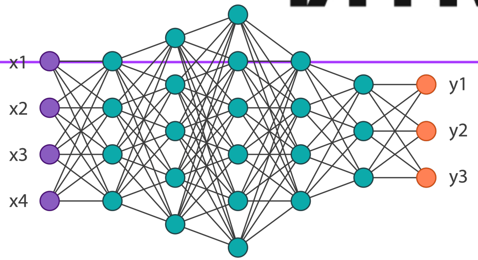
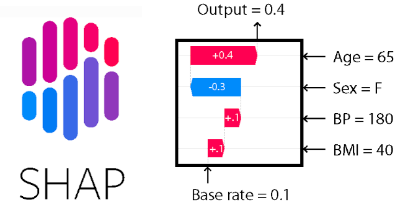
Modeling Neural Networks as Open Games for Robustness and Explainability

Maria Zaitseva, ITMO, 2025

Problem of Interpretability

- Neural networks remain to be **black boxes:** not straightforward enough to figure out why the model makes a particular choice
- Classical **game theory** is a useful explanatory tool (see SHAP), however its monolithic nature limits wide-scale application
- **Compositional Game Theory** [1] (CGT) is a promising foundation for a general method of analysis of neural networks analysis

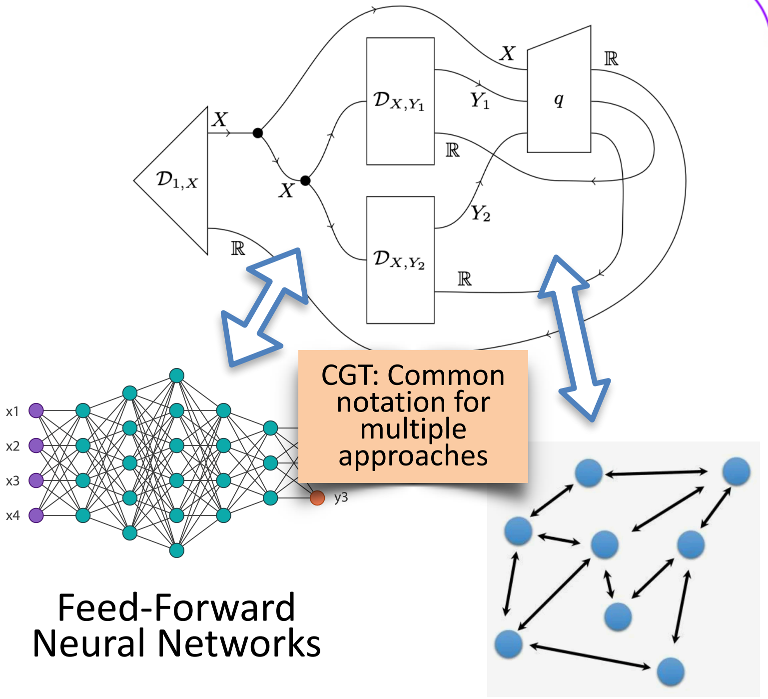




[1] N. Ghani, J. Hedges, V. Winschel, P. Zahn ***Compositional game theory*** // Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science. – 2018. – pp. 472–481.

Why Compositional Game Theory?

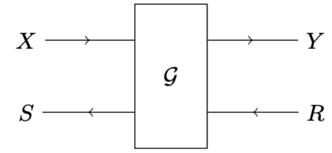
- **Compositionality:** split the system into sub-games (Open games) and analyze; combine systems arbitrarily
- Naturally captures backward pass semantics
- Possible to extend to multi-agent systems by linking to existing research
- Enables application of game theory algorithms for model verification and explainable AI



MARL

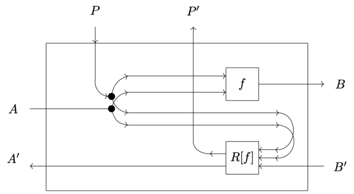
Modeling Neural Networks as Open Games

Combine



Open Games [1]

