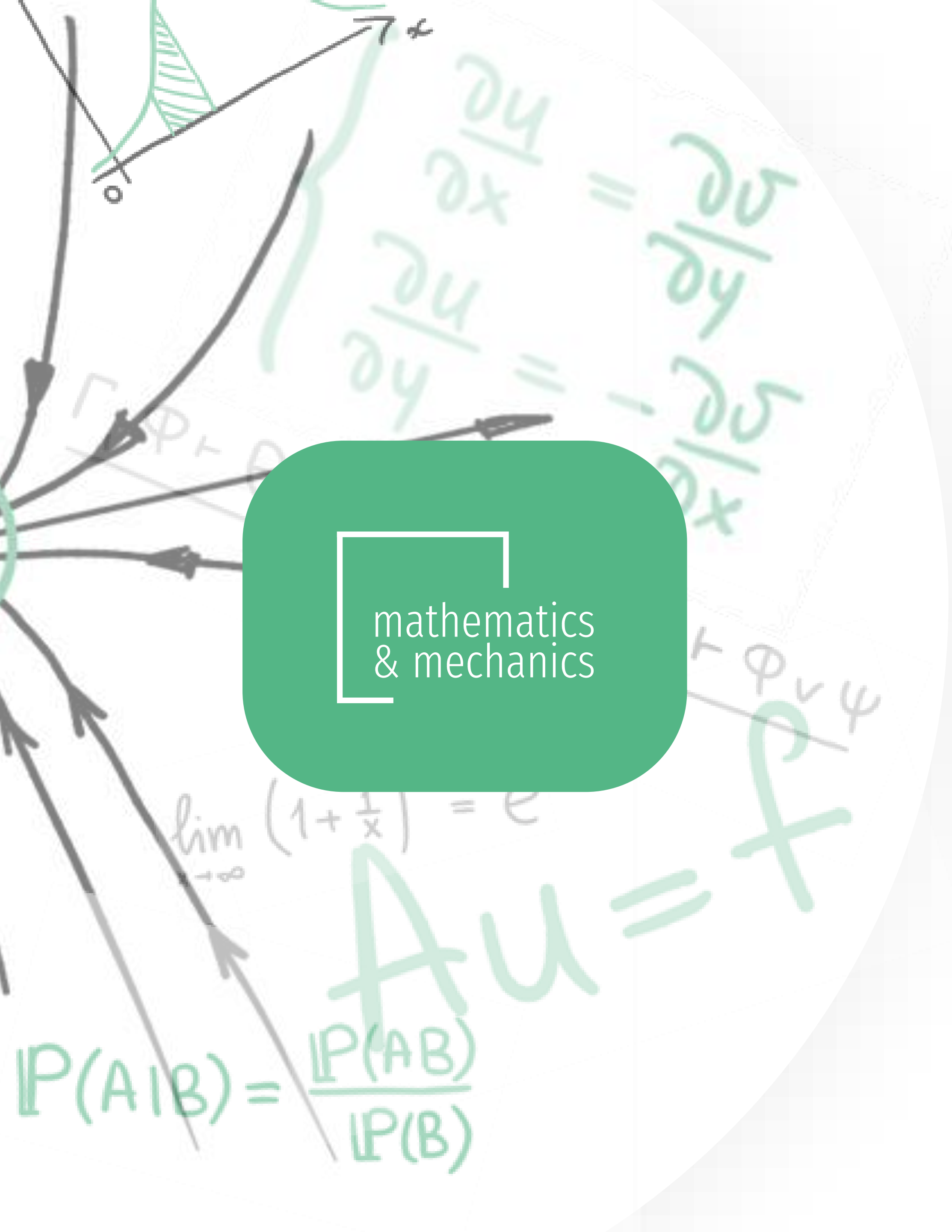


INVARIANT RISKS WITHOUT KNOWLEDGE OF THE ENVIRONMENT

Authors: Bratenkov Miron & Bondarenko Ivan.

