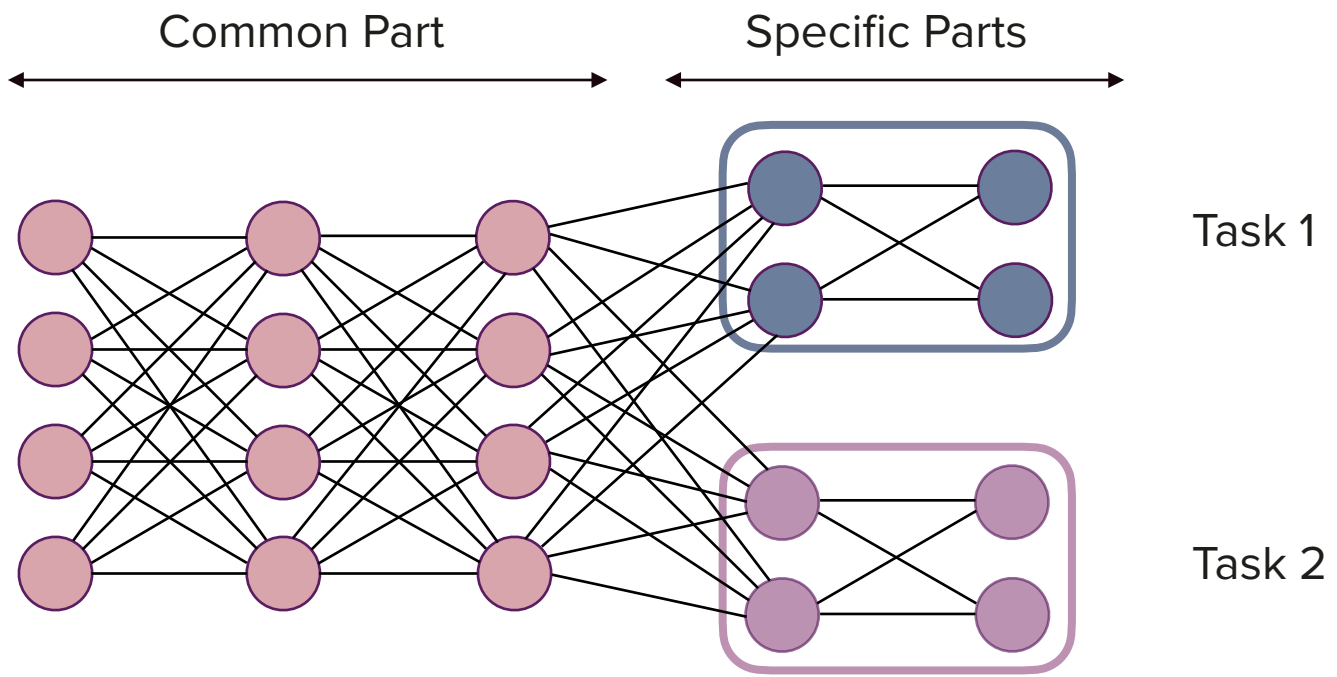
Analysing Multi-Task Regression via Random Matrix Theory

Romain ILBERT, Malik Tiomoko, Cosme Louart, Vasilii Feofanov,
Themis Palpanas, Ievgen Redko

NEURAL INFORMATION
PROCESSING SYSTEMS

Multi-Task Regression : Definition

- Multi-task learning: inspired by human intelligence, enabling knowledge transfer
- It leverages shared information across tasks to boost overall performance.
- Key benefits: enhanced accuracy and structured representations from diverse and multimodal data.
- Successfully applied in fields like computer vision, NLP, and biology.



