Analysing Multi-Task Regression via Random Matrix Theory

Romain Ilbert[1,2] **Malik Tiomoko**[1] **Cosme Louart**[3] **Ambroise Odonnat**[1,4] **Vasilii Feofanov**[1] **Themis Palpanas**[2] **Ievgen Redko**[1]

[1]Huawei Noah's Ark Lab [2]LIPADE, Paris Descartes University [3]School of Data Science, The Chinese University of Hong Kong, Shenzhen, China [4]Inria, Univ. Rennes 2, CNRS, IRISA

SPOTLIGHT

NEURAL INFORMATION PROCESSING SYSTEMS

TL;DR

- We introduce a **theoretical framework** for **multi-task regression** using **random matrix theory**, providing analytic solutions and insights into task relationships.
- By framing multi-task optimization as a **regularization technique**, we enable **single-task models** to leverage **multi-task learning** benefits.
- Our analysis offers **consistent estimations** of training and testing errors, facilitating effective **hyperparameter optimization**.
- Experiments on **synthetic** and **real-world datasets** in **regression** and **multivariate time series forecasting** demonstrate that our method significantly **improves** univariate models when incorporated into the training loss.

Problem Setup

Goal: Leverage shared information across multiple related tasks to improve overall performance in multi-task regression.

- Consider T regression tasks, each with its own input space $\mathcal{X}^{(t)} \subset \mathbb{R}^d$ and output space $\mathcal{Y}^{(t)} \subset \mathbb{R}^q$, for $t = 1, \dots, T$.
- For each task t, we have n_t training examples comprising a feature matrix $\mathbf{X}^{(t)} \in \mathbb{R}^{d \times n_t}$, and response matrix $\mathbf{Y}^{(t)} \in \mathbb{R}^{q \times n_t}$.
- Linear multi-task regression model:

$$\mathbf{Y}^{(t)} = \frac{\mathbf{X}^{(t)\top}\mathbf{W}_t}{\sqrt{Td}} + \boldsymbol{\varepsilon}^{(t)}, \quad \forall t = 1, \dots, T,$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times q}$ is the signal-generating hyperplane for task t, and $\boldsymbol{\varepsilon}^{(t)}$ is a noise matrix with entries drawn from $\mathcal{N}(0, \boldsymbol{\Sigma}_N)$.

Weights Decomposition

- Each $\mathbf{W}_t$ is decomposed into:

$$\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t,$$

where $\mathbf{W}_0$ is the common component shared across all tasks, and $\mathbf{V}_t$ is the task-specific deviation.

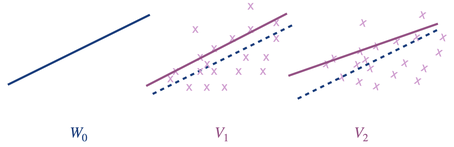
- Objective: Estimate $\mathbf{W}_0$ and $\{\mathbf{V}_t\}$ by solving:

$$\min_{\mathbf{W}_0, \{\mathbf{V}_t\}, \lambda} \frac{1}{2\lambda}\|\mathbf{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\mathbf{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^{T}\left\|\mathbf{Y}^{(t)} - \frac{\mathbf{X}^{(t)\top}\mathbf{W}_t}{\sqrt{Td}}\right\|_F^2,$$

where λ and γ_t are regularization parameters controlling the trade-off between shared and task-specific components.

The objective function consists of three components:

- A regularization term for $\mathbf{W}_0$ to mitigate overfitting,
- Task-specific regularization terms controlling deviations $\mathbf{V}_t$ from the shared weight matrix $\mathbf{W}_0$.
- A loss term quantifying the error between the predicted outputs and the actual responses for each task.



W_0 : **A shared component that enhances the performance of each task**

Main Theoretical Results

Interpretation and Key Insights

- We decompose train and test risks into **signal** and **noise** terms to gain insights into our model's behavior and to find the optimal regularization parameter λ^*.
- λ^* indicates the point at which we want to **leverage multivariate information**.
- The **signal term** capture how well the model learns the underlying tasks, while the **noise term** represent the impact of noise on performance.
- This competition between signal and noise allows us to determine the optimal λ^*.