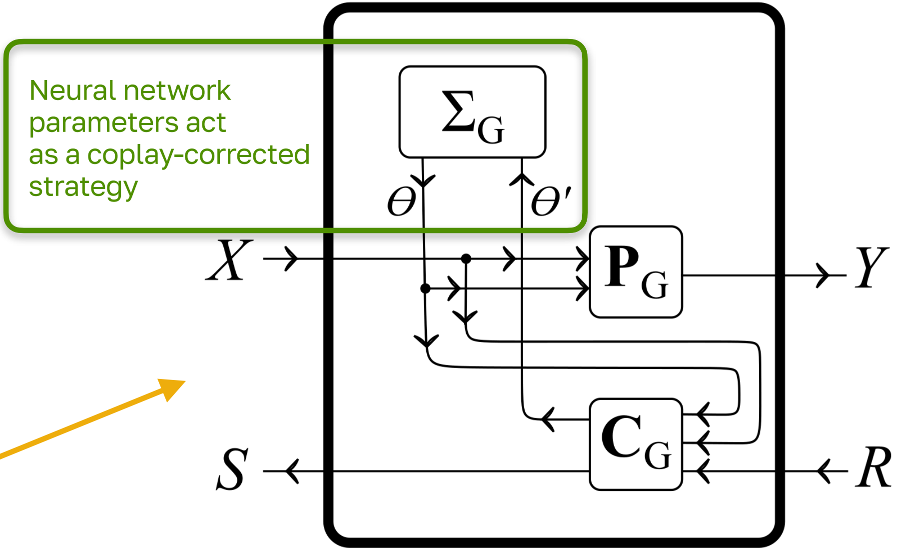
with



Para(Lens) [2]

and get this

Parametric Open Game

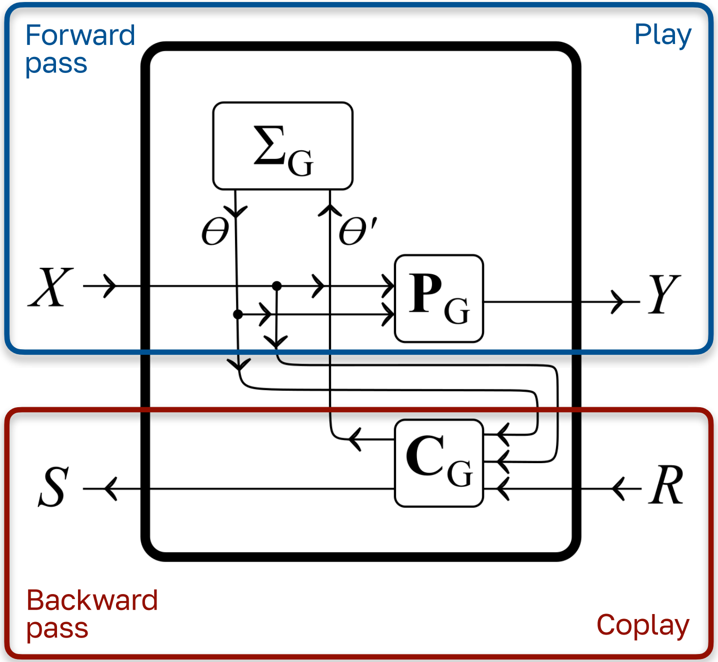


Lenses are a well-known structure in functional programming languages. It is a tool for abstracting *access:* reading/writing databases, JSON, XML, other composite data storage.

Parametric lenses [2] are lenses extended to support learning semantics for deep neural networks.

[2] Cruttwell, G. S. H., Gavranovic, B., Ghani, N., Wilson, P., & Zanasi, F. (2024). **Deep Learning with Parametric Lenses** // arXiv preprint arXiv:2404.00408. https://doi.org/10.48550/

Structure of Parametric Open Games



$$\mathcal{P} : (\underbrace{X}_{\text{observations}}, \underbrace{S}_{\text{state}}) \to (\underbrace{Y}_{\text{choice}}, \underbrace{R}_{\text{response (reward)}})$$

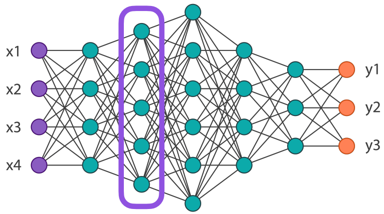
Open game maps observations + state to choices + rewards

$$\mathcal{P} = (\underbrace{\Sigma_G}_{\text{strategy profile}}, \underbrace{P_G}_{\text{play function}}, \underbrace{C_G}_{\text{coplay function}}, \underbrace{B_G}_{\text{best response function}})$$

Structure: tuple of functions

Example:

Feed-Forward Layer



Forward pass

X – inputs

Y – outputs

Backward pass

S – incoming error gradient

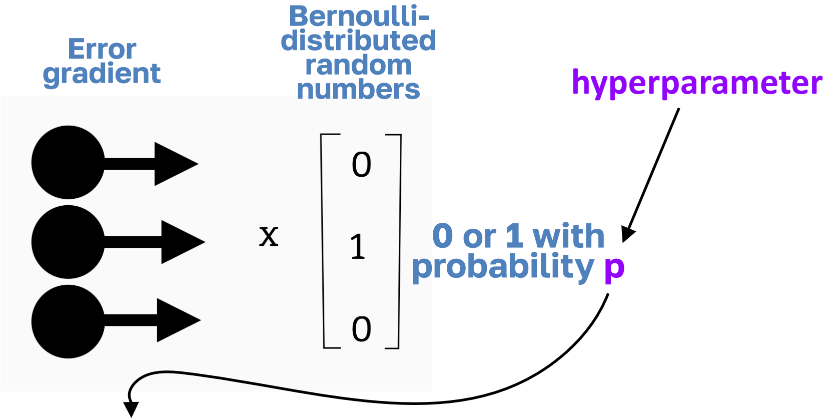
R – back-propagated gradient

Experiment

- In the experiment we evaluate a regularization technique called ***gradient dropout***, which is similar to the classical dropout, *but neurons remain active on forward pass*
- **Goal:** increase input noise robustness
- We observe performance metrics (MSE, SMAPE, accuracy, ROC AUC) and loss curves for varying values of hyperparameter *p*
- Metrics are measured against input noise of increasing value

$$\left(\frac{\partial L}{\partial w_{ji}^{(k)}}\right)_{\text{modified}} = m_j \cdot \delta_j^{(k)} a_i^{(k-1)}$$