Multi-Task Regression : Problem Setup

- We consider *T tasks* with the input space $\mathcal{X}^{(t)} \subset \mathbb{R}^d$ and the output space $\mathcal{Y}^{(t)} \subset \mathbb{R}^q$
- We consider n_t samples
- We consider the input matrix $\mathbf{X}^{(t)} = \left[\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{n_t}^{(t)}\right] \in \mathbb{R}^{d \times n_t}, \mathbf{x}_i^{(t)} \in \mathcal{X}^{(t)}$
- We consider the the output matrix $\mathbf{Y}^{(t)} = \left[\mathbf{y}_1^{(t)}, \ldots, \mathbf{y}_{n_t}^{(t)}\right] \in \mathbb{R}^{q \times n_t}, \mathbf{y}_i^{(t)} \in \mathcal{Y}^{(t)}$

Regression : To learn $\mathbf{W}_t \in \mathbb{R}^{d \times q}$ such that :

$$\forall t \in \{1, \ldots, T\}, \quad \mathbf{Y}^{(t)} = \frac{\mathbf{X}^{(t)}\mathbf{W}_t}{\sqrt{d}} + \epsilon^{(t)}$$

with $\epsilon^{(t)} \in \mathbb{R}^{n_t \times q}, \epsilon_i^{(t)} \sim \mathcal{N}(0, \Sigma_N), \Sigma_N \in \mathbb{R}^{q \times q}$

Multi-Task Regression : Regularization objective

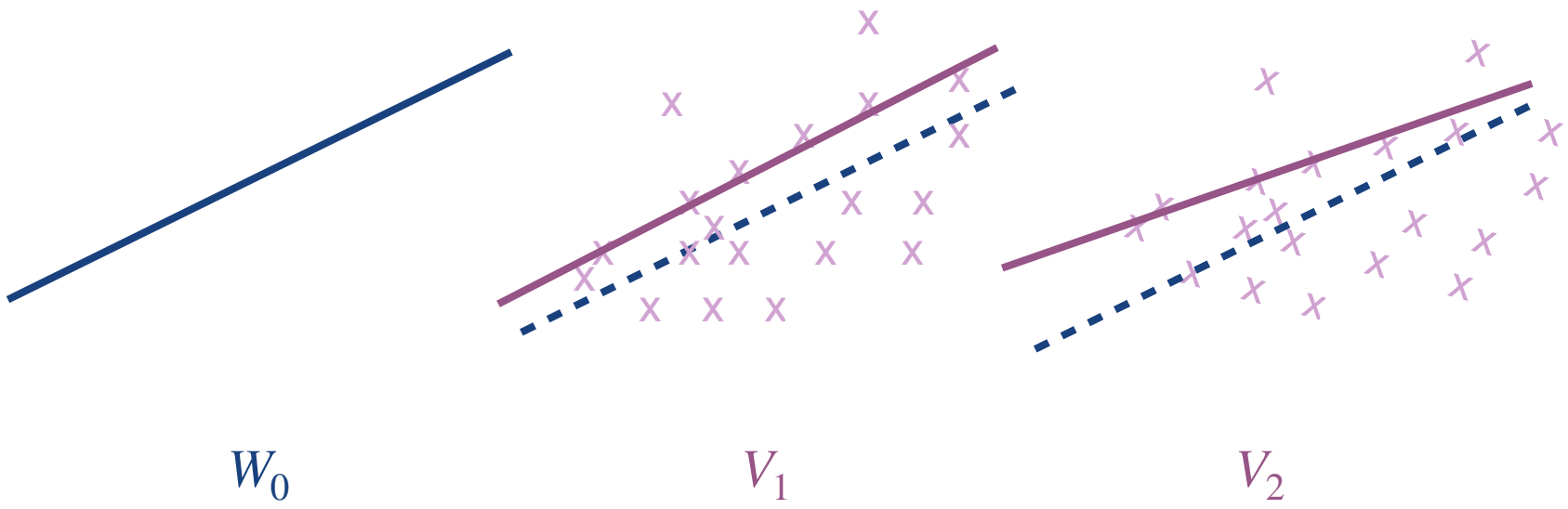
- We propose to decompose the weights to learn into $\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t$
- $\mathbf{W}_0 \in \mathbb{R}^{d \times q}$ is a common matrix that captures information across all the *T* tasks
- $\mathbf{V}_t \in \mathbb{R}^{d \times q}$ is a task-specific matrix which captures deviations specific to task *t*

