USC Viterbi
School of Engineering

CSCI 699: Safe Learning-Enabled Autonomous Systems

Units: 4.0
Spring 2024 MonWed 5:00-6:50pm

Location: TBA

Instructor: Prof. Dr. Lars Lindemann

Office: SAL 328
Office Hours: TBA
Contact Info: llindema@usc.edu

Catalogue Description

Certifying input-output properties of neural networks; neural networks within autonomy; verification of learning-enabled autonomous systems; safe learning-enabled planning and control.

Course Description

Autonomous systems are engineered systems that operate without or only with little human intervention. Applications can be found in self-driving cars, intelligent transportation, and robotics. Modern autonomous systems are learning-enabled, i.e., they use neural networks to reason and learn from data. In fact, learning-enabled systems promise to enable many future technologies. However, initial excitement was quickly met with disappointment and safety concerns as the fragility of learning-enabled systems became apparent, e.g., slowing down autonomous driving research. This fragility raises fundamental questions for the research community regarding the safety of learning-enabled systems that we will study in this course.

In this course, we will study mathematical techniques to analyze and design learning-enabled autonomous systems. The course will consist of three main parts. 1. We will certify input-output properties of neural networks by computing reachable sets and Lipschitz constants via complete methods (satisfiable modulo theory solver, mixed integer linear programming), incomplete methods (bound propagation, abstract interpretation, semidefinite programming), and statistical methods (conformal prediction). In the process, we study the efficiency-precision trade-off. 2. We will study learning-enabled components within an autonomous systems, e.g., when neural networks are used for perception, prediction, and control. We then verify safety of the learning-enabled autonomous system using concepts from systems theory (barrier functions, hybrid systems reachability) and formal methods (temporal logics) along with tools learned in the first part. 3. We will design safe planning and control algorithms for learning-enabled systems using techniques from control theory (control barrier functions, model predictive control, reinforcement learning) along with tools in the first and second parts. If time permits, we will study advanced topics to enhance the safety of autonomous systems, e.g., adversarial, distributionally robust, and risk-aware learning.

The course is designed to be interactive and engaging with a mix of lectures — given by the instructor — and student presentations of recent research papers. The student presentations have a “reading group” format with the goal to stimulate a critical discussion of the paper. In this format, one or two students (who read the paper in detail) will act as an expert moderator that walk the class through the paper and ask critical questions to initiate a dialogue and deepen everyone’s understanding of the paper. Therefore, everyone is require to read each paper before class. The course will also contain a project in which students will implement the learned tools on a learning-enabled autonomous systems of their choice. The course is designed for PhD students in robotics, control theory, machine learning, optimization, cyber-physical and autonomous systems, and formal methods.

Learning Objectives

By the end of this course, students will be able to understand and use complete methods (satisfiable modulo theory solver, mixed integer linear programming), incomplete methods (bound propagation, abstract interpretation, semidefinite programming), and statistical methods (conformal prediction) to certify input-output properties and robustness of neural networks. In doing so, students will be able to assess limitations of each method and understand the efficiency-precision trade-off. Students will be able to verify safety of the learning-enabled autonomous system using concepts from systems theory (barrier functions, hybrid systems reachability) and formal methods (temporal logics). Students will be able to design safe planning and control algorithms for learning-enabled systems using techniques from control theory (control barrier functions, model predictive control, reinforcement learning). Students will apply the learned techniques to design a safe learning-enabled autonomous system in a simulator of their choice. Throughout the course, students will learn to critically assess limitations of the learned techniques. Besides these technical objectives, the students will gain experience in reading, understanding, analyzing, and communicating research by the “reading group” format of the student presentations. Additionally, students will become familiar with the main conferences and journals in the area of safe learning-enabled autonomy.