27 march 2025

mathematics & mechanics

PROBLEM

Most machine learning models are based on the minimization of empirical risk, under the assumption that all data belong to a single distribution.

The minimization of empirical risk results in low predictive quality for data that do not conform to the distribution of the training set, a challenge known as the "Data Shift Problem" or "Out-of-Distribution Generalization" (OOD).

INVARIANT RISK MINIMIZATION

1. Developed to address the issue of data shifts and enhance model robustness against them.

2. Requires the partitioning of training data into different environments.

3. Mathematically represents the learning model as a composition of a data transformation model and a classifier model.

Mathematical notation:

$$\min_{\substack{\Phi:\to H \\ w:H\to Y}} \sum_{e\in E} R^e(w\circ\Phi)$$

$$\text{Where} \quad w \in \underset{\hat{w}:H\to Y}{argmin} R^e(\hat{w}\circ\Phi) \quad \text{for all} \quad e\in E$$

$$\text{Where} \quad \Phi - \text{data transformation model,}$$

$$w - \text{classifier,} \quad E - \text{set of environments,}$$

$$R - \text{error.}$$

Practically applicable formula:

$$L(\theta) = L_u(\theta) + \lambda \cdot Penalty(\theta)$$

source: arXiv:1907.02893

mathematics
& mechanics

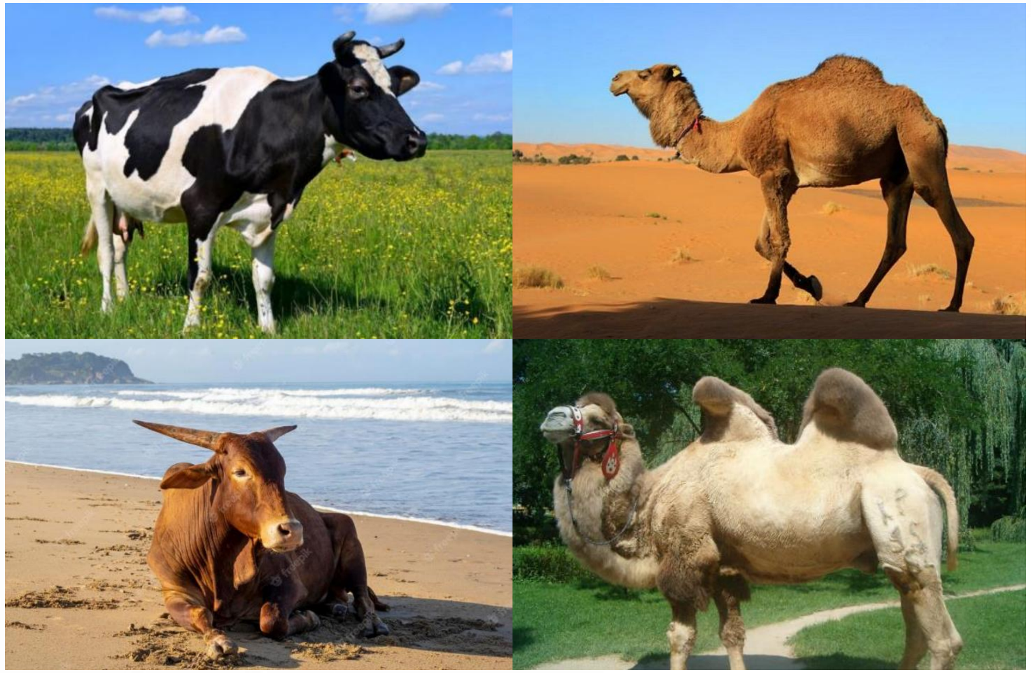
The research aim

Extend the invariant risk minimization paradigm to problems without data partitioning across environments and improve the resilience of models trained within these paradigms to data shifts.