The following minimization problem governed by a parameter λ that controls the balance between the common and specific components of $\mathbf{W} = \left[\mathbf{W}_1, \ldots, \mathbf{W}_T\right]^\top \in \mathbb{R}^{Td \times q}$:

$$\mathbf{W}_0^*, \{\mathbf{V}_t^*\}_{t=1}^T, \lambda^* = \arg\min \frac{1}{2\lambda}\|\mathbf{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\mathbf{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^{T}\left\| \mathbf{Y}^{(t)} - \frac{\mathbf{X}^{(t)}\mathbf{W}_t}{\sqrt{d}} \right\|_F^2$$

with $\gamma = [\gamma_1, \ldots, \gamma_T]$

Multi-Task Regression : W decomposition



Main contributions and results

- **Random Matrix Theory**: Exact computation of train and test risks and decomposition of test risk into signal (effectiveness) and noise (negative transfer) terms.

- **Test Risk Optimization**: We show how the signal and noise terms compete with each other depending on λ for which we obtain an optimal value optimizing the test risk.

- **We derived a closed-form solution for λ^*** based on data covariances, signal-generating hyperplanes, noise levels, and dataset size.

- **We demonstrate the patterns observed in real-world regression problems with linear models also apply to neural networks** in the context of multivariate time series forecasting (MTSF).

- **By obtaining λ^*, we make a simple univariate linear model outperform the current sota models.**

Assumptions

Assumption 1: *Concentrated Random Vector*

We assume that there exists two constants $C, c > 0$ (independent of dimension *d*) such that, for any task *t*, for any 1-Lipschitz function f, any feature vector $\mathbf{x}^{(t)}$ verifies :

$$\forall t > 0 : \mathbb{P}(\,|f(\mathbf{x}^{(t)}) - \mathbb{E}[f(\mathbf{x}^{(t)})]\,| \geq t) \leq Ce^{-(t/c)^2}, \quad \mathbb{E}[\mathbf{x}^{(t)}] = 0 \quad and \quad Cov[\mathbf{x}^{(t)}] = \Sigma^{(t)}$$

Assumption 2: *High-Dimensional Asymptotics*

As $d \to \infty, n_t = \mathcal{O}(d)$ and $T = \mathcal{O}(1)$. More specifically, we assume that $\frac{n}{d} \xrightarrow{\textbf{a.s.}} c_0 < \infty$ **with** $n = \sum_{t=1}^{T} n_t$.

Main Theoretical Results

Theorem 1 (Asymptotic train and test risk). *Assuming that the training data vectors* $\mathbf{x}_i^{(t)}$ *and the test data vectors* $\mathbf{x}^{(t)}$ *are concentrated random vectors, and given the growth rate assumption (Assumption 2), it follows that:*

$$\mathcal{R}_{test}^{\infty} = \underbrace{\frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-\frac{1}{2}} \bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}) \mathbf{A}^{-\frac{1}{2}} \mathbf{W}\right)}{Td}}_{\textit{signal term}} + \underbrace{\frac{\operatorname{tr}(\boldsymbol{\Sigma}_n \bar{\mathbf{Q}}_2)}{Td} + \operatorname{tr}\left(\boldsymbol{\Sigma}_n\right)}_{\textit{noise terms}}. \tag{ATR}$$

In addition, the asymptotic risk on the training data is given by

$$\mathcal{R}_{train}^{\infty} \leftrightarrow \frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-1/2} \bar{\tilde{\mathbf{Q}}} \mathbf{A}^{-1/2} \mathbf{W}\right)}{Tn} - \frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{A}^{-1/2} \bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_{Td}) \mathbf{A}^{-1/2} \mathbf{W}\right)}{Tn} + \frac{\operatorname{tr}\left(\boldsymbol{\Sigma}_n \bar{\mathbf{Q}}_2\right)}{Tn},$$

where $\bar{\tilde{\mathbf{Q}}}$, $\bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_{Td})$ *and* $\bar{\mathbf{Q}}_2$ *are respectively deterministic equivalents for* $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{Q}}^2$ *and* $\mathbf{Q}^2$.

