Enhancing Multivariate Time Series Forecasting via Multi-Task Learning and Random Matrix Theory

Romain Ilbert[*1,2] **Malik Tiomoko**[1] **Cosme Louart**[3]
Vasilii Feofanov [1] **Themis Palpanas**[2] **Ievgen Redko**[1]
[1]Huawei Noah's Ark Lab, Paris, France [2]LIPADE, Paris Descartes University, Paris, France
[3] School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China

Abstract

We present a novel approach to multivariate time series forecasting by framing it as a multi-task learning problem. We propose an optimization strategy that enhances single-channel predictions by leveraging information across multiple channels. Our framework offers a closed-form solution for linear models and connects forecasting performance to key statistical properties using advanced analytical tools. Empirical results on both synthetic and real-world datasets demonstrate that integrating our method into training loss functions significantly improves univariate models by effectively utilizing multivariate data within a multi-task learning framework.

1 Introduction

Multivariate time series forecasting (MTSF) is crucial in fields where predictions depend on multiple interrelated variables. Capturing dependencies among channels with varying dynamics poses significant challenges. Traditional methods either treat channels independently—losing cross-channel information—or use complex multivariate models that are computationally intensive and prone to overfitting, especially with limited data. We reframe MTSF as a multi-task learning (MTL) problem by treating each time series channel as a related task. This approach leverages shared knowledge to enhance single-channel predictions through effective information transfer. By balancing shared and specific components, we improve forecasting accuracy of state-of-the-art univariate models.

Our Approach. We frame MTSF within a multi-task learning framework by treating each of the T time series channels as a separate task. For each task t, we have training data $\mathbf{X}^{(t)} \in \mathbb{R}^{d \times n_t}$ and responses $\mathbf{Y}^{(t)} \in \mathbb{R}^{q \times n_t}$. We model the response using a linear signal-plus-noise model: [24]:

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

where $\boldsymbol{\varepsilon}^{(t)}$ represents noise, and $\mathbf{W}_t$ combines shared and task-specific components:

$$\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t. \tag{2}$$

Here, $\mathbf{W}_0$ captures common information, while $\mathbf{V}_t$ captures deviations unique to each task.

Theoretical Analysis and Optimization. We aim to retrieve the common and specific hyperplanes, $\mathbf{W}_0$ and $\mathbf{V}_t$, through the following minimization problem:

$$\mathbf{W}_0^*, \{\mathbf{V}_t^*\}_{t=1}^T, \lambda^* = \text{argmin} \quad \frac{1}{2\lambda}\|\mathbf{W}_0\|_F^2 + \frac{1}{2}\sum_{t=1}^T \frac{\|\mathbf{V}_t\|_F^2}{\gamma_t} + \frac{1}{2}\sum_{t=1}^T \left\| \mathbf{Y}^{(t)} - \frac{{\mathbf{X}^{(t)}}^\top \mathbf{W}_t}{\sqrt{Td}} \right\|_F^2. \tag{3}$$

[*]Correspondence to: romain.ilbert@hotmail.fr.

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

where $\gamma = [\gamma_1, \ldots, \gamma_T]$ serves as a regularization term for each task, controlling the degree of overfitting (small γ_t) or underfitting (large γ_t). This formulation balances shared and task-specific components, optimizing the performance of the MTL framework for forecasting.

Contributions. Our main contributions are:

1. We develop an optimization framework for MTSF within the MTL paradigm, providing closed-form solutions for linear models.
2. Using random matrix theory, we derive exact expressions for training and test risks, offering insights into balancing shared and task-specific learning.
3. We present a data-dependent solution for optimal hyperparameters, informed by data covariances and noise levels, facilitating practical implementation.
4. We validate our approach on real datasets, showing significant improvements over univariate models and achieving performance comparable to state-of-the-art multivariate models.

Applications and Results. We validate our framework on multivariate time series forecasting, showing that our multi-task learning approach enhances univariate forecasting models like `PatchTST` [22] and DLinear [36] by leveraging shared learning across channels. Our method achieves performance comparable to state-of-the-art multivariate models such as `SAMformer` [12] and `iTransformer` [17], demonstrating the effectiveness of treating MTSF as a multi-task learning problem.

2 Related Work

Multi-Task Learning and Theoretical Bounds. Multi-task learning (MTL), introduced by Caruana [3], improves performance by sharing information across tasks. Recent research has focused on task diversity [11] and theoretical risk bounds, exploring task correlations via contrasts [16] and transfer distances [20], providing insights into generalization error and task similarity [21]. However, these studies are often theoretical and lack practical algorithms for optimizing hyperparameters—gaps we address in this work by offering practical performance evaluations and tuning insights.

Random Matrix Theory in MTL. Random Matrix Theory (RMT) provides precise estimates for metrics like train/test risk in high-dimensional settings [1, 30]. While RMT has been applied to MTL for classification [31], its use in forecasting remains underexplored. We build on these findings to tackle multivariate forecasting, focusing on hyperparameter selection within an optimized framework, drawing inspiration from studies on negative transfer [35].

Multivariate Time Series Forecasting. RMT has seen limited application in Multi-Task Sequential Forecasting (MTSF), a common task in fields like healthcare [4], energy [33], and finance [28]. Methods range from classical models [5, 29] and ARIMA [2] to deep learning approaches [12, 22, 25, 34, 38]. While univariate models [22] are often used for multivariate forecasting, we enhance such models by integrating regularization from RMT and MTL.

3 Framework

Notation. Matrices are denoted by uppercase bold letters (e.g., $\mathbf{A}$), vectors by lowercase bold letters (e.g., $\mathbf{v}$), and scalars by regular type (e.g., a). The Kronecker product is denoted by $\mathbf{A} \otimes \mathbf{B}$. The canonical vector of $\mathbb{R}^T$ is $\mathbf{e}_t^{[T]}$, and $\mathrm{Diag}(\mathbf{x})$ denotes a diagonal matrix with elements of $\mathbf{x}$. The Frobenius norm of a matrix $\mathbf{M}$ is $\|\mathbf{M}\|_F = \sqrt{\mathrm{tr}(\mathbf{M}^\top \mathbf{M})}$. The training data is denoted by:

$$\mathbf{Y} = [\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(T)}] \in \mathbb{R}^{q \times n}, \quad \mathbf{Z} = \sum_{t=1}^{T} (\mathbf{e}_t^{[T]} \mathbf{e}_t^{[T]\top}) \otimes \mathbf{X}^{(t)} \in \mathbb{R}^{Td \times n},$$

where $n = \sum_{t=1}^{T} n_t$ is the total number of samples across all tasks.

3.1 Multi-Task Model

We solve the multi-task forecasting problem by finding $\hat{\mathbf{W}} = [\hat{\mathbf{W}}_1^\top, \ldots, \hat{\mathbf{W}}_T^\top]^\top \in \mathbb{R}^{dT \times q}$ under the assumption $\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t$, where $\mathbf{W}_0$ is the shared component. The optimization problem is:

$$\min_{(\mathbf{W}_0, \mathbf{V}) \in \mathbb{R}^{d \times q} \times \mathbb{R}^{dT \times q}} \mathcal{J}(\mathbf{W}_0, \mathbf{V}),$$

where

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

This convex optimization problem has a unique solution as shown in Appendix B.