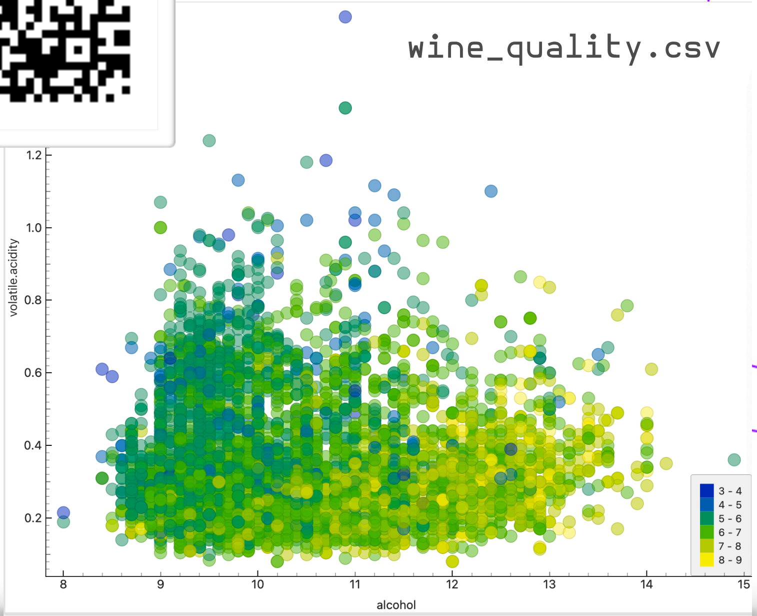
$$m_j \sim \text{Bernoulli}(p)$$



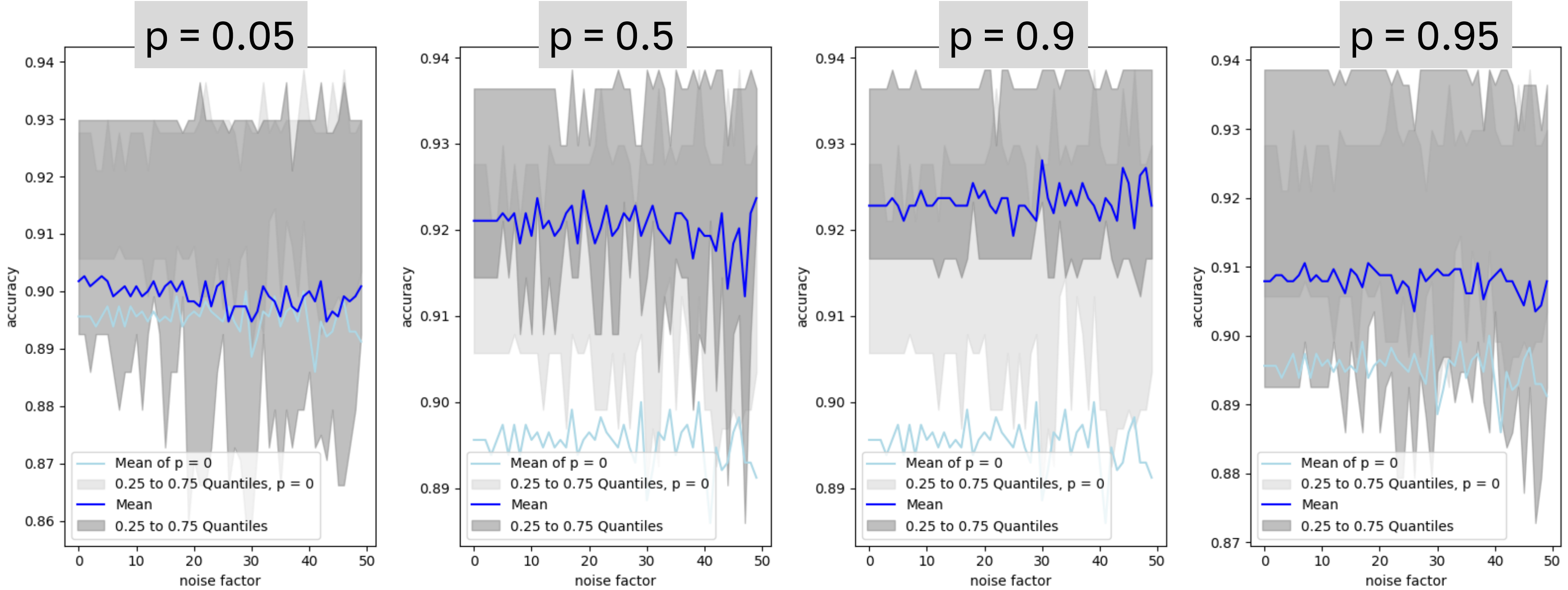
$$\text{p} \in \{0, 0.01, 0.05, 0.5, 0.9, 0.95, 0.99\}$$

Data Used for Training

- 10 datasets
- **Size:** 569–7608 samples
- **Tasks:** regression, classification (2, 4 classes)
- **Sources:**
  UCI ML Repo, Kaggle, curated collections
  PMLB (Penn’s machine learning benchmark),
  Tabular benchmark
- Both synthetic and real world data



