative dynamic programming principle. This outcome leads to an analytical characterization with the help of a nonlinear elliptic PDE. We conclude by showing that the value function serves as the unique viscosity solution for this PDE, generalizing the comparison principle to this setting.

Labart Céline (Université Savoie Mont-Blanc). Numerical study for optimal switching problems.

In this talk we deal with a numerical method for solving optimal switching problems (OSP) in presence of switching costs. Using the randomization method, we link the value function of the OSP to a non standard BSDE. We propose a numerical method combining a backward discretization and a least square regression method to approximate the solution of the non standard BSDE. We provide a rate of convergence for our algorithm. This is a joint work with Marie-Amelie Morlais.

Li Xinyu (University of Oxford). An α-potential game framework for N-player dynamic games.

Game theory has a long history, but computing Nash equilibria in dynamic non-cooperative games remains highly challenging. In this talk, we introduce a new framework for dynamic N-player games, called α-potential games, which avoids the homogeneity assumptions and infinite-population limits commonly imposed in mean field games. In an α-potential game, the change in a player's objective under a unilateral deviation coincides with the change in an α-potential function up to an error of order α. This reduces the computation of approximate Nash equilibria to a stochastic control problem, substantially simplifying both analysis and computation. Moreover, the parameter α captures key structural features of the game, including the population size, interaction strength, and degree of heterogeneity across players. We illustrate the framework through two examples, highlighting recent theoretical and algorithmic developments. For crowd-motion network games, we show that α=0 for all symmetric interaction networks. For asymmetric networks, we quantify the precise polynomial and logarithmic decay rates of α in terms of the number of players, network degree, and the decay rate of interaction asymmetry. We also apply the α-potential game framework to an N-player portfolio selection game under a mean-variance criterion, proving that the game is a potential game and explicitly constructing its Nash equilibrium. Our analysis allows

for heterogeneous preference parameters, extending beyond the mean-field interaction structures commonly studied in the literature. Finally, for general stochastic differential games, the associated optimization problem is embedded into a conditional McKean-Vlasov control problem to analyze the α-NE. A verification theorem is established to construct α-NE based on solutions to an infinite-dimensional Hamilton-Jacobi-Bellman equation, which is reduced to a system of ordinary differential equations for linear-quadratic games.

Nefzaoui Blanchard Cyril (LaMME, Université Evry-Paris Saclay). Piecewise constant policy approximation for the Quantile Hedging problem.

We study the numerical approximation of the quantile hedging problem in a Markovian diffusion framework modeling a complete and linear market. In this setting, the quantile hedging problem, a stochastic target problem , writes as a stochastic control problem for an augmented state variable (X, P), where P, controlled through its volatility coefficient in a degenerate and unbounded fashion, encodes the targeted success probability and, as such, is constrained to lie in $[0, 1]$.

Our analysis targets approximation procedures that are both implementable and compatible with the state constraint $P \in [0, 1]$: (i) truncation of the control set, and (ii) piecewise-constant-in-time controls (PCPT). As PCPT controls lead to conditionally Gaussian distributions, they do not satisfy the state constraint. One needs, in addition to a raw PCPT control, to consider the exit time from the interior of $[0, 1]$, and nullify the control after this time. Then, we rely on one-step dynamic programming principles and derive a suitable generator inequality for a regularised PCPT value function.

Under standard regularity assumptions on the coefficients and a Lipschitz condition on the payoff, we obtain explicit error bounds on the interior interval and optimise the regularisation parameters. This yields a quantitative rate for the combined effect of control truncation and PCPT time discretisation. We propose an implementation of the Deep BSDE algorithm and illustrate the convergence rate. This is joint work with Cyril Bénézet and Sergio Pulido.

Ocello Antonio (ENSAE Paris, CREST, IP Paris). Convergence Guarantees for Score-Based Generative Models: From Continuous Diffusions to Discrete Data.