Recommended Preparation

No strict prerequisites are required. After all, the most important prerequisite is the student's interest in the topic, motivation, and commitment to learning. Nonetheless, background in optimization is highly recommended, e.g., convex and combinatorial optimization (at the level of CSCI 675) or optimization for information and data sciences (at the level of EE 588). Additionally, some background in systems theory and logic may be helpful, e.g., as (partially) taught in linear systems theory (at the level EE 585), nonlinear and adaptive control (at the level of EE 587), and introduction to artificial intelligence (at the level of CSCI 360).

Course Notes

The final grade will be determined based on attendance, homeworks, course presentations, and a course project.

Technological Proficiency and Hardware/Software Required

Course work and projects require standard computing software (Matlab, Python, C++).

Required Readings and Supplementary Materials

Required readings and supplementary materials will be announced during the course.

Optional Readings and Supplementary Materials

Optional readings and supplementary materials will be announced during the course.

Description of Assignments and How They Will Be Assessed

There are three types of assignments: student presentations, homeworks, and course project.

- Students will be assigned to research papers that are to be presented in class. Students will have to walk the class through the paper and ask critical questions to initiate a dialogue and deepen everyone's understanding of the paper. Grading will be based on the ability of the students to convey the main points of the paper and to create a stimulating dialogue, e.g., by asking critical questions.
- Homeworks will be assigned and need to be solved individually by each student. Homeworks consist of pen and paper problems as well as coding assignments.
- The course project is a research project, i.e., it is the students responsibility to i) define a problem with a scope aligned with the course topic, and ii) to propose a viable solution and illustrate it. It is required that students implement their solution on an autonomous systems simulator of their choice (the course instructor will provide suggestions on suitable simulators). A final project report has to be submitted and a 10-15 minute research presentation has to be given in front of the class. The course project can be performed in teams of at most two students. An optimal outcome of the course project would be a conference submission.

Participation

Participation is mandatory and is part of the grading (see below).

Grading Breakdown

Including the above detailed assignments, how will students be graded overall? Participation should be no more than 15%, unless justified for a higher amount. All must total 100%.

Assignment	Points	% of Grade
Participation	10	10
Homework 1	10	10
Homework 2	10	10
Homework 3	10	10

Presentations	30	30
Course Project	30	30
TOTAL	100	100

Attendance

Attendance is mandatory.

Academic Integrity

Unless otherwise noted, this course will follow the expectations for academic integrity as stated in the USC Student Handbook. The general USC guidelines on Academic Integrity and Course Content Distribution are provided in the subsequent “Statement on Academic Conduct and Support Systems” section.

Use of Generative AI in this Course

Generative AI is not permitted: Since creating, analytical, and critical thinking skills are part of the learning outcomes of this course, all assignments should be prepared by the student working individually or in groups as described on each assignment. Students may not have another person or entity complete any portion of the assignment. Developing strong competencies in these areas will prepare you for a competitive workplace. Therefore, using AI-generated tools is prohibited in this course, will be identified as plagiarism, and will be reported to the Office of Academic Integrity.

Course Evaluations

[Course evaluation occurs at the end of the semester university-wide. It is an important review of students ’ experience in the class. The process and intent of the end-of-semester evaluation should be provided. In addition, a mid-semester evaluation is recommended practice for early course correction.]

Course Schedule

Below is a tentative list of the course schedule that outlines the aforementioned main three parts of the course The schedule may change during the course based on preferences of students. The suggested readings/preparations is an exhaustive list of related work — we may not be able to cover all papers.