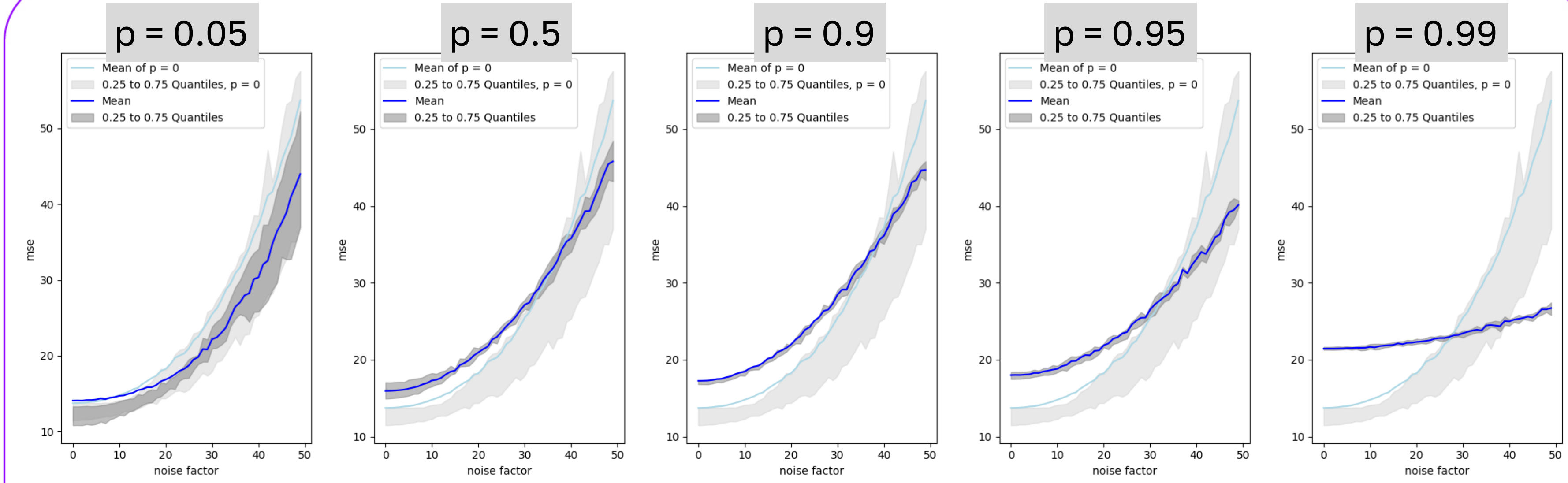
Classification Performance



Accuracy on Dataset #1: wisc_bc_data.csv
- Binary Classification

For some values of p, gradient dropout yields a performance boost

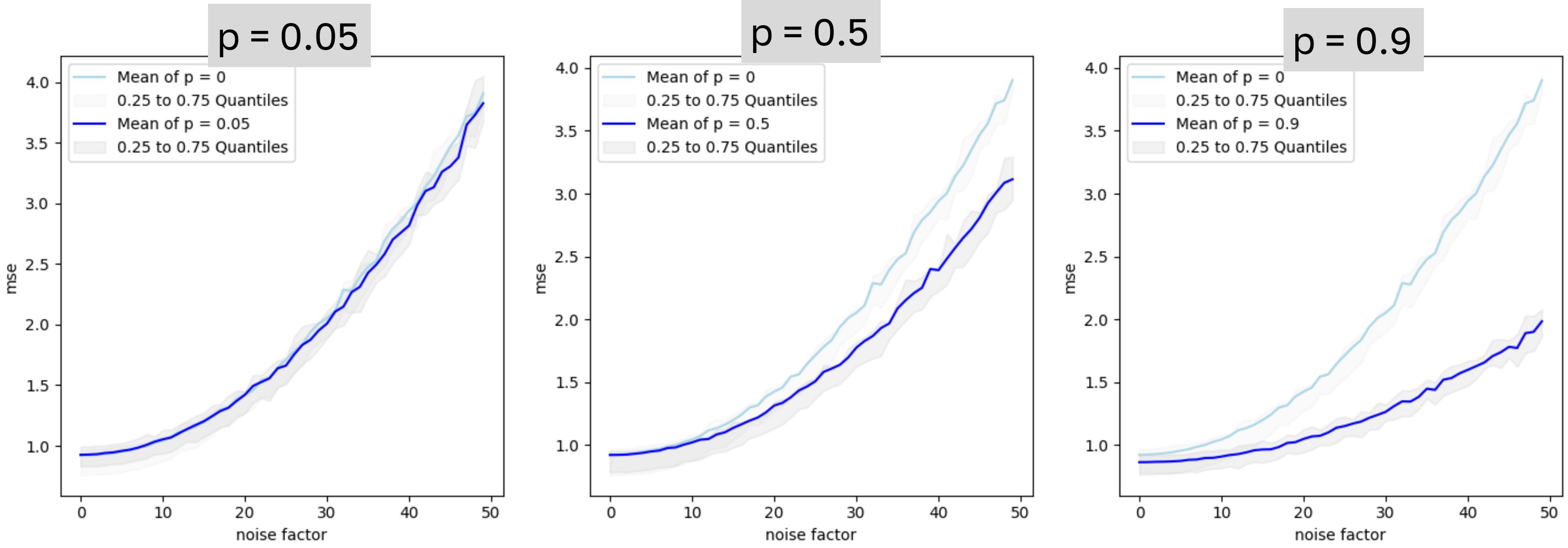
Regression Performance (1/2)



MSE on Dataset #3: StudentPerformanceFactors.csv
- Regression

Increasing p yields improved robustness at the cost of increased prediction error

Regression Performance (2/2)



MSE on Dataset #6: wine_quality.csv
- Regression

Improvement definitely manifested: lower error, higher robustness

Conclusion

- Now working on a paper for the **Games** journal
- In progress: game theory neural network architecture evaluation software based on **open-game-engine** [3]
- Future work: fuse FFNN theory with multi-agent reinforcement learning for wide-scale explainability

[3] https://github.com/CyberCat-Institute/open-game-engine

Thank you for your attention

Maria Zaitseva, ITMO, 2025