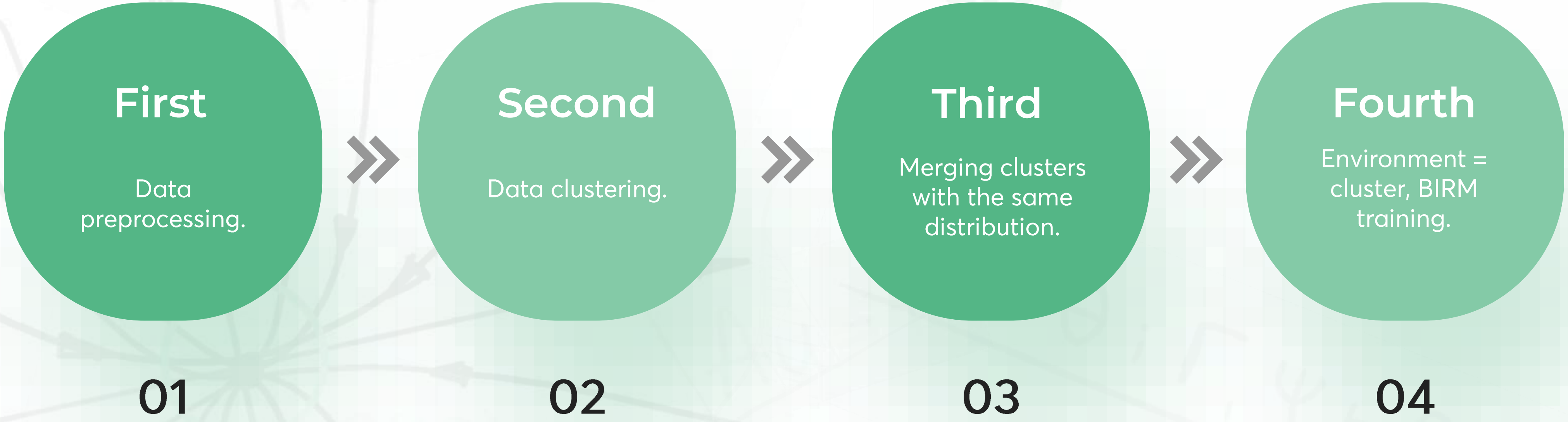
ENVIRONMENT

In the context of environmental factors, one should understand them as a set of characteristics that can introduce spurious correlations in the predictions generated by the model.

Environments, in this context, should be understood as data sets in which the environmental features display a same distribution.



LEARNING ALGORITHM



First

Data preprocessing.

01

Second

Data clustering.

02

Third

Merging clusters with the same distribution.

03

Fourth

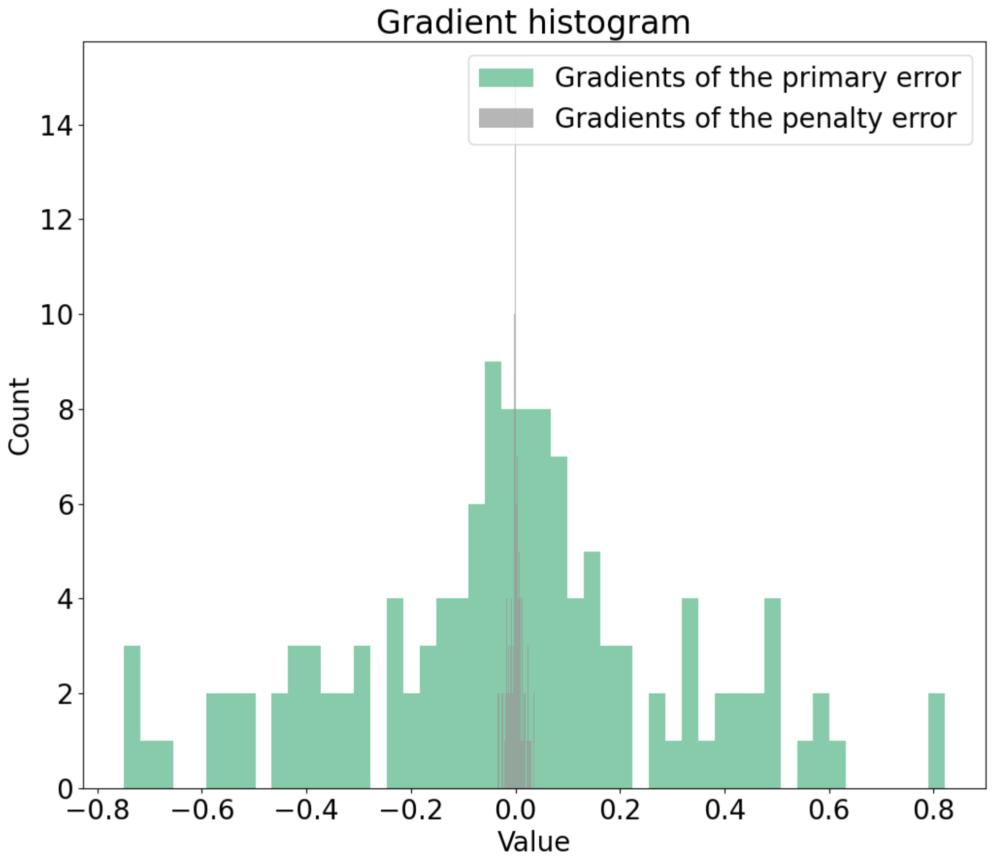
Environment = cluster, BIRM training.