	Topics/Daily Activities	**Readings/Preparation**	**Deliverables**

Week 1	Course logistics, Fragility of Neural networks (NNs), Desirable input-output properties of NNs and their certification	[1], [2], [3], [4], [5], [6]	
Week 2	**Part 1: Certification of NNs** Complete certification methods using satisfiable modulo theory solver (SMT)	[7], [8], [9], [10]	
Week 3	Complete certification methods using mixed integer linear programming (MILP)	[11], [12], [13], [14]	
Week 4	Sound certification methods using semidefinite programming (SDP)	[15], [16], [17]	
Week 5	Sound certification methods using abstract interpretation and bound propagation	[18], [19], [20], [21], [22]	
Week 6	Statistical certification methods using conformal prediction (CP)	[23], [24], [25]	
Week 7	**Part 2: Verification of learning-enabled autonomous systems** Neural Network Dynamical Systems (NNDS), Barrier functions for NNDS and their synthesis	[26], [27], [28], [29]	
Week 8	Verifying Neural Network Controlled Dynamical Systems (NNCDS) using Hybrid System Reachability	[30], [31], [32]	
Week 9	Verifying Neural NNCDS using abstract models	[33], [34], [35], [36], [37]	
Week 10	Verifying NNCDS using CP and Risk Measures	[38], [39], [40], [41], [42]	
Week 11	**Part 3: Safe Planning and Control for Learning-Enabled Systems** Planning in Dynamic Environments using CP	[43], [44], [45]	
Week 12	PAC-Bayes Control	[46], [47], [48]	
Week 13	Vision-based Control and Out-of-Distribution Detection	[49], [50], [51], [52]	
Week 14	Learning Control Barrier Functions from Expert Demonstrations	[53], [54], [55]	
Week 15	Safe Reinforcement Learning	[56], [57]	
FINAL			Refer to the final exam schedule in the USC *Schedule of Classes* at classes.usc.edu.

Fragility of Neural Networks and overview:

[1] “Intriguing Properties of Neural Networks” by Szegedy et. al.

[2] “Explaining and Harnessing Adversarial Examples” by Goodfellow et. al.

[3] “Measuring Neural Net Robustness with Constraints” by Bastani et. al.

Certification of Neural Networks (Overview and toolbox)

[4] “When to Trust AI: Advances and Challenges for Certification of Neural Networks” by Kwiatkowska et. al.

[5] “Algorithms for Verifying Deep Neural Networks” by Liu et. al.

[6] “NNV: The Neural Network Verification Tool for Deep Neural Networks and Learning-Enabled Cyber-Physical Systems” by Tran et. al.

Part 1: Certification of Neural Networks

SMT:

[7] “Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks” by Ehlers

[8] “Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks” by Katz et. al.

[9] “Safety Verification of Deep Neural Networks” by Huang et. al.

[10] “The Marabou Framework for Verification and Analysis of Deep Neural Networks” by Katz et. al.

MILP:

[11] “Output Range Analysis for Deep Neural Networks” by Dutta et. al.

[12] “An Approach to Reachability Analysis for Feed-Forward ReLU Neural Networks” by Lomuscio et. al.

[13] “Evaluating Robustness of Neural Networks with Mixed Integer Programming” by Tjeng et. al.

[14] “Deep Neural Networks and Mixed Integer Linear Optimization” by Fischetti et. al.

SDP:

[15] “An Introduction to Neural Network Analysis via Semidefinite Programming” by Fazlyab et. al.

[16] “Safety Verification and Robustness Analysis of Neural Networks via Quadratic Constraints and Semidefinite Programming” by Fazlyab et. al.

[17] “Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks” by Fazlyab et. al.
Bound Propagation:

Bound Propagation:

[18] “Efficient Neural Network Robustness Certification with General Activation Functions” by Zhang et. al.

[19] “Beta-CROWN: Efficient Bound Propagation with Per-Neuron Split Constraints for Neural Network Robustness Verification” by Wang et. al.

Abstract Interpretation:

[20] “AI^2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation” by Gehr et. al.

[21] “An Abstract Domain for Certifying Neural Networks” by Singh et. al.

[22] “Fast and Effective Robustness Certification” by Singh et. al.

Conformal Prediction:

[23] “A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification” by Angelopoulos et. al.

[24] “Conformal Prediction Regions for Time Series using Linear Complementarity Programming” by Cleaveland et. al.

[25] “Conformal Prediction under Covariate Shift” by Tibshirani et. al.