3.2 Assumptions

Concentrated Random Vectors. We assume feature vectors $\mathbf{x}_i^{(t)}$ are concentrated random vectors, meaning they satisfy concentration inequalities and have zero mean with bounded covariance $\mathbf{\Sigma}^{(t)}$. This assumption holds for Gaussian vectors, vectors uniformly distributed on spheres, and certain Lipschitz transformations, capturing a wide range of practical data settings [26].

High-Dimensional Asymptotics. We consider a high-dimensional asymptotic regime where $n_t = \mathcal{O}(d)$ and $T = \mathcal{O}(1)$, assuming $n/d \xrightarrow{\text{a.s.}} c_0 < \infty$. This setting is common in applications such as telecommunications, finance, and machine learning [7, 23].

More details about these assumptions and their implications can be found in Appendix A.

4 Main Theoretical Results

4.1 Performance Estimation

Given training data $\mathbf{X} \in \mathbb{R}^{n \times d}$ and response $\mathbf{Y} \in \mathbb{R}^{n \times q}$, we define the training and test risks as:

$$\mathcal{R}_{train}^{\infty} = \frac{1}{Tn} \mathbb{E}\left[\|\mathbf{Y} - g(\mathbf{X})\|_2^2 \right], \quad \mathcal{R}_{test}^{\infty} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{y}^{(t)} - g(\mathbf{x}^{(t)})\|_2^2].$$

where: $g(\mathbf{x}^{(t)}) = \frac{1}{Td} \left(\mathbf{e}_t^{[T]} \otimes \mathbf{x}^{(t)} \right)^\top \mathbf{AZQY}$, and $\mathbf{Q} = \left(\frac{\mathbf{Z}^\top \mathbf{AZ}}{Td} + \mathbf{I}_{Td} \right)^{-1}$.

To analyze $\mathcal{R}_{train}^{\infty}$ and $\mathcal{R}_{test}^{\infty}$, we use a deterministic equivalent of $\mathbf{Q}$, denoted $\bar{\mathbf{M}}$, which approximates $\mathbf{Q}$ in a linear form. This approach allows estimating key quantities like $\frac{1}{d}\operatorname{tr}(\mathbf{AQ})$ using the coresolvent $\tilde{\mathbf{Q}} = \left(\frac{\mathbf{A}^{\frac{1}{2}} \mathbf{Z} \mathbf{Z}^\top \mathbf{A}^{\frac{1}{2}}}{Td} + \mathbf{I}_{Td} \right)^{-1}$.

Using Lemma 1, we derive deterministic equivalents to characterize the asymptotic behavior of test risks, leading to the following theorem.

Theorem 1 (Asymptotic Test Risk)**.** *Under the assumptions of concentrated random vectors and high-dimensional asymptotics, the asymptotic test risk is given by:*

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

where $\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})$ *and* $\bar{\mathbf{Q}}_2$ *are deterministic equivalents for specific matrix forms of* $\tilde{\mathbf{Q}}$ *and* $\mathbf{Q}$.

The proof and detailed derivations are provided in Appendix D.

4.2 Error Contribution Analysis

We decompose the test risk into signal and noise components.

Signal Term. Approximated as $\operatorname{tr}(\mathbf{W}^\top (\mathbf{A\Sigma} + \mathbf{I})^{-2} \mathbf{W})$, where $\mathbf{\Sigma} = \sum_{t=1}^{T} \frac{n_t}{d} \mathbf{\Sigma}^{(t)}$. The matrix $(\mathbf{A\Sigma} + \mathbf{I})^{-2}$ amplifies both independent and cross terms, enhancing the multi-task effect and decreasing the test risk. The cross terms, critical for multi-task learning, are maximized when task

Table 1: MTL regularization results. Models marked with † are state-of-the-art multivariate benchmarks. We compare MTL-regularized models to their unregularized versions. MSEs are averaged over 3 seeds, * indicates significant improvement (p < 0.05) and **bold** values show the top performers.

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

covariances are aligned (i.e., $\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_v = \mathbf{I}_d$), boosting performance. Larger Fisher distances between tasks reduce these beneficial correlations.

Noise Term. Approximated as $\operatorname{tr}(\mathbf{\Sigma}_N(\mathbf{A}^{-1} + \mathbf{\Sigma})^{-1})$, the noise term lacks cross-task interactions due to independent noise, affecting only diagonal elements and contributing to negative transfer. It increases with sample size and λ, hindering task transfer—a critical factor in multi-task learning.

5 Application to Multivariate Time Series Forecasting

We apply our theoretical framework to Multivariate Time Series Forecasting (MTSF), extending previous work from a linear setting to neural networks. The presented results assume optimal lambda values, as discussed in Appendix I.

Motivation. Many current MTSF models are univariate, neglecting the multivariate information available in benchmarks. Our framework is well-suited for MTSF models that utilize a linear layer for historical data projection, improving forecasting accuracy. Additionally, under the assumption that input vectors are concentrated, we aim to explore whether our theoretical results, developed for linear models, could extend to neural networks. Since neural networks are Lipschitz continuous functions, they preserve the concentration property of inputs, making it plausible that our theory could apply in this context as well.

Our approach. We modify the loss function to include feature-specific transformations f_t and a shared transformation f_0. For a neural network f with inputs $\mathbf{X}$, we compare univariate models with and without MTL regularization: $f_t^{MTL}(\mathbf{X}^{(t)}) = f_t(\mathbf{X}^{(t)}) + f_0(\mathbf{Y})$. The regularized loss is:

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \sum_{t=1}^{T} \|\mathbf{Y}^{(t)} - f_t^{MTL}(\mathbf{X}^{(t)})\|_F^2 + \lambda \|f_0(\mathbf{X})\|_F^2 + \sum_{t=1}^{T} \gamma_t \|f_t(\mathbf{X}^{(t)})\|_F^2.$$

This balances fitting multivariate series with shared dynamics and specific channels.

Results. Experiments on various datasets show MTL regularization improves `PatchTST`, `DLinearU`, and `Transformer` models compared to their unregularized versions and state-of-the-art multivariate models like `SAMformer` and `iTransformer`. The best performances are achieved by `PatchTST` and `DLinearU` with MTL regularization, often outperforming state-of-the-art models. Detailed results are in Table 1 and Appendix H.3.

6 Conclusions and Future Works

In this paper, we proposed the application of multi-task learning to multivariate time series forecasting, providing a theoretical foundation and practical insights. By deriving a closed-form solution for linear multi-task optimization and employing Random Matrix Theory, we quantified training and testing risks, revealing key insights into high-dimensional multi-task regression. Our approach, validated on traditional real-world forecasting benchmarks, demonstrates the potential of multi-task learning trough a regularization term to enhance forecasting accuracy of the state-of-the-art univariate models.

References

[1] Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York, NY, 2010. ISBN 978-1-4419-0660-1. doi: 10.1007/978-1-4419-0661-8.

[2] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., USA, 1990. ISBN 0816211043.

[3] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[4] Paulius Čepulionis and Kristina Lukoševičiūtė. Electrocardiogram time series forecasting and optimization using ant colony optimization algorithm. *Mathematical Models in Engineering*, 2 (1):69–77, Jun 2016. ISSN 2351-5279. URL https://www.extrica.com/article/17229.

[5] Renyi Chen and Molei Tao. Data-driven prediction of general hamiltonian dynamics via learning exactly-symplectic maps. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1717–1727. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/chen21r.html.

[6] Si-An Chen, Chun-Liang Li, Sercan O Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

[7] Romain Couillet and Merouane Debbah. *Random matrix methods for wireless communications*. Cambridge University Press, 2011.

[8] Romain Couillet, Yagmur Gizem Cinar, Eric Gaussier, and Muhammad Imran. Word representations concentrate and this is good news. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 208–219, 2020.

[9] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[10] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *IEEE Transactions on Information Theory*, 69(3):1824–1852, 2022.

[11] Romain Ilbert, Thai V. Hoang, and Zonghua Zhang. Data augmentation for multivariate time series classification: An experimental study. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, pages 128–139, 2024. doi: 10.1109/ICDEW61823.2024.00023.

[12] Romain Ilbert, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux, Giuseppe Paolo, Themis Palpanas, and Ievgen Redko. Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention, 2024.

[13] Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, Vasilii Feofanov, Themis Palpanas, and Ievgen Redko. Analysing multi-task regression via random matrix theory with application to time series forecasting, 2024. URL https://arxiv.org/abs/2406.10327.

[14] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=cGDAkQo1C0p.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[16] Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

[17] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

[18] Cosme Louart and Romain Couillet. Spectral properties of sample covariance matrices arising from random matrices with independent non identically distributed columns. *arXiv preprint arXiv:2109.02644*, 2021.

[19] Max Planck Institute. Weather dataset, n.d. URL https://pems.dot.ca.gov/.

[20] Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.

[21] Minh-Toan Nguyen and Romain Couillet. Asymptotic bayes risk of semi-supervised multitask learning on gaussian mixture. In *International Conference on Artificial Intelligence and Statistics*, pages 5063–5078. PMLR, 2023.

[22] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

[23] M. Potters, J.-P. Bouchaud, and L. Laloux. Financial applications of random matrix theory: Old laces and new pieces. *Acta Physica Polonica B*, 36(9):2767, 2005.

[24] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pages 1444–1452. PMLR, 2013.

[25] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL https://www.sciencedirect.com/science/article/pii/S0169207019301888.

[26] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning*, 2020.

[27] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. *arXiv preprint arXiv:2006.05667*, 2020.

[28] Gaurang Sonkavde, Deepak Sudhakar Dharrao, Anupkumar M. Bongale, Sarika T. Deokate, Deepak Doreswamy, and Subraya Krishna Bhat. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*, 11(3), 2023. ISSN 2227-7072. doi: 10.3390/ijfs11030094. URL https://www.mdpi.com/2227-7072/11/3/94.

[29] Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, and Amaury Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16):2861–2869, 2007. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2006.06.015. URL https://www.sciencedirect.com/science/article/pii/S0925231207001610. Neural Network Applications in Electrical Engineering Selected papers from the 3rd International Work-Conference on Artificial Neural Networks (IWANN 2005).

[30] Terence Tao. *Topics in Random Matrix Theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012. ISBN 978-0821885079.

[31] Malik Tiomoko, Romain Couillet, and Hafiz Tiomoko. Large dimensional analysis and improvement of multi task learning, 2020.

[32] Malik Tiomoko, Ekkehard Schnoor, Mohamed El Amine Seddik, Igor Colin, and Aladin Virmaux. Deciphering lasso-based classification through a large dimensional analysis of the iterative soft-thresholding algorithm. In *International Conference on Machine Learning*, pages 21750–21778, 2022.

[33] UCI. Electricity dataset, n.d. URL https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014.

[34] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

[35] Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Ré, and Weijie J. Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers, 2023.

[36] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[37] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pages 11106–11115. AAAI Press, 2021.

[38] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*, 2022.

Appendix

Table of Contents

A Discussion about the Main Assumptions **9**

A.1 On the zero-mean assumption 9

A.2 Concentration of Random Vectors 9

A.3 High-Dimensional Asymptotics 9

B Minimization Problem **9**

B.1 Computation of $\hat{\mathbf{W}}_t$ and $\hat{\mathbf{W}}_0$ 9

C Lemma 1 and proof with Random Matrix Theory **11**

C.1 Lemma 1 11

C.2 Deterministic equivalent of the resolvent $\tilde{\mathbf{Q}}$ 11

C.3 Deterministic equivalent of bilinear forms of the resolvent 12

C.4 Estimation of the deterministic equivalent of $\mathbf{Q}^2$ 14

D Risk Estimation (Proof of Theorem 1) **15**

D.1 Test Risk 15

D.2 Train Risk 16

E Interpretation and insights of the theoretical analysis **16**

E.1 Analysis of the test risk 16

E.2 Interpretation of the signal term 17

E.3 Interpretation and insights of the noise terms 17

E.4 Simplified Model Insights 17

E.5 Optimal Lambda 18

F Theoretical Estimations **18**

F.1 Estimation of the training and test risk 18

F.2 Estimation of the noise covariance 19

G Multi-Task Regression **19**

G.1 Related Work 19

G.2 Empirical vs. Theoretical Comparison 19

H Multivariate Time Series Forecasting **19**

H.1 Architecture and Training Parameters 19

H.2 Datasets 20

H.3 Additional Experiments. 21

I Limitations **24**

A Discussion about the Main Assumptions

A.1 On the zero-mean assumption

We would like to note that we are assuming that both the noise ε and the feature $\mathbf{x}_i^{(t)}$ have zero mean. This is a common assumption in many statistical models and it simplifies the analysis. However, this assumption is not restrictive. In practice, if the data or the response variable are not centered, we can always preprocess the data by subtracting the mean. This preprocessing step brings us back to the zero-mean setting that we consider in our theoretical analysis.

A.2 Concentration of Random Vectors

Our theoretical analysis is based on the assumption that the data are *concentrated random vectors*. This means that in high-dimensional spaces, the data exhibit stable statistical properties, even after complex transformations. This assumption is more realistic and encompassing than traditional Gaussian assumptions, as it applies to a wider class of distributions and data types encountered in practice.

The justification for this assumption stems from the behavior of neural networks in tasks like image recognition and natural language processing. Neural networks act as Lipschitz continuous functions, preserving distances between inputs and outputs to some extent, which leads to concentrated representations. Recent studies have demonstrated that both real-world datasets and synthetic data generated by Generative Adversarial Networks (GANs) exhibit concentration phenomena [8, 27, 32]..

By considering concentrated random vectors, we can more accurately model the behavior of algorithms on complex, high-dimensional data without relying on restrictive distributional assumptions. This approach captures the stability and structure inherent in modern datasets, providing a solid foundation for our theoretical results.

A.3 High-Dimensional Asymptotics

We also assume a *high-dimensional asymptotic* regime where both the feature dimension d and the sample size n grow large together, with their ratio d/n approaching a finite constant. This assumption reflects the nature of modern datasets, where collecting more data often involves increasing both the number of samples and the number of features, as seen in genomics and image analysis.

This high-dimensional perspective is more appropriate than classical settings where n grows while d remains fixed, which may not capture the complexities and interactions present in large-scale data. By allowing both d and n to increase, we account for the intricate relationships and potential overfitting issues that can arise in high dimensions.

Under this regime, the variance of our estimators scales as $O\left(\frac{1}{\sqrt{dn}}\right)$, implying that as both d and n become large, the variance diminishes, and empirical results closely match theoretical predictions. This scaling offers better reliability compared to scenarios where only n increases, reinforcing the practical relevance of our theoretical findings.

B Minimization Problem

B.1 Computation of $\hat{\mathbf{W}}_t$ and $\hat{\mathbf{W}}_0$

The proposed multi task regression finds $\hat{\mathbf{W}} = [\hat{\mathbf{W}}_1^\top, \dots, \hat{\mathbf{W}}_k^\top]^\top \in \mathbb{R}^{dT\times q}$ which solves the following optimization problem using the additional assumption of relatedness between the tasks ($\mathbf{W}_t = \mathbf{W}_0 + \mathbf{V}_t$ for all tasks t):

$$\min_{(\mathbf{W}_0,\mathbf{V})\in\mathbb{R}^d\times\mathbb{R}^{d\times T}\times\mathbb{R}^T} \mathcal{J}(\mathbf{W}_0, \mathbf{V}) \tag{4}$$

where

$$\mathcal{J}(\mathbf{W}_0,\mathbf{V}) \equiv \frac{1}{2\lambda}\mathrm{tr}\left(\mathbf{W}_0^\top\mathbf{W}_0\right)+\frac{1}{2}\sum_{t=1}^{T}\frac{\mathrm{tr}\left(\mathbf{V}_t^\top\mathbf{V}_t\right)}{\gamma_t}+\frac{1}{2}\sum_{t=1}^{T}\mathrm{tr}\left(\boldsymbol{\xi}_t^\top\boldsymbol{\xi}_t\right)$$

$$\boldsymbol{\xi}_t=\mathbf{Y}^{(t)}-\frac{{\mathbf{X}^{(t)}}^\top\mathbf{W}_t}{\sqrt{Td}},\quad \forall t\in\{1,\ldots,T\}.$$

The Lagrangian introducing the lagrangian parameters for each task t, $\boldsymbol{\alpha}_t\in\mathbb{R}^{n_t\times q}$ reads as

$$\begin{aligned}\mathcal{L}(\mathbf{W}_0,\mathbf{V}_t,\boldsymbol{\xi}_t,\boldsymbol{\alpha}_t)&=\frac{1}{2\lambda}\mathrm{tr}\left(\mathbf{W}_0^\top\mathbf{W}_0\right)+\frac{1}{2}\sum_{t=1}^{T}\frac{\mathrm{tr}\left(\mathbf{V}_t^\top\mathbf{V}_t\right)}{\gamma_t}+\frac{1}{2}\sum_{t=1}\mathrm{tr}\left(\boldsymbol{\xi}_t^\top\boldsymbol{\xi}_t\right)\\&+\sum_{t=1}^{T}\mathrm{tr}\left(\boldsymbol{\alpha}_t^\top\left(\mathbf{Y}^{(t)}-\frac{{\mathbf{X}^{(t)}}^\top(\mathbf{W}_0+\mathbf{V}_t)}{\sqrt{Td}}-\boldsymbol{\xi}_t\right)\right)\end{aligned}$$

Differentiating with respect to the unknown variables $\hat{\mathbf{W}}_0$, $\hat{\mathbf{V}}_t$, $\boldsymbol{\xi}_t$, $\boldsymbol{\alpha}_t$ and $\mathbf{b}_t$, we get the following system of equation

$$\begin{aligned}\frac{1}{\lambda}\hat{\mathbf{W}}_0-\sum_{t=1}^{T}\frac{\mathbf{X}^{(t)}\boldsymbol{\alpha}_t}{\sqrt{Td}}&=0\\\frac{1}{\gamma_t}\hat{\mathbf{V}}_t-\frac{\mathbf{X}^{(t)}\boldsymbol{\alpha}_t}{\sqrt{Td}}&=0\\\boldsymbol{\xi}_t-\boldsymbol{\alpha}_t&=0\\\mathbf{Y}^{(t)}-\frac{{\mathbf{X}^{(t)}}^\top\hat{\mathbf{W}}_0}{\sqrt{Td}}-\frac{{\mathbf{X}^{(t)}}^\top\hat{\mathbf{V}}_t}{\sqrt{Td}}-\boldsymbol{\xi}_t&=0\end{aligned}$$

Plugging the expression of $\hat{\mathbf{W}}_0$, $\hat{\mathbf{V}}_t$ and $\boldsymbol{\xi}_t$ into the expression of $\mathbf{Y}^{(t)}$ gives

$$\mathbf{Y}^{(t)}=\lambda\sum_{t=1}^{T}\frac{{\mathbf{X}^{(t)}}^\top\mathbf{X}^{(t)}}{Td}\boldsymbol{\alpha}_t+\gamma_t\frac{{\mathbf{X}^{(t)}}^\top\mathbf{X}^{(t)}}{Td}\boldsymbol{\alpha}_t+\boldsymbol{\alpha}_t$$

which can be rewritten as

$$\mathbf{Y}^{(t)}=(\lambda+\gamma_t)\frac{{\mathbf{X}^{(t)}}^\top\mathbf{X}^{(t)}}{Td}\boldsymbol{\alpha}_t+\lambda\sum_{v\neq t}\frac{{\mathbf{X}^{(t)}}^\top\mathbf{X}^{(v)}}{Td}\boldsymbol{\alpha}_v+\boldsymbol{\alpha}_t$$

With $\mathbf{Y}=[{\mathbf{Y}^{(1)}}^\top,\ldots,{\mathbf{Y}^{(T)}}^\top]^\top\in\mathbb{R}^{n\times q}$, $\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1^\top,\ldots,\boldsymbol{\alpha}_k^\top]^\top\in\mathbb{R}^{n\times q}$, $\mathbf{Z}=\sum_{t=1}^{T}\mathbf{e}_t^{[T]}{\mathbf{e}_t^{[T]}}^\top\otimes\mathbf{X}^{(t)}\in\mathbb{R}^{Td\times n}$, this system of equations can be written under the following compact matrix form:

$$\mathbf{Q}^{-1}\boldsymbol{\alpha}=\mathbf{Y}$$

with $\mathbf{Q}=\left(\frac{\mathbf{Z}^\top\mathbf{A}\mathbf{Z}}{Td}+\mathbf{I}_n\right)^{-1}\in\mathbb{R}^{n\times n}$, and $\mathbf{A}=\left(\mathcal{D}_{\boldsymbol{\gamma}}+\lambda\mathbb{1}_T\mathbb{1}_T^\top\right)\otimes\mathbf{I}_d\in\mathbb{R}^{Td\times Td}$.

Solving for $\boldsymbol{\alpha}$ then gives:

$$\boldsymbol{\alpha}=\mathbf{Q}\mathbf{Y}$$

Moreover, using $\hat{\mathbf{W}}_t=\hat{\mathbf{W}}_0+\hat{\mathbf{V}}_t$, the expression of $\mathbf{W}_t$ becomes:

$$\begin{aligned}\hat{\mathbf{W}}_t&=\left({\mathbf{e}_t^{[T]}}^\top\otimes\mathbf{I}_d\right)\frac{\mathbf{A}\mathbf{Z}\boldsymbol{\alpha}}{\sqrt{Td}},\\\hat{\mathbf{W}}_0&=\left(\mathbb{1}_T^\top\otimes\lambda\mathbf{I}_d\right)\frac{\mathbf{Z}\boldsymbol{\alpha}}{\sqrt{Td}}.\end{aligned}$$

C Lemma 1 and proof with Random Matrix Theory

C.1 Lemma 1

Lemma 1 (Deterministic equivalents for $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{Q}}\mathbf{M}\tilde{\mathbf{Q}}$ and $\mathbf{Q}^2$ for any $\mathbf{M} \in \mathbb{R}^{n\times n}$)**.** *Under the concentrated random vector assumption for each feature vector $\mathbf{x}_i^{(t)}$ and under the growth rate assumption (Assumption* **??***), for any deterministic $\mathbf{M} \in \mathbb{R}^{n\times n}$, we have the following convergence:*

$$\tilde{\mathbf{Q}} \leftrightarrow \bar{\tilde{\mathbf{Q}}}, \qquad \tilde{\mathbf{Q}}\mathbf{M}\tilde{\mathbf{Q}} \leftrightarrow \bar{\tilde{\mathbf{Q}}}_2(\mathbf{M}), \qquad \mathbf{Q}^2 \leftrightarrow \bar{\mathbf{Q}}_2$$

where $\bar{\tilde{\mathbf{Q}}}_2$, $\bar{\tilde{\mathbf{Q}}}$ and $\bar{\mathbf{Q}}_2$ are defined as follows

$$\bar{\tilde{\mathbf{Q}}} = \left(\sum_{t=1}^{T} \frac{c_0\mathbf{C}^{(t)}}{1+\delta_t} + \mathbf{I}_{Td}\right)^{-1}, \quad \delta_t = \frac{1}{Td}\mathrm{tr}\left(\mathbf{\Sigma}^{(t)}\bar{\tilde{\mathbf{Q}}}\right), \quad \mathbf{C}^{(t)} = \mathbf{A}^{\frac{1}{2}}\left(\mathbf{e}_t^{[T]} \otimes \mathbf{\Sigma}^{(t)}\right)\mathbf{A}^{\frac{1}{2}}$$

$$\bar{\tilde{\mathbf{Q}}}_2(\mathbf{M}) = \bar{\tilde{\mathbf{Q}}}\mathbf{M}\bar{\tilde{\mathbf{Q}}} + \frac{1}{Td}\sum_{t=1}^{T}\frac{d_t}{1+\delta_t}\bar{\tilde{\mathbf{Q}}}\mathbf{C}^{(t)}\bar{\tilde{\mathbf{Q}}}, \qquad \mathbf{d} = \left(\mathbf{I}_T - \frac{1}{Td}\Psi\right)^{-1}\Psi(\mathbf{M}) \in \mathbb{R}^T$$

$$\bar{\mathbf{Q}}_2 = \mathbf{I}_n - \mathit{Diag}_{t\in[T]}(v_t\mathbf{I}_{n_t}), \quad v_t = \frac{1}{Td}\frac{\mathrm{tr}(\mathbf{C}^{(t)}\bar{\tilde{\mathbf{Q}}})}{(1+\delta_t)^2} + \frac{1}{Td}\frac{\mathrm{tr}\left(\mathbf{C}^{(t)}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_n)\right)}{(1+\delta_t)^2}$$

where

$$\Psi(M) = \left(\frac{n_t}{Td}\frac{\mathrm{tr}\left(\mathbf{C}^{(t)}\bar{\tilde{\mathbf{Q}}}\mathbf{M}\bar{\tilde{\mathbf{Q}}}\right)}{1+\delta_t}\right)_{t\in[T]} \in \mathbb{R}^T, \qquad \Psi = \left(\frac{n_t}{Td}\frac{\mathrm{tr}\left(\mathbf{C}^{(t)}\bar{\tilde{\mathbf{Q}}}\mathbf{C}^{(t')}\bar{\tilde{\mathbf{Q}}}\right)}{(1+\delta_t)(1+\delta_{t'})}\right)_{t,t'\in[T]} \in \mathbb{R}^{T\times T},$$

C.2 Deterministic equivalent of the resolvent $\tilde{\mathbf{Q}}$

The evaluation of the expectation of linear forms on $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{Q}}^2$ can be found in the literature. To find a result that meets exactly our setting, we will cite [18] that is a bit more general since it treats cases where $\mathbb{E}[x_i^{(t)}] \neq 0$ for $t \in [T]$ and $i \in [n_t]$. Unlike the main paper, and to be more general, the study presented below is "quasi asymptotic" meaning that the results are true for finite value of d, n. Let us first rewrite the general required hypotheses, adapting them to our setting. For that purpose, we consider in the rest of this paper a certain asymptotic $I \subset \{(d,n), d \in \mathbb{N}, n \in \mathbb{N}\} = \mathbb{N}^2$ satisfying:

$$\{d, \exists\, n \in \mathbb{N} :\ (d,n) \in I\} = \mathbb{N} \qquad \text{and} \qquad \{n, \exists\, d \in \mathbb{N} :\ (d,n) \in I\} = \mathbb{N}.$$

such that n and d can tend to ∞ but with some constraint that is given in the first item of Assumption 1 below. Given two sequences $(a_{d,n})_{d,n\in I}, (b_{d,n})_{d,n\in I} > 0$, the notation $a_{d,n} \leq O(b_{d,n})$ (or $a \leq O(b)$) means that there exists a constant $C > 0$ such that for all $(d,n) \in I$, $a_{d,n} \leq Cb_{d,n}$.

Assumption 1. *There exists some constants $C, c > 0$ independent such that:*

- $n \leq O(d)$
- $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \mathbb{R}^{Td\times n}$ *has independent columns*
- *for any $(d,n) \in I$, and any $f : \mathbb{R}^{Td\times n} \to \mathbb{R}$ 1-Lipschitz for the euclidean norm:*

$$\mathbb{P}\left(|f(\mathbf{Z}) - \mathbb{E}[f(\mathbf{Z})]| \geq t\right) \leq Ce^{-ct^2}.$$

- $\forall i \in \{n, \exists d \in \mathbb{N}, (d,n) \in I\}$*:* $\|\mathbb{E}[\mathbf{z}_i]\| \leq O(1)$.

Theorem 2 ([18], Theorem 0.9.)**.** *Given $T \in \mathbb{N}$, $\mathbf{Z} \in \mathbb{R}^{Td\times n}$ and two deterministic $A \in \mathbb{R}^{Td\times Td}$, we note $\tilde{\mathbf{Q}} \equiv (\frac{1}{Td}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^\top\mathbf{A}^{\frac{1}{2}} + I_{Td})^{-1}$. If $\mathbf{Z}$ satisfies Assumption 1 and $\mathbf{M} \in \mathbb{R}^{Td\times Td}$ is a deterministic matrix satisfying $\|\mathbf{M}\|_F \leq 1$, one has the concentration:*

$$\mathbb{P}\left(\left|\mathrm{tr}(M\tilde{\mathbf{Q}}) - \mathrm{tr}(M\bar{\tilde{\mathbf{Q}}}_{\boldsymbol{\delta}(\mathbf{S})}(\mathbf{S}))\right| \geq t\right) \leq Ce^{-ct^2},$$

where $\mathbf{S} = (\mathbf{S}_1, \dots, \mathbf{S}_n) = (\mathbb{E}[\mathbf{z}_1\mathbf{z}_1^\top], \dots, \mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top])$*, for* $\boldsymbol{\delta} \in \mathbb{R}^n$*,* $\bar{\tilde{\mathbf{Q}}}_{\boldsymbol{\delta}}$ *is defined as:*

$$\bar{\tilde{\mathbf{Q}}}_{\boldsymbol{\delta}}(\mathbf{S}) = \left(\frac{1}{Td} \sum_{i \in [n]} \frac{\mathbf{A}^{\frac{1}{2}}\mathbf{S}_i\mathbf{A}^{\frac{1}{2}}}{1 + \boldsymbol{\delta}_i} + I_{Td} \right)^{-1},$$

and $\boldsymbol{\delta}(\mathbf{S})$ *is the unique solution to the system of equations:*

$$\forall i \in [n] : \quad \boldsymbol{\delta}(\mathbf{S})_i = \frac{1}{n} \mathrm{tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{S}_i\mathbf{A}^{\frac{1}{2}} \bar{\tilde{\mathbf{Q}}}_{\boldsymbol{\delta}(\mathbf{S})} \right).$$

We end this subsection with some results that will be useful for next subsection on the estimation of bilinear forms on $\tilde{\mathbf{Q}}$.

Lemma 2 ([18], Lemmas 4.2, 4.6)**.** *Under the setting of Theorem 2, given a deterministic vector* $\mathbf{u} \in \mathbb{R}^{Td}$ *such that* $\|\mathbf{u}\| \leq O(1)$ *and two deterministic matrices* $\mathbf{U}, \mathbf{V}$ *such that* $\|\mathbf{U}\|, \|\mathbf{V}\| \leq O(1)$ *and a power* $r > 0$*,* $r \leq O(1)$*:*

- $\mathbb{E}\left[\left| \mathbf{u}^\top \mathbf{U} \tilde{\mathbf{Q}}_{-i} \mathbf{V} \mathbf{z}_i \right|^r \right] \leq O(1)$
- $\mathbb{E}\left[\left| \frac{1}{Td} \mathbf{z}_i^\top \mathbf{U} \tilde{\mathbf{Q}}_{-i} \mathbf{V} \mathbf{z}_i - \mathbb{E}\left[\frac{1}{Td} \mathrm{tr}\left(\Sigma_i \mathbf{U} \bar{\tilde{\mathbf{Q}}} \mathbf{B} \right) \right] \right|^r \right] \leq O\left(\frac{1}{d^{\frac{r}{2}}} \right).$

C.3 Deterministic equivalent of bilinear forms of the resolvent

To simplify the expression of the following theorem, we take $\mathbf{A} = \mathbf{I}_{Td}$. One can replace $\mathbf{Z}$ with $\mathbf{A}^{\frac{1}{2}}\mathbf{Z}$ to retrieve the result necessary for the main paper.

Theorem 3. *Under the setting of Theorem 2, with* $\mathbf{A} = \mathbf{I}_{Td}$*, one can estimate for any deterministic matrices* $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{Td}$ *such that* $\|\mathbf{U}\|, \|\mathbf{V}\| \leq O(1)$ *and any deterministic vector* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{Td}$ *such that* $\|\mathbf{u}\|, \|\mathbf{v}\| \leq 1$*, if one notes* $\mathbf{B} = \frac{1}{Td}\mathbf{V}$ *or* $\mathbf{B} = \mathbf{u}\mathbf{v}^\top$*, one can estimate:*

$$\left| \mathbb{E}\left[\mathrm{tr}(\mathbf{B}\tilde{\mathbf{Q}}\mathbf{U}\tilde{\mathbf{Q}}) \right] - \Psi(\mathbf{U}, \mathbf{B}) - \frac{1}{Td}\Psi(\mathbf{U})^\top \left(\mathbf{I}_n - \frac{1}{Td}\Psi \right)^{-1} \Psi(\mathbf{B}) \right| \leq O\left(\frac{1}{\sqrt{d}} \right) \tag{5}$$

where we noted:

- $\bar{\tilde{\mathbf{Q}}} \equiv \bar{\tilde{\mathbf{Q}}}_{\boldsymbol{\delta}}(\mathbf{S})$, $\boldsymbol{\delta} = \boldsymbol{\delta}(\mathbf{S})$,
- $\Psi \equiv \frac{1}{Td} \left(\frac{\mathrm{tr}\left(\mathbf{S}_i \bar{\tilde{\mathbf{Q}}} \mathbf{S}_j \bar{\tilde{\mathbf{Q}}} \right)}{(1+\delta_i)(1+\delta_j)} \right)_{i,j \in [n]} \in \mathbb{R}^{n,n}$
- $\forall \mathbf{U} \in \mathbb{R}^{n \times n}$ *:* $\Psi(\mathbf{U}) \equiv \frac{1}{Td} \left(\frac{\mathrm{tr}\left(\mathbf{U} \bar{\tilde{\mathbf{Q}}} \mathbf{S}_i \bar{\tilde{\mathbf{Q}}} \right)}{1+\delta_i} \right)_{i \in [n]} \in \mathbb{R}^n$
- $\forall \mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times n}$ *:* $\Psi(\mathbf{U}, \mathbf{V}) \equiv \frac{1}{Td} \mathrm{tr}\left(\mathbf{U} \bar{\tilde{\mathbf{Q}}} \mathbf{V} \bar{\tilde{\mathbf{Q}}} \right) \in \mathbb{R}$

If there exist $T < n$ dinstinct matrices $\mathbf{C}_1, \dots, \mathbf{C}_T$ such that:

$$\{\mathbf{S}_1, \dots, \mathbf{S}_n\} = \{\mathbf{C}_1, \dots, \mathbf{C}_T\},$$

and if we denote $\forall t \in [T]\ n_t = \#\{i \in [n] \mid \mathbf{S}_i = \mathbf{C}_t\}$ and:

$$P \equiv \left(I_T - \left(\frac{n_t n_v}{(Td)^2} \frac{\mathrm{tr}\left(\mathbf{S}_t \bar{\tilde{\mathbf{Q}}} \mathbf{S}_v \bar{\tilde{\mathbf{Q}}} \right)}{(1+\delta_t)(1+\delta_v)} \right)_{t,v \in [T]} \right)^{-1} \in \mathbb{R}^{T,T}$$

$$\forall \mathbf{U} \in \mathbb{R}^{Td \times Td} : \quad \bar{\tilde{\mathbf{Q}}}_2(\mathbf{U}) \equiv \bar{\tilde{\mathbf{Q}}} \mathbf{U} \bar{\tilde{\mathbf{Q}}} + \frac{1}{(Td)^2} \sum_{t,v=1}^{T} \frac{\mathrm{tr}(\mathbf{S}_t \bar{\tilde{\mathbf{Q}}} \mathbf{U} \bar{\tilde{\mathbf{Q}}}) P_{t,v} \bar{\tilde{\mathbf{Q}}} \mathbf{S}_v \bar{\tilde{\mathbf{Q}}}}{(1+\delta_t)(1+\delta_v)},$$

the result of Theorem 3 rewrites:

$$\left\| \mathbb{E}\left[\tilde{\mathbf{Q}} \mathbf{U} \tilde{\mathbf{Q}} \right] - \bar{\tilde{\mathbf{Q}}}_2(\mathbf{U}) \right\| \leq O\left(\frac{1}{\sqrt{d}} \right) \tag{6}$$

Proof. Given $i \in [n]$, let us note $\mathbf{Z}_{-i} = (\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, 0, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_n)$ and $\tilde{\mathbf{Q}}_{-i} = (\frac{1}{Td}\mathbf{Z}_{-i}\mathbf{Z}_{-i}^\top + \mathbf{I}_{Td})^{-1}$, then we have the identity:

$$\tilde{\mathbf{Q}} - \tilde{\mathbf{Q}}_{-i} = \frac{1}{Td}\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i} \qquad \text{and} \qquad \tilde{\mathbf{Q}}\mathbf{z}_i = \frac{\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i}{1 + \frac{1}{Td}\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i}. \tag{7}$$

Given $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{Td}$, such that $\|\mathbf{u}\|, \|\mathbf{v}\| \leq 1$, let us express:

$$\mathbb{E}\left[\frac{1}{Td}\mathbf{u}^\top\left(\tilde{\mathbf{Q}} - \bar{\tilde{\mathbf{Q}}}\right)\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\left(\frac{\mathbf{S}_i}{1+\boldsymbol{\delta}_i} - \mathbf{z}_i\mathbf{z}_i^\top\right)\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] \tag{8}$$

$$\tag{9}$$

First, given $i \in [n]$, let us estimate thanks to (7):

$$\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] = \mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] - \frac{1}{Td}\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right]$$

Hölder inequality combined with Lemma 2 allows us to bound:

$$\frac{1}{Td}\left|\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right]\right| \leq \frac{1}{Td}\mathbb{E}\left[\left|\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{z}_i\right|^2\right]^{\frac{1}{2}}\mathbb{E}\left[\left|\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right|^2\right]^{\frac{1}{2}} \leq O\left(\frac{1}{d}\right),$$

one can thus deduce:

$$\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] = \mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] + O\left(\frac{1}{d}\right) = \mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}_{-i}\mathbf{v}\right] + O\left(\frac{1}{d}\right). \tag{10}$$

Second, one can also estimate thanks to Lemma 7:

$$\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] = \mathbb{E}\left[\frac{\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}}{1 + \frac{1}{Td}\mathbf{z}_i^\top\mathbf{Q}_{-i}\mathbf{z}_i}\right] \quad = \mathbb{E}\left[\frac{\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}}{1+\delta_i}\right] + O\left(\frac{1}{\sqrt{d}}\right),$$

again thanks to Hölder inequality combined with Lemma 2 that allow us to bound:

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{\delta_i - \frac{1}{Td}\mathbf{z}_i^\top\mathbf{Q}_{-i}\mathbf{z}_i}{(1+\delta_i)\left(1+\frac{1}{Td}\mathbf{z}_i^\top\mathbf{Q}_{-i}\mathbf{z}_i\right)}\right| |\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}|\right] \\ &\qquad \leq \mathbb{E}\left[\left|\delta_i - \frac{1}{Td}\mathbf{z}_i^\top\mathbf{Q}_{-i}\mathbf{z}_i\right|^2\right]^{\frac{1}{2}}\mathbb{E}\left[|\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}|^2\right]^{\frac{1}{2}} \ \leq \ O\left(\frac{1}{\sqrt{d}}\right),\end{aligned}$$

The independence between $\mathbf{z}_i$ and $\tilde{\mathbf{Q}}_{-i}$ (and $\bar{\tilde{\mathbf{Q}}}$) then allow us to deduce (again with formula (7)):

$$\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] = \mathbb{E}\left[\frac{\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}_{-i}\mathbf{v}}{1+\delta_i}\right] + \frac{1}{Td}\mathbb{E}\left[\frac{\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{v}}{1+\delta_i}\right] + O\left(\frac{1}{\sqrt{d}}\right). \tag{11}$$

Let us inject (10) and (11) in (8) to obtain (again with an application of Hölder inequality and Lemma 2 that we do not detail this time):

$$\begin{aligned}\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}\left(\frac{\mathbf{S}_i}{1+\boldsymbol{\delta}_i} - \mathbf{z}_i\mathbf{z}_i^\top\right)\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{v}\right] &= \frac{1}{Td}\mathbb{E}\left[\frac{\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\mathbf{z}_i^\top\bar{\tilde{\mathbf{Q}}}\mathbf{U}\tilde{\mathbf{Q}}\mathbf{z}_i\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{v}}{\left(1+\frac{1}{n}\mathbf{z}_i^\top\tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\right)^2}\right] + O\left(\frac{1}{\sqrt{d}}\right), \\ &= \frac{1}{Td}\frac{\mathbb{E}\left[\mathbf{u}^\top\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\tilde{\mathbf{Q}}_{-i}\mathbf{v}\right]}{(1+\delta_i)^2}\operatorname{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\bar{\tilde{\mathbf{Q}}}\right) + O\left(\frac{1}{\sqrt{d}}\right),\end{aligned}$$

Putting all the estimations together, one finally obtains:

$$\left\|\mathbb{E}\left[\tilde{\mathbf{Q}}\mathbf{U}\tilde{\mathbf{Q}}\right] - \mathbb{E}\left[\bar{\tilde{\mathbf{Q}}}\mathbf{U}\bar{\tilde{\mathbf{Q}}}\right] - \frac{1}{(Td)^2}\sum_{i=1}^n\frac{\operatorname{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\mathbf{U}\bar{\tilde{\mathbf{Q}}}\right)}{(1+\delta_i)^2}\mathbb{E}\left[\tilde{\mathbf{Q}}_{-i}\mathbf{S}_i\tilde{\mathbf{Q}}_{-i}\right]\right\| \leq O\left(\frac{1}{\sqrt{d}}\right) \tag{12}$$

One then see that if we introduce for any $\mathbf{V} \in \mathbb{R}^{n\times n}$ the block matrices:

- $\theta = \frac{1}{Td}(\frac{\mathbb{E}\left[\mathrm{tr}(\mathbf{S}_j\tilde{\mathbf{Q}}\mathbf{S}_i\tilde{\mathbf{Q}}^Y)\right]}{(1+\delta_i)(1+\delta_j)})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$
- $\theta(\mathbf{V}) = \frac{1}{Td}(\frac{\mathbb{E}\left[\mathrm{tr}(\mathbf{V}\tilde{\mathbf{Q}}\mathbf{S}_i\tilde{\mathbf{Q}}^Y)\right]}{1+\delta_i})_{i\in[n]} \in \mathbb{R}^n$,
- $\theta(\mathbf{U},\mathbf{V}) = \frac{1}{Td}\mathbb{E}\left[\mathrm{tr}(\mathbf{V}\tilde{\mathbf{Q}}\mathbf{U}\tilde{\mathbf{Q}}^Y)\right] \in \mathbb{R}$,

then, if $\|\mathbf{V}\| \leq O(1)$, multiplying (12) with $\mathbf{V}$ and taking the trace leads to:

$$\theta(\mathbf{U},\mathbf{V}) = \Psi(\mathbf{U},\mathbf{V}) + \frac{1}{Td}\Psi(\mathbf{U})^\top\theta(\mathbf{V}) + O\left(\frac{1}{\sqrt{d}}\right), \tag{13}$$

Now, taking $\mathbf{U} = \frac{\mathbf{S}_1}{1+\delta_1}, \dots, \frac{\mathbf{S}_n}{1+\delta_n}$, one gets the vectorial equation:

$$\theta(\mathbf{V}) = \Psi(\mathbf{V}) + \frac{1}{Td}\Psi\theta(\mathbf{V}) + O\left(\frac{1}{\sqrt{d}}\right),$$

When $(I_{Td} - \frac{1}{Td}\Psi)$ is invertible, one gets $\theta(\mathbf{V}) = (I_{Td} - \frac{1}{Td}\Psi)^{-1}\Psi(\mathbf{V}) + O\left(\frac{1}{\sqrt{d}}\right)$, and combining with (13), one finally obtains:

$$\theta(\mathbf{U},\mathbf{V}) = \Psi(\mathbf{U},\mathbf{V}) + \frac{1}{Td}\Psi(\mathbf{U})^\top(I_{Td} - \frac{1}{Td}\Psi)^{-1}\Psi(\mathbf{V}) + O\left(\frac{1}{\sqrt{d}}\right).$$

□

C.4 Estimation of the deterministic equivalent of $\mathbf{Q}^2$

Theorem 4. *Under the setting of Theorem 3, one can estimate:*

$$\left\|\mathbb{E}\left[\mathbf{Q}^2\right] - \mathbf{I}_n + \mathcal{D}_v\right\| \leq O\left(\frac{1}{\sqrt{d}}\right), \tag{14}$$

with, $\forall i \in [n]$*:*

$$v_i \equiv \frac{1}{Td}\frac{\mathrm{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\right)}{(1+\delta_i)^2} + \frac{1}{Td}\frac{\mathrm{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_n)\right)}{(1+\delta_i)^2}$$

Proof. The justifications are generally the same as in the proof of Theorem 3, we will thus allow ourselves to be quicker in this proof.

Using the definition of $\mathbf{Q} = \left(\frac{\mathbf{Z}^\top\mathbf{A}\mathbf{Z}}{Td} + \mathbf{I}_n\right)^{-1}$, we have that

$$\frac{\mathbf{Z}^\top\mathbf{Z}}{Td}\mathbf{Q} = \left(\frac{\mathbf{Z}^\top\mathbf{Z}}{Td} + \mathbf{I}_n - \mathbf{I}_n\right)\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{Td} + \mathbf{I}_n\right)^{-1} = \mathbf{I}_n - \mathbf{Q} \tag{15}$$

and one can then let appear $\tilde{\mathbf{Q}}$ thanks to the relation:

$$\mathbf{Z}\mathbf{Q} = \tilde{\mathbf{Q}}\mathbf{Z}, \tag{16}$$

that finally gives us:

$$\mathbf{Q} = \mathbf{I}_n - \frac{1}{Td}\mathbf{Z}^\top\mathbf{Z}\mathbf{Q} = \mathbf{I}_n - \frac{1}{Td}\mathbf{Z}^\top\tilde{\mathbf{Q}}\mathbf{Z}$$

One can then express:

$$\begin{aligned}\mathbf{Q}^2 &= \mathbf{I}_n - \frac{2}{Td}\mathbf{Z}^\top\tilde{\mathbf{Q}}\mathbf{Z} + \frac{1}{(Td)^2}\mathbf{Z}^\top\tilde{\mathbf{Q}}\mathbf{Z}\mathbf{Z}^\top\tilde{\mathbf{Q}}\mathbf{Z} \\ &= \mathbf{I}_n - \frac{1}{Td}\mathbf{Z}^\top\tilde{\mathbf{Q}}\mathbf{Z} - \frac{1}{Td}\mathbf{Z}^\top\tilde{\mathbf{Q}}^2\mathbf{Z}.\end{aligned}$$

Given $i, j \in [n]$, $i \neq j$, let us first estimate (thanks to Hölder inequality and Lemma 2):

$$\frac{1}{Td}\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}\mathbf{z}_j\right] = \frac{1}{Td}\frac{\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}_{-i,j}\mathbf{z}_j\right]}{(1+\delta_i)(1+\delta_j)} + O\left(\frac{1}{\sqrt{d}}\right) \leq O\left(\frac{1}{\sqrt{d}}\right),$$

since $\mathbb{E}[z_i] = \mathbb{E}[z_j] = 0$. Now, we consider the case $j = i$ to get:

$$\frac{1}{Td}\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}\mathbf{z}_i\right] = \frac{1}{Td}\frac{\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}_{-i}\mathbf{z}_i\right]}{(1+\delta_i)^2} + O\left(\frac{1}{\sqrt{d}}\right) = \frac{1}{Td}\frac{\operatorname{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}\right)}{(1+\delta_i)^2} + O\left(\frac{1}{\sqrt{d}}\right).$$

As before, we know that $\frac{1}{Td}\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}\mathbf{z}_j\right] \leq O\left(\frac{1}{\sqrt{d}}\right)$ if $i \neq j$. Considering $i \in [n]$, we thus are left to estimate:

$$\frac{1}{Td}\mathbb{E}\left[\mathbf{z}_i^\top \tilde{\mathbf{Q}}^2\mathbf{z}_j\right] = \frac{1}{Td}\frac{\operatorname{tr}\left(\mathbf{S}_i\bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_n)\right)}{(1+\delta_i)^2} + O\left(\frac{1}{\sqrt{d}}\right)$$

□

D Risk Estimation (Proof of Theorem 1)

D.1 Test Risk

The expected value of the MSE of the test data $\mathbf{x} \in \mathbb{R}^{T\times Td}$ concatenating the feature vector of all the tasks with the corresponding response variable $\mathbf{y} \in \mathbb{R}^{T\times Tq}$ reads as

$$\begin{aligned}
\mathcal{R}_{test}^{\infty} &= \frac{1}{T}\mathbb{E}[\|\mathbf{y} - g(\mathbf{x})\|_2^2] \\
&= \frac{1}{T}\mathbb{E}\left[\|\frac{\mathbf{x}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon} - \frac{\mathbf{x}^\top\mathbf{AZQY}}{Td}\|_2^2\right] \\
&= \frac{1}{T}\mathbb{E}\left[\|\frac{\mathbf{x}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon} - \frac{\mathbf{x}^\top\mathbf{AZQ}(\frac{\mathbf{Z}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon})}{Td}\|_2^2\right] \\
&= \frac{1}{T}\mathbb{E}\left[\|\frac{\mathbf{x}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon} - \frac{\mathbf{x}^\top\mathbf{AZQZ}^\top\mathbf{W}}{Td\sqrt{Td}} - \frac{\mathbf{x}^\top\mathbf{AZQ}\boldsymbol{\varepsilon}}{Td}\|_2^2\right] \\
&= \frac{1}{T}\mathbb{E}\left[\frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{W}\right)}{Td} - \frac{2\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{AZQZ}^\top\mathbf{W}\right)}{(Td)^2} + \operatorname{tr}\left(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\right) + \frac{\operatorname{tr}\left(\mathbf{W}^\top\mathbf{ZQZ}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{AZQZ}^\top\mathbf{W}\right)}{(Td)^3} + \right. \\
&\qquad \left. \frac{\operatorname{tr}\left(\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{AZQ}\boldsymbol{\varepsilon}\right)}{(Td)^2}\right] \\
&= \frac{1}{T}\mathbb{E}\left[\frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{W}\right)}{Td} - \frac{2\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{A}^{\frac{1}{2}}(\mathbf{I}_{Td} - \tilde{\mathbf{Q}})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}\right)}{Td} + \right. \\
&\qquad \left. \operatorname{tr}\left(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\right) + \frac{\operatorname{tr}\left(\mathbf{W}^\top\mathbf{ZQZ}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{AZQZ}^\top\mathbf{W}\right)}{(Td)^3} + \frac{\operatorname{tr}\left(\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{AZQ}\boldsymbol{\varepsilon}\right)}{(Td)^2}\right] \\
&= \frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{W}\right)}{Td} - 2\frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{A}^{\frac{1}{2}}(\mathbf{I}_{Td} - \tilde{\mathbf{Q}})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}\right)}{Td} + \operatorname{tr}\left(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\right) + \frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{W}\right)}{Td} \\
&\qquad - 2\frac{\operatorname{tr}\left(\mathbf{W}^\top\boldsymbol{\