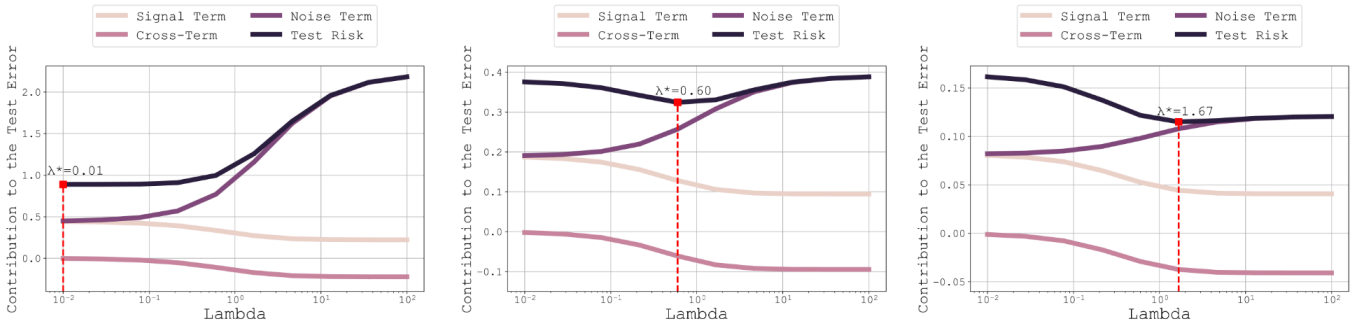
Main Theoretical Results

$$\mathcal{R}_{test}^{\infty} = \mathbf{D}_{IL}\left(\|\mathbf{W}_1\|_2^2 + \|\mathbf{W}_2\|_2^2\right) + \mathbf{C}_{MTL}\mathbf{W}_1^\top \mathbf{W}_2 + \mathbf{N}_{NT}\mathrm{tr}\boldsymbol{\Sigma}_n$$

where the diagonal term (independent learning) $\mathbf{D}_{IL}$, the cross term (multi-task learning) $\mathbf{C}_{MTL}$, and the noise term (negative transfer) $\mathbf{N}_{NT}$ have closed-form expressions depending on γ and λ:

$$\mathbf{D}_{IL} = \frac{(c_0(\lambda+\gamma)+1)^2 + c_0^2\lambda^2}{(c_0(\lambda+\gamma)+1)^2 - c_0^2\lambda^2}, \quad \mathbf{C}_{MTL} = \frac{-2c_0\lambda(c_0(\lambda+\gamma)+1)}{(c_0(\lambda+\gamma)+1)^2 - c_0^2\lambda^2}$$

$$\mathbf{N}_{NT} = \frac{(c_0(\lambda+\gamma)^2 + (\lambda+\gamma) - c_0\lambda^2)^2 + \lambda^2}{\left((c_0(\lambda+\gamma)+1)^2 - c_0^2\lambda^2\right)^2}$$



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

Figure 1: Test loss contributions $\mathbf{D}_{IL}$, $\mathbf{C}_{MTL}$, $\mathbf{N}_{NT}$ across three sample size regimes. Test risk exhibits decreasing, increasing, or convex shapes based on the regime. Optimal values of λ from theory are marked.