Main Theorem: Asymptotic Train and Test Risks
The test and train risks decompose into signal and noise contributions:

Test Risk:

$$\mathcal{R}_{\text{test}}^{\infty} = \underbrace{\frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-\frac{1}{2}} \tilde{\mathbf{Q}}_2(\mathbf{A}) \mathbf{A}^{-\frac{1}{2}} \mathbf{W}\right)}{Td}}_{\text{Signal Term}} + \underbrace{\frac{\operatorname{tr}(\boldsymbol{\Sigma}_N \tilde{\mathbf{Q}}_2)}{Td}}_{\text{Noise Terms}} + \operatorname{tr}(\boldsymbol{\Sigma}_N).$$

Train Risk:

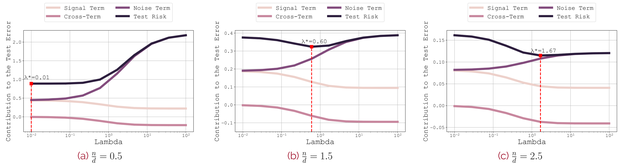
$$\mathcal{R}_{\text{train}}^{\infty} = \underbrace{\frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-\frac{1}{2}} \tilde{\mathbf{Q}} \mathbf{A}^{-\frac{1}{2}} \mathbf{W}\right)}{Tn}}_{\text{Signal Term}} - \underbrace{\frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-\frac{1}{2}} \tilde{\mathbf{Q}}_2(\mathbf{I}_{Td}) \mathbf{A}^{-\frac{1}{2}} \mathbf{W}\right)}{Tn}}_{\text{Signal Term}} + \underbrace{\frac{\operatorname{tr}\left(\boldsymbol{\Sigma}_N \tilde{\mathbf{Q}}_2\right)}{Tn}}_{\text{Noise Term}}.$$

Error Contribution Analysis

The signal-noise competition makes the test-curve convex and allows us to identify λ^* to solve our optimization problem.

Implications:

- Increasing λ promotes shared learning by emphasizing the signal term.
- However, a too large λ can amplify the noise term, degrading performance.
- Therefore, finding λ^* is crucial to balance enhancing the signal and suppressing the noise.



(a) $\frac{n}{d} = 0.5$ (b) $\frac{n}{d} = 1.5$ (c) $\frac{n}{d} = 2.5$

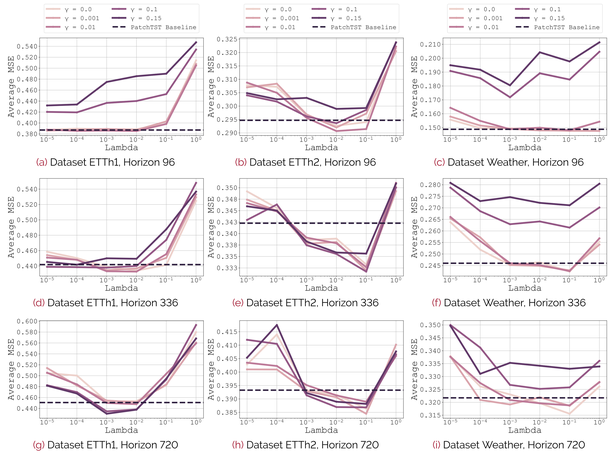
Test loss contributions $\mathcal{D}_{\text{IL}}$, $\mathcal{C}_{\text{MTL}}$, $\mathcal{N}_{\text{NT}}$ across three sample size regimes. The convexity of the test risk curve allows us to identify λ^* to solve our optimization problem.

Beyond the Case of Linear Models: Multivariate Time Series Forecasting

- We apply our theoretical framework to Multivariate Time Series Forecasting . We compare models with and without MTL regularization.
- Evaluation conducted on common open-source benchmarks of various scales.

Dataset	H	with MTL regularization			without MTL regularization					
		PatchTST	LinearU	Transformer	PatchTST	LinearU	LinearT	Transformer	SAMformer	iTransformer
ETTh1	96	0.385	**0.367***	0.368	0.387	0.397	0.386	0.370	0.381	0.386
	192	0.422	**0.405***	0.407*	0.424	0.422	0.437	0.411	0.409	0.441
	336	0.433*	0.431	0.433	0.442	0.431	0.481	0.437	**0.423**	0.487
	720	0.430*	0.454	0.455*	0.451	0.428	0.519	0.470	**0.427**	0.503
ETTh2	96	0.291	**0.267***	0.270	0.295	0.294	0.333	0.273	0.295	0.297
	192	0.346*	**0.331***	0.337	0.351	0.361	0.477	0.339	0.340	0.380
	336	**0.332***	0.367	0.366*	0.342	0.361	0.594	0.369	0.350	0.428
	720	**0.384***	0.412	0.405*	0.393	0.395	0.831	0.428	0.391	0.427
Weather	96	**0.148**	0.149*	0.154*	0.149	0.196	0.196	0.170	0.197	0.174
	192	**0.190**	0.206*	0.198*	0.193	0.243	0.237	0.214	0.235	0.221
	336	**0.242***	0.249*	0.258	0.246	0.283	0.283	0.260	0.276	0.278
	720	**0.316***	0.326*	0.331	0.322	0.339	0.345	0.326	0.334	0.358