04

PROBLEM - CHOICE OF HYPERPARAMETER λ

1. The static λ does not provide a qualitative improvement and may hinder the training process.
2. The selection of hyperparameters is of paramount importance during the training phase.
3. Gradual adjustments to hyperparameters demand a significant amount of time for optimization.

SOLUTION - ADAPTIVE HYPERPARAMETER



$$L(\theta) = L_u(\theta) + \lambda \cdot Penalty(\theta)$$

$$D(\nabla_\theta Penalty(\theta_n)|_{\text{last layer}}) = \sigma_p$$

$$D(\nabla_\theta L_u(\theta_n)|_{\text{last layer}}) = \sigma_u$$

$$\lambda = \frac{\sigma_u}{\sigma_p}$$

$$\lambda_{n+1} = \alpha \cdot \lambda_n + \lambda$$

SYNTHETIC PROBLEM

IN 480 TIMES

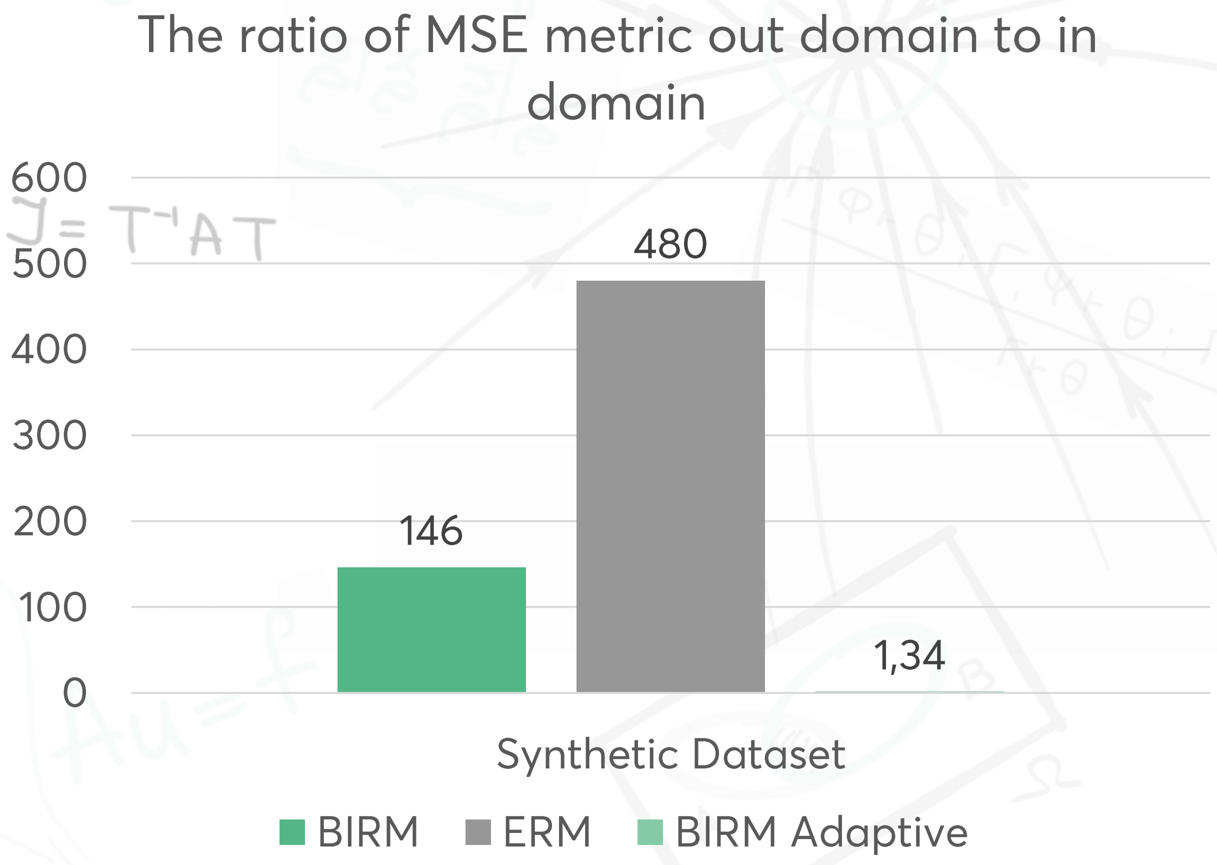
MSE WITH ERM ON OUT MORE THEN MSE ON IN DOMAIN