Comparison between practice and theory

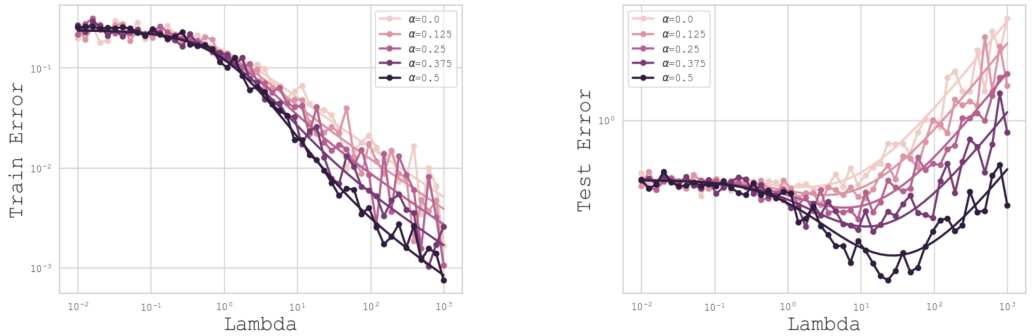
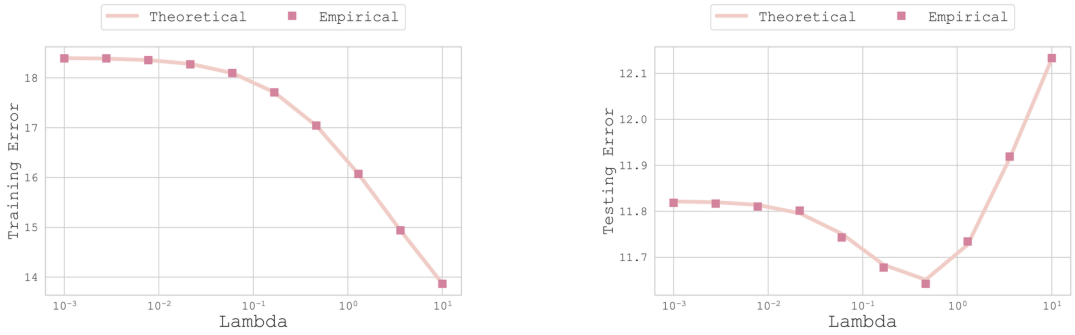


Figure 2: Empirical and theoretical train and test MSE as functions of the parameter λ for different values of α. The smooth curves represent the theoretical predictions, while the corresponding curves with the same color show the empirical results, highlighting that the empirical observations indeed match the theoretical predictions.

