POLITECNICO DI MILANO

PhD in MODELLI E METODI MATEMATICI PER L'INGEGNERIA / MATHEMATICAL MODELS AND METHODS IN ENGINEERING - 38th cycle

PNRR_352 Research Field: : PHYSICS-INFORMED DATA AUGMENTATION FOR MACHINE LEARNING APPLICATIONS

Monthly net income of PhDscholarship (max 36 months)
€ 1450.0 In case of a change of the welfare rates during the three-year period, the amount could be modified.

Context of the research activity	
Motivation and objectives of the research in this field	Numerical solvers are getting increasing importance as data generators in those contexts where data are scarce and machine learning or deep learning algorithms must be trained. On the other hand, enforcing physical constraints or equations governing the phenomenon at hand has become possible also when relying on neural networks, thanks to the so-called physics-informed neural networks. Scientific machine learning nowadays provides several strategies to blend physics-based modeling (e.g., relying on dynamical systems, or differential problems) with data analysis tools to enhance decision support, forecasting and maintenance. This project focuses on the construction, the analysis and the application of synthetic data generation algorithms. Synthetic data can then be used to supplement, augment and in some cases replace real data when training Machine/Deep Learning models. Additionally, it enables the testing of Machine Learning or other data dependent software systems without the risk of exposure that comes with data disclosure. Through this doctoral fellowship, we are planning to interact with Intellico engineers and data scientists, with the aim of developing synthetic data generation algorithms to enhance the performance of machine/deep learning algorithms that must be ultimately devised as end-to-end solutions for customers. This research contributes to the PNRR aim to increase knowledge and competencies on

	advanced simulation and big data as a key enabling technology. It also contributes to fostering collaboration between universities and Industry.
Methods and techniques that will be developed and used to carry out the research	During the past decade, synthetic data has been usually generated by treating each column in a table as a randomvariable, modeling a joint multivariate probability distribution, and then sampling from that distribution. Decision trees and Bayesian networks, as well as copulas, have been used to model discrete or continuous variables, respectively, however featuring severe computational issues and limiting the synthetic data¿s fidelity. Generative models using variational autoencoders, as well as generative adversarial networks, nowadays offer better performances and flexibility in representing data of increasing complexity. Starting from these techniques, the project will focus on the construction of a unique framework to perform synthetic data generation, in which hierarchical generative modeling and recursive sampling techniques within the framework of deep learning will be mainly explored, possibly embedding physical constraints and insights under the form of a system of equations describing the process or the system at hand. Multi-fidelity strategies relying on neural networks will also be considered to leverage the knowledge arising from data sources or models with different fidelity. Practical application will involve, e.g., predictive maintenance (e.g., in the energy sector, regarding both storage and production, as well as in health monitoring of civil infrastructures) and image processing for supporting clinical decisions. Synthetic data are expected to improve the performance of time series-type algorithms as well as image processing in the aforementioned contexts.
Educational objectives	The research will be carried out within a team of numerical analysis with experience on physics informed machine learning and neural network algorithms. An internship of at least 6 months at Intellico srl is planned, where the student will have the possibility to interact with an industrially focused team. Within this lively and stimulating academic and industrial environment, the

	doctoral student will become a skillful data scientist, with expertise in advanced tecniques.
Job opportunities	Data scientists are the most in-demand job today, among high-qualification jobs. In all industrial and business sectors, the demand for data scientists continues to outpace supply and dominates both the US and the European job market.
Composition of the research group	5 Full Professors 6 Associated Professors 4 Assistant Professors 15 PhD Students
Name of the research directors	Luca Formaggia, Andrea Manzoni

Contacts
Prof. Luca Formaggia: luca.formaggia@polimi.it, Prof. Andrea Manzoni: andrea1.manzoni@polimi.it

Additional support - Financial aid per PhD student per year (gross amount)			
	1st year	**2nd year**	**3rd year**
Housing - Foreign Students	1500.0 € per student	0.0 € per student	0.0 € per student
	max number of financial aid available: 3, given in order of merit ..		
Housing - Out-of-town residents (more than 80Km out of Milano)	--		

Scholarship Increase for a period abroad	
Amount monthly	725.0 €
By number of months	6

National Operational Program for Research and Innovation	
Company where the candidate will attend the stage (name and brief description)	Intellico srl
By number of months at the company	6
Institution or company where the candidate will spend the period abroad (name and brief description)	
By number of months abroad	6

Additional information: educational activity, teaching assistantship, computer availability, desk availability, any other information
Educational activities (purchase of study books and material, funding for participation to courses, summer schools, workshops and conferences): financial aid per PhD student per year 1st year: max 1.970,27 euros

2nd year: max 1.970,27 euros
3rd year: max 1.970,27 euros

The PhD students are encouraged to take part in activities related to teaching, within the limits allowed by the regulations.1 individual PC per student + several shared PC.
Access to one cluster with 32 processors and 384 GB RAM, and to several multi-processor servers and 1 individual desk per student are granted.