IN 146 TIMES

MSE WITH BIRM ON OUT MORE THEN MSE ON IN DOMAIN

IN 1,34 TIMES

MSE WITH BIRM ADAPTIVE ON OUT MORE THEN MSE ON IN DOMAIN



The ratio of MSE metric out domain to in domain

THE PROBLEM OF WEATHER PREDICTION

IN 1,35 TIMES

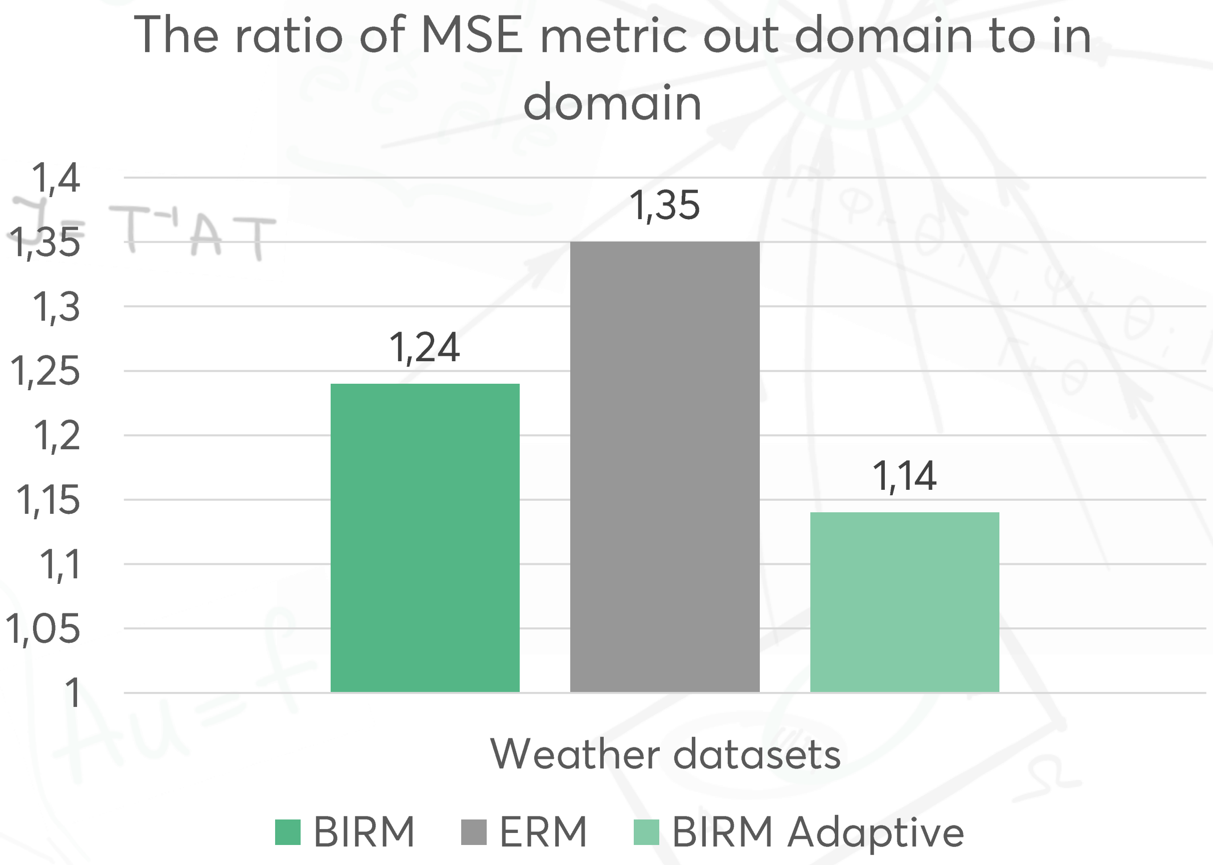
MSE WITH ERM ON OUT MORE THEN MSE ON IN DOMAIN