Our multi-task regularization makes a univariate linear model state-of-the-art in multivariate time series forecasting and improves transformer-based model performances as well.



(a) Dataset ETTh1, Horizon 96 (b) Dataset ETTh2, Horizon 96 (c) Dataset Weather, Horizon 96

(d) Dataset ETTh1, Horizon 336 (e) Dataset ETTh2, Horizon 336 (f) Dataset Weather, Horizon 336

(g) Dataset ETTh1, Horizon 720 (h) Dataset ETTh2, Horizon 720 (i) Dataset Weather, Horizon 720

The non-linear model curves align with the theoretical patterns observed in linear models, and our MTL regularization enhances transformer performance across four forecasting horizons.

Main References

- **Ilbert et al.** - **ICML 2024 (Oral)**
  SAMformer
- **Tiomoko et al.** - **PrePrint 2020**
  Large Dimensional Analysis
- **Yang et al.** - **ICML 2023**
  Precise High-Dimensional Asymptotics

Romain Ilbert

Malik Tiomoko

Multi-Task Regression

We apply our theoretical framework to the **Appliance Energy** dataset, containing 138 time series of dimension 24, and reformulate it as a **multi-task regression** by selecting two features as tasks and setting $\gamma_1 = \gamma_2 = \gamma$.

Optimal Regularization Parameter λ^*:

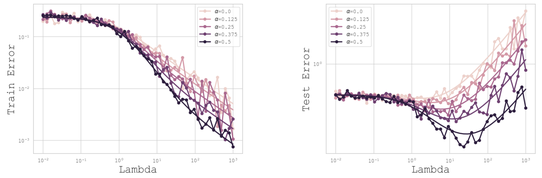
$$\lambda^* = \frac{n}{d}\text{SNR} - \frac{\gamma}{2}, \quad \text{where} \quad \text{SNR} = \frac{\|\mathbf{W}_1\|_2^2 + \|\mathbf{W}_2\|_2^2}{\operatorname{tr}\boldsymbol{\Sigma}_N} + \frac{\mathbf{W}_1^\top \mathbf{W}_2}{\operatorname{tr}\boldsymbol{\Sigma}_N}.$$

Interpretation and Key Insights

- α represents the **cosine similarity** between $\mathbf{W}_1$ and $\mathbf{W}_2$, i.e., $\alpha = \frac{\mathbf{W}_1^\top \mathbf{W}_2}{\|\mathbf{W}_1\|_2 \|\mathbf{W}_2\|_2}$.
- **As α increases**, the tasks become more similar, and λ^* **increases**.
- Larger λ^* reduces the penalty on the **shared component**, enhancing the use of multivariate information.
- Thus, similar tasks benefit from shared learning, which **improves performance**.

$$\mathbf{W}_1 \sim \mathcal{N}(0, I_p), \quad \mathbf{W}_2 = \alpha \mathbf{W}_1 + \sqrt{1-\alpha^2}\mathbf{W}_1^\perp,$$

where $\mathbf{W}_1^\perp$ is any vector orthogonal to $\mathbf{W}_1$, and $\alpha \in [0, 1]$.



Theoretical predictions (smooth curves) closely match the empirical results (dots). As α increases, λ^* also increases, indicating less penalty on the shared component.

Experimental Results:

- Experiments confirm that **optimal λ^* increases with task similarity** α.
- Theoretical predictions closely match empirical results, demonstrating the **accuracy** of our analysis.
- Our framework effectively guides **hyperparameter selection** in real-world multi-task learning.

Conclusion and Takeaways

- We developed a theoretical framework providing **optimal regularization parameters** for multi-task regression, effectively balancing signal and noise to enhance performance.
- Our theoretical predictions **closely align with empirical results**
- Applying our framework to **multivariate time series forecasting**, we demonstrate that **simpler models can achieve state-of-the-art results**.

Paper

Correspondence to romain.ilbert@hotmail.fr