# Tri-Level Navigator: LLM-Empowered Tri-Level Learning for Time Series OOD Generalization

Chengtao Jian, Kai Yang, Yang Jiao

School of Computer Science and Technology, Tongji University, Shanghai, China

NeurIPS 2024, October 18, 2024

同濟大學
TONGJI UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

# Outline

# Introduction

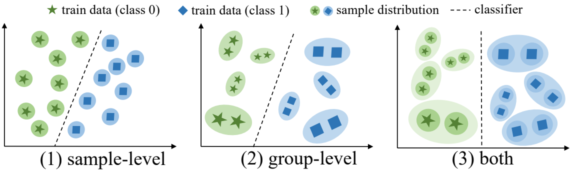
- **Background**: Out-of-Distribution (OOD) generalization is a key challenge in machine learning when training and test distributions differ significantly.
- **Problem**: Time series data present unique challenges for OOD generalization due to temporal dependencies and dynamic changes.
- **Goal**: We propose a novel tri-level learning framework (TTSO) based on Large Language Models (LLMs) for time series OOD generalization.

# Motivation

- **Current State**: Extensive research has focused on OOD generalization in vision and text domains, but limited work exists for time series.
- **Pre-trained Models**: LLMs like GPT demonstrate strong generalization and representation learning capabilities across modalities, including time series.
- **Solution**: Our TTSO framework leverages LLMs and tri-level optimization to address time series OOD generalization challenges.

# Related Work

- **OOD Generalization**: Previous work focuses on sample-level or group-level uncertainties.
- **LLMs in Time Series**: An emerging field where LLMs show promise in time series analysis.
- **Contribution**: TTSO uniquely integrates sample-level and group-level uncertainties into a tri-level learning problem with LLMs to enhance OOD generalization.



# Tri-Level Learning Framework

- **Overview**: TTSO addresses sample and group uncertainties through tri-level optimization:

$$
\begin{aligned}
\min_{\boldsymbol{\theta},\boldsymbol{q},\boldsymbol{\delta}} \quad & \textstyle\sum_{i=1}^{K} q_i \ell_{\text{con}}\left(\boldsymbol{\theta}, \boldsymbol{\delta}; \mathcal{D}_{S_i}\right) \\
\text{s.t.} \quad & \boldsymbol{q} = \arg\max_{\boldsymbol{q}' \in \Delta^K} \textstyle\sum_{i=1}^{K} q_i' \ell_{\text{con}}\left(\boldsymbol{\theta}, \boldsymbol{\delta}; \mathcal{D}_{S_i}\right) \\
& \text{s.t. } d(\boldsymbol{p}, \boldsymbol{q}') \leq \tau \\
& \qquad \boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta}' \sim p(\boldsymbol{\delta}'; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})} \textstyle\sum_{i=1}^{K} q_i' \ell_{\text{align}}(\boldsymbol{\theta}, \boldsymbol{\delta}'; \mathcal{D}_{S_i}) \\
& \qquad \text{s.t. } \|\boldsymbol{\mu}\| \leq C_1, \|\boldsymbol{\sigma}\| \leq C_2, \textstyle\sum_{m=1}^{M} \pi_m = 1, \pi_m \geq 0,
\end{aligned} \tag{1}
$$

- **Advantages**: Tackles both sample and group uncertainties, providing strong theoretical and practical generalization benefits.

# Stratified Localization Algorithm

- **Step 1**: Use gradient descent and Taylor approximation to transform the tri-level problem into a single-layer optimization problem.
- **Step 2**: Generate cutting planes to approximate the non-convex feasible region.
- **Step 3**: Iteratively refine the solution using the cutting planes.
- **Step 4**: Update variables and add new cutting planes every k iterations to ensure tighter approximation.
- **Step 5**: Return the optimized parameters after satisfying convergence criteria.

# Theoretical Analysis

- **Convergence Analysis**: The Stratified Localization Algorithm (SLA) is proven to converge to an $\epsilon$-stationary point.
- **Iteration Complexity**: The iteration complexity is given by:

$$T(\epsilon) \sim \mathcal{O}\left(t_1 + \frac{L^2(m(\alpha_1^2+\alpha_2^2+\alpha_3^2)+\sigma_1^2+\sigma_2^2+\sigma_3^2)^2}{4m^2(\epsilon - F(\boldsymbol{\theta}^{T_1}, \boldsymbol{q}^{T_1}, \boldsymbol{\delta}^{T_1}) + F^*)^2}\right), \tag{2}$$

- **Generalization Bound**: Based on VC (Vapnik-Chervonenkis) dimension theory, the generalization properties is given by:

$$\begin{aligned} \epsilon_T(\hat{h}) \leq 3\epsilon_T(h_T^*) + \lambda + d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{P}_\mathsf{C}, \mathbb{P}_\mathsf{T}) \\ + \max_{i,j} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{P}_{S_i}, \mathbb{P}_{S_j}) + C(\delta, m, d), \end{aligned} \tag{3}$$

where $\lambda$ and $C(\delta, m, d)$ is a statistical term. $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ is a metric function which measures differences in distribution. $\epsilon_{S_i}(h)$ and $\epsilon_T(h)$ is the source error and the target error.

# Experiments

- **Datasets**: Experiments conducted on six real-world time series datasets (HHAR, PAMAP, WESAD, etc.).
- **Baselines**: Compared against traditional OOD methods (ERM, IRM, GroupDRO) and time-series-specific approaches (AdaRNN, GILE, DIVERSIFY, DFDDG, CCDG).
- **Results**: TTSO achieves an maximum 4.9% improvement in performance on time series classification in OOD scenarios.

# Conclusion

- The TTSO framework integrates sample-level and group-level uncertainties with LLMs for time series OOD generalization.
- TTSO demonstrates significant improvements in real-world datasets, outperforming existing methods.
- Future Work: Explore the application of TTSO on other modalities beyond time series, such as image, audio.