Part 2: Verification of Learning-Enabled Autonomous Systems

Barrier Functions:

[26] “FOSSIL: A Software Tool for The Formal Synthesis of Lyapunov Functions and Barrier Certificates using Neural Networks” by Abate et. al.

[27] “Formal Synthesis of Lyapunov Neural Networks” by Abate et. al.
[28] “Simulation-Guided Lyapunov Analysis for Hybrid Dynamical Systems” by Kapinski et. al.
[29] “Safety Guarantees for Neural Network Dynamic Systems via Stochastic Barrier Functions” by Mazouz et. al.

Hybrid System Reachability:
[30] “Verisig: Verifying Safety Properties of Hybrid Systems with Neural Network Controllers” by Ivanov et. al.
[31] “Verifying the Safety of Autonomous Systems with Neural Network Controllers” by Ivanov et. al.
[32] “Case Study: Verifying the Safety of an Autonomous Racing Car with a Neural Network Controller” by Ivanov et. al.

Abstract Models and Reachability:
[33] “Reach-SDP: Reachability Analysis of Closed-Loop Systems with Neural Network Controllers via Semidefinite Programming” by Hu et. al.
[34] “One-Shot Reachability Analysis of Neural Network Dynamical Systems” by Chen et. al.
[35] “Reachability Analysis for Neural Feedback Systems using Regressive Polynomial Rule Inference” by Dutta et. al.
[36] “ReachNN: Reachability Analysis of Neural-Network Controlled Systems” by Huang et. al.
[37] “Safety Verification of Cyber-Physical Systems with Reinforcement Learning Control”

Conformal Prediction:
[38] “Risk of Stochastic Systems for Temporal Logic Specifications” by Lindemann et. al.
[39] “Risk Verification of Stochastic Systems with Neural Network Controllers” by Cleaveland et. al.
[40] “Data-Driven Reachability Analysis of Stochastic Dynamical Systems with Conformal Inference” by Hashemi et. al.
[41] “Statistical Verification of Autonomous Systems using Surrogate Models and Conformal Inference” by Qin et. al.
[42] “Conformal Prediction for STL Runtime Verification” by Lindemann et. al.

Part 3: Safe Planning and Control for Learning-Enabled Systems

Conformal Prediction:
[43] “Safe Planning in Dynamic Environments using Conformal Prediction” by Lindemann et. al.
[44] “Adaptive Conformal Prediction for Motion Planning among Dynamic Agents” by Dixit et. al.
[45] “Conformal Predictive Safety Filter for RL Controllers in Dynamic Environments” by Strawn et. al.

PAC-Bayes Control:
[46] “PAC-Bayes Control: Synthesizing Controllers that Provably Generalize to Novel Environments” by Majumdar et. al.
[47] “PAC-Bayes Control: Learning Policies that Provably Generalize to Novel Environments” by Majumdar et. al.
[48] “Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability” by Farid et. al.

Vision-based Control and Out-of-Distribution Detection:
[49] “Probably approximately correct vision-based planning using motion primitives” by Veer et. al.
[50] “Failure Prediction with Statistical Guarantees for Vision-Based Robot Control” by Farid et. al.
[51] “Task-driven Out-of-Distribution Detection with Statistical Guarantees for Robot Learning” by Farid. et. al.
[52] “Generalized Out-of-Distribution Detection: A survey” by Yang et. al.

Learning Control Barrier Functions:
[53] “Learning Control Barrier Functions from Expert Demonstrations” by Robey et. al.
[54] “Learning Hybrid Control Barrier Functions from Data” by Lindemann et. al.

[55] "Learning Robust Output Control Barrier Functions from Safe Expert Demonstrations" by Lindemann et. al.

Safe Reinforcement Learning:
[56] "Sim-to-Lab-to-Real: Safe reinforcement learning with Shielding and Generalization Guarantees" by Hsu et. al.
[57] "Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning" by Brunke et. Al.

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Course Content Distribution and Synchronous Session Recordings Policies
USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.