Score-based generative models, also known as diffusion models, can be naturally formulated in continuous time as the time-reversal of a stochastic differential equation. In this talk, we present this formalism and highlight its connection with stochastic optimal control, where generation is interpreted as steering a reference diffusion toward a target distribution. Within this framework, we discuss the convergence bounds established by Conforti, Durmus, and Gentiloni-Silveri (2025), providing non-asymptotic guarantees under minimal regularity assumptions. We then extend this perspective to discrete

data generation via Discrete Markov Probabilistic Models (DMPMs). Here, the forward process is a continuous-time Markov chain on discrete states, and the reverse-time jump intensity is governed by a discrete analogue of the score function, characterized as a conditional expectation of the forward process. We present convergence guarantees in this discrete setting. This unified view connects diffusion models, optimal control, and discrete generative modeling within a rigorous convergence framework. This talk is based on joint work with Le-Tuyet-Nhi Pham, Dario Shariatian, Giovanni Conforti, and Alain Oliviero Durmus (ICML 2025).

Oudjane Nadia (EDF R&D). Optimizing and learning over the space of probability measures to manage flexibilities in power systems.

With the massive integration of renewable energies (photovoltaic (PV) and wind power) into the power grid, new uncertainties are impacting the power balance. At the same time, advances in smart technologies and batteries offer new flexibilities with the possibility of controlling the consumption of a large number of electrical appliances (electric vehicle recharging, heat pumps, etc.). In this framework, a major technical challenge is to manage this large number of flexible agents to help in balancing the system. By using mean-field approximations, we shift from analyzing individual agents to studying their distribution, making it easier to control large populations. We then consider a mean-field control problem formulated as an optimization problem over the space of probability measures. To address the challenges posed by unknown environments, we introduce adaptive algorithms based on principles from both online convex optimization and convex reinforcement learning theory.

Pang Guodong (Rice University). New Relative Value Iteration and Q-Learning Algorithms for Ergodic Risk Sensitive Control of Markov Chains.

In this talk, we will present new Jacobi-like relative value iteration (RVI) algorithms for the ergodic risk-sensitive control (ERSC) problem of discrete-time Markov chains, and the associated Q-learning algorithms. In the case of finite state space, we prove the iterates of the new RVI algorithms converge geometrically. We employ the entropy variational formula in order to tackle the multiplicative nature of the risk-sensitive Bellman operator, albeit with an additional optimization problem over a corresponding set of probability vectors. We then develop the entropy-based risk-sensitive Q-learning algorithms corresponding to the existing and new Jacobi-like RVI algorithms. These Q-learning algorithms have two components: the usual Q-function iterates and the new probability iterates arising from the entropy-variational formula. We prove the convergence of the coupled iterates by investigating the multi-scale stochastic approximations for these iterates.

Pham Huyen (Ecole Polytechnique). Learning generative dynamics with soft law constraints: a McKean-Vlasov FBSDE approach.

We study how to learn stochastic generative dynamics from empirical distributional snapshots observed at several time points. Instead of imposing hard marginal constraints, we introduce soft law penalties, which are more robust when the observed laws are noisy or sample-based. The resulting McKean-Vlasov control problem is characterised by a forward-backward SDE whose backward component has deterministic-time jumps driven by law-gradient forces. We discuss a sample-based deep BSDE solver and illustrate the approach on synthetic transport, latent-space face generation, and human motion examples, with perspectives for quantitative finance applications.

Ruelland Justin (Sorbonne University, LPSM). Option-Surface Fitting as a Dynamical Optimal Transport Problem.

We study the well-posedness of fitting an option-price surface to a finite set of European option prices using semimartingale optimal transport (SMOT). We establish market conditions ensuring existence, uniqueness, and stability of the solution to the corresponding optimal transport problem. The fitting is exact, and the resulting surface is arbitrage-free. The problem is addressed through its associated dual formulation, which involves a Hamilton-Jacobi-Bellman equation and leads us to investigate properties of the corresponding solution map, including differentiability and convexity. In high dimensions, solving the problem numerically becomes computationally demanding, as it requires optimising an objective function whose evaluation involves solving an HJB equation. We discuss potential approaches to overcome this difficulty.

Sananes Simon (Sorbonne University, LPSM). Stochastic Policy Gradient Methods in the Uncertain Volatility Model.