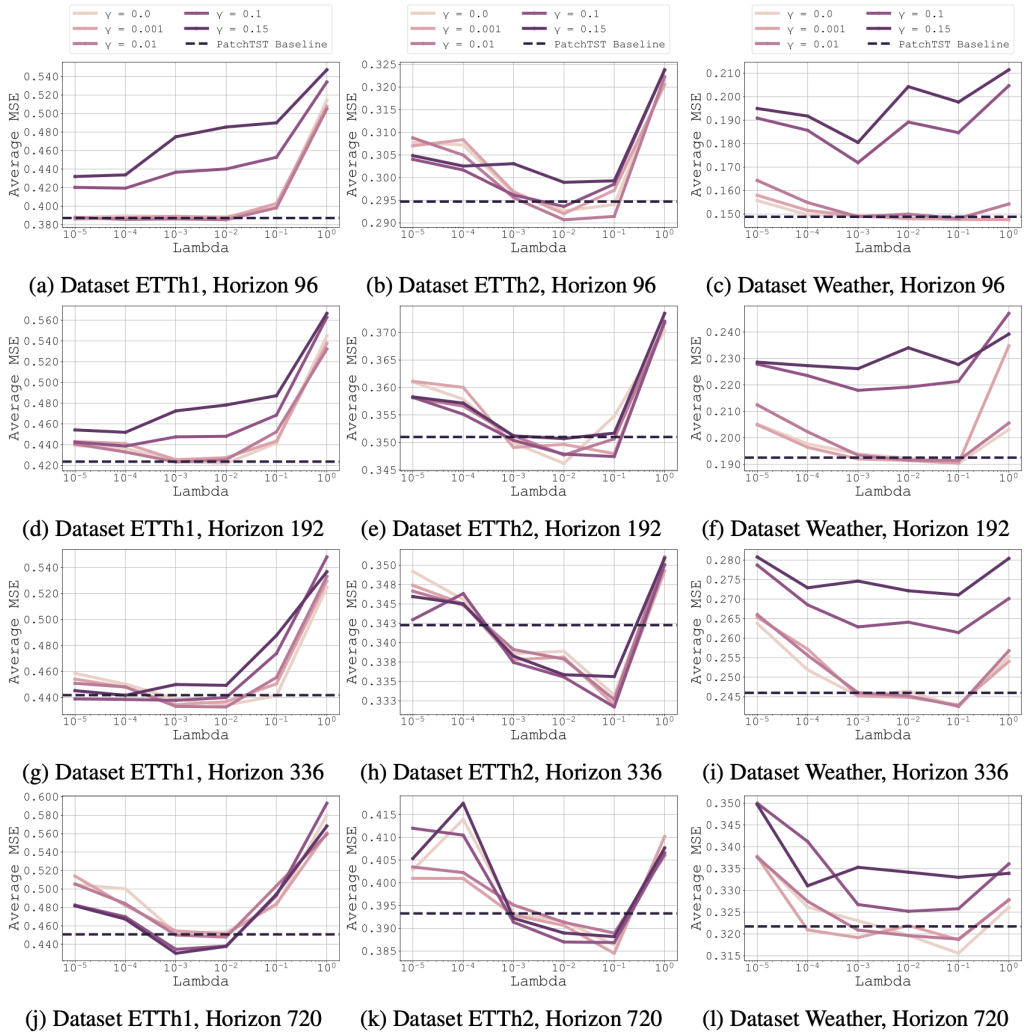


(a) Training MSE

(b) Testing MSE

Figure 3: Theoretical vs Empirical MSE as function of regularization parameter λ. Close fit between the theoretical and the empirical predictions which underscores the robustness of the theory in light of varying assumptions as well as the accuracy of the suggested estimates. We consider the first two channels as the the two tasks and $d = 144$. 95 samples are used for the training and 42 samples are used for the test.

Application to Multivariate Time Series Forecasting



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

(d) Dataset ETTh1, Horizon 192 (e) Dataset ETTh2, Horizon 192 (f) Dataset Weather, Horizon 192

(g) Dataset ETTh1, Horizon 336 (h) Dataset ETTh2, Horizon 336 (i) Dataset Weather, Horizon 336

(j) Dataset ETTh1, Horizon 720 (k) Dataset ETTh2, Horizon 720 (l) Dataset Weather, Horizon 720

Figure 4: Results of our optimization method on different datasets and horizons averaged across 3 different seeds for each gamma and lambda values for the `PatchTST` baseline

Application to Multivariate Time Series Forecasting

Dataset	H	with MTL regularization			without MTL regularization					
		PatchTST	DLinearU	Transformer	PatchTST	DLinearU	DLinearM	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 regularization approach allows for the efficient use of univariate models in a multivariate context.
- We show that our method improves performance over PatchTST and DLinear compared to independent application to each channel.
- It enables univariate models to reach SOTA performance similar to multivariate models like SAMformer and iTransformer.

Conclusions and Future Works

- Explored linear multi-task learning with a closed-form solution for an optimization problem leveraging information across multiple tasks.
- Applied Random Matrix Theory to derive asymptotic training and testing risks.
- Provided insights into high-dimensional multi-task learning regression.
- Successfully applied theoretical analysis to multi-task regression and multivariate forecasting on synthetic and real-world datasets.
- Laid a solid foundation for future research using random matrix theory with more complex models, including deep neural networks, within the multi-task learning framework.