IN 1,24 TIMES

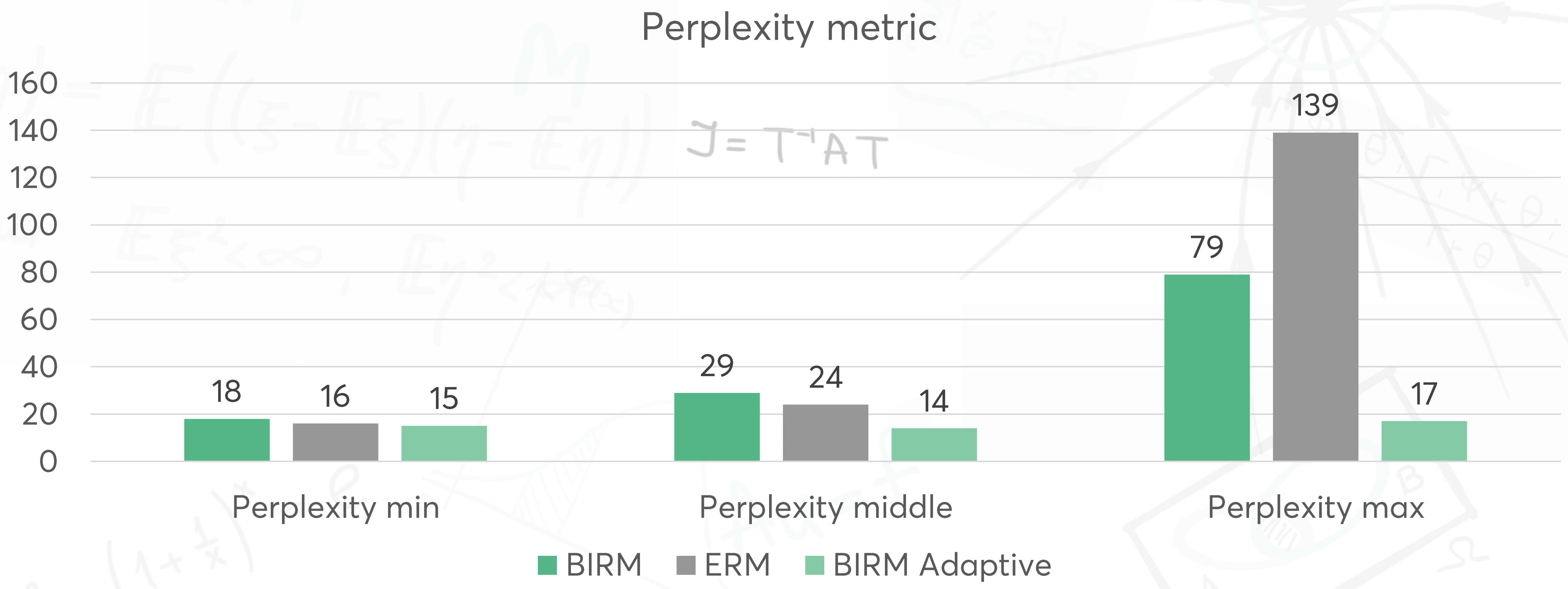
MSE WITH BIRM ON OUT MORE THEN MSE ON IN DOMAIN