The multidimensional Uncertain Volatility Model (UVM) leads to robust option pricing problems under joint volatility and correlation uncertainty. Their numerical resolution is challenging because the associated stochastic control problem is high-dimensional. We propose a backward actor-critic stochastic policy-gradient scheme tailored to this setting, combining a discrete dynamic programming principle with policy-gradient optimization and neural-network approximations of both the value function and the control policy. A key ingredient is a policy parameterization that enforces positive definiteness of the correlation matrix by construction. Beyond the robust price, the spatial gradient of the trained critic yields an approximation of the associated superhedging strategy, whose quality we assess a posteriori through a dual upper bound. Numerical experiments on a range of multidimensional derivatives show that the method yields accurate prices, remains computationally efficient, and compares favorably with existing Monte Carlo and machine-learning benchmarks.

Sartor Enrico (Laboratoire des Signaux et Systèmes, Université Paris-Saclay, CentraleSupélec). A Particle Perspective on Partially Observed Stochastic Control.

Partially observed stochastic control couples decision-making with online inference: controls affect both the hidden state and the information available through observations. Since partial observation destroys the Markovian structure of the state process, one typically reformulates the problem on the belief space, using the conditional law of the hidden state as the new state variable. In this talk, I will discuss an exploratory particle-based perspective on this formulation. Motivated by nonlinear filtering, where beliefs are approximated by weighted particles, the goal is to investigate whether such finite-dimensional representations can provide a useful framework for stochastic optimal control, and to outline some of the analytical questions that arise from this viewpoint.

Siska David (School of Mathematics, University of Edinburgh). Convergence of actor-critic algorithms for discrete and continuous time RL.

First we will review recent results on convergence of actor-critic algorithms for Markov Decision Problems formulated on general state and action spaces. In general, under Q-function realisability and assuming exact integral evaluations we can show sub-linear convergence for double-loop actor critic. For single loop actor critic additional structural assumptions are needed. Depending on how current research progresses we will discuss how these results extend to continuous time RL. This is joint work with Galen Cao, Lukasz Szpruch, Yufei Zhang and Denis Zorba.

Yan Haoze (UC Berkeley). Controlled Hawkes JumpâDiffusions and CT-DDPG via Markov Approximation.

We study stochastic control of multivariate Hawkes-driven stochastic differential equations with machine learning algorithms. Due to the path dependence of the memory of the Hawkes intensity, this problem does not fall within classical stochastic control theory outside particular Markovian kernels. We first develop a finite-dimensional Markovianization procedure and algorithm to approximate multivariate Hawkes processes with mixtures of exponential kernels. We prove the convergence of the Markovianized approximation of the Hawkes process, its intensity, and the value of the problem to the original non-Markovian processes and the value of the primal problem. We then formulate continuous-time deterministic policy gradient learning on the Markovianized approximation of the problem, called Hawkes-CT DDPG. We propose a model-free algorithm to solve the non-Markovian Hawkes-driven optimization by observing only the event times of the process, the realization of the solution to the SDE, and a chosen set of decay filters, while the Hawkes kernel coefficients remain unknown. We compare our

reinforcement learning Hawkes-CT DDPG method with classical reinforcement learning techniques under three different types of kernels: simple exponential, Erlang, and power-law kernels.

Yang Yanzhao (University of Oxford). Adaptive Partitioning and Learning for Stochastic Control of Diffusion Processes.

We study reinforcement learning for controlled diffusion processes with unbounded continuous state spaces, bounded continuous actions, and polynomially growing rewards– settings that arise naturally in finance, economics, and operations research. To overcome the challenges of continuous and high-dimensional domains, we introduce a model-based algorithm that adaptively partitions the joint state-action space. The algorithm maintains estimators of drift, volatility, and rewards within each partition, refining the discretization whenever estimation bias exceeds statistical confidence. This adaptive scheme balances exploration and approximation, enabling efficient learning in unbounded domains. Our analysis establishes regret bounds that depend on the problem horizon, state dimension, reward growth order, and a newly defined notion of zooming dimension tailored to unbounded diffusion processes. The bounds recover existing results for bounded settings as a special case, while extending theoretical guarantees to a broader class of diffusion-type problems. Finally, we validate the effectiveness of our approach through numerical experiments, including applications to high-dimensional problems such as multi-asset mean-variance portfolio selection.