IN 1,14 TIMES

MSE WITH BIRM ADAPTIVE ON OUT MORE THEN MSE ON IN DOMAIN



TEXT GENERATION PROBLEM



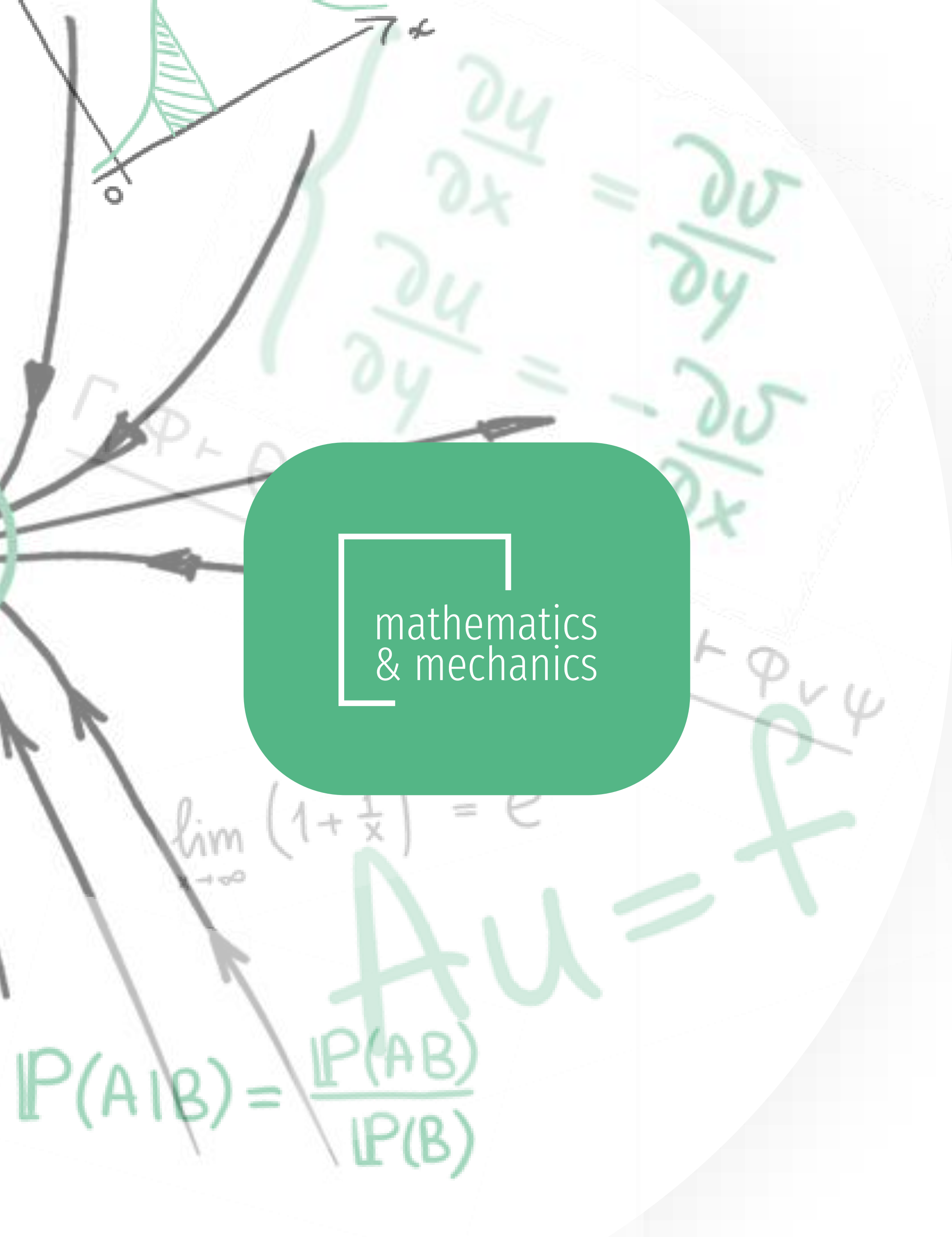
mathematics
& mechanics

CONCLUSION

1. The proposed algorithm for training models demonstrates resilience to data shifts, leveraging clustering techniques combined with Invariant Risk Minimization (IRM).

2. A modification of the paradigm for model training focused on minimizing invariant risk through adaptive hyperparameter tuning has been introduced.

3. Results are presented, indicating that models trained using the proposed algorithm exhibit enhanced robustness against data shifts.



mathematics & mechanics

THANKS FOR ATTENTION

27 march 2025

SYNTHETIC DATASET

$$\begin{cases} X_{inv} \sim N(E_2, I_2) \\ Y = 1^T \cdot X_{inv} + N(0, 0.1) \\ X_{env} = Y + N(E_2, pe \cdot I_2) \\ X = (X_{inv}; X_{env}) \end{cases}$$

$$pe_{train} = [0.1; 0.3; 0.5; 0.7; 0.9]$$

$$pe_{val} = [0.4; 0.8]$$

$$pe_{test} = [10; 100]$$

QUALITY METRICS

1. In-domain – the data distribution corresponds to the distribution of the training set.

2. Out-of-domain – the data distribution falls outside that of the training set.

3. Quality metric – the ratio of the mean squared error for out-of-domain data to the mean squared error for in-domain data.

$$Metrica = \frac{MSE_{out\,domain}}{MSE_{in\,domain}}$$

mathematics
& mechanics

WEATHER DATASET

1. The weather dataset comprises pairs of meteorological characteristics (123 values) and the target variable, which is temperature at a height of 2 meters.

2. K-means clustering was performed with 30 clusters based on the length of tokens.

3. Out-of-distribution data was sourced from regions with snowy and polar climates after May 14, 2019. The training data includes tropical, humid, and temperate climates from September 1 to April 8, 2019.

4. The data was obtained from the Shifts Challenge 2019 competition.

mathematics
& mechanics

TEXT DATASET

1. The Taiga dataset is a corpus comprising artistic texts, poetry, news articles, and other types of texts.

2. K-means clustering was performed with a specification of three clusters based on token length.

3. The average number of tokens per example across the clusters is as follows: 43, 235, and 800.

4. Data that falls outside the distribution corresponds to the cluster with the highest average number of tokens per example. The remaining clusters serve as the training data.

QUALITY METRICS

1. The lower the PPL (Perplexity), the more confident the model is in its predictions.
2. Conversely, the higher the PPL, the more the model struggles to make accurate predictions based on the input data.

Perplexity:

$$PPL(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t} \log p(x_i|x_{<i})\right\}$$

Where $\log p(x_i|x_{<i})$ denotes the logarithmic likelihood of the i-th token, conditioned on the tokens with indices $< i$.

TABLE DATA OF RESULTS

Synthetic Dataset

	BIRM	ERM	Adaptive BIRM
In domain	0,022	0,01	1,7
Out domain	3,2	4,8	2,3

Weather Dataset

	BIRM	ERM	Adaptive BIRM
In domain	0,39	0,36	0,43
Out domain	0,49	0,48	0,49
train	0,38	0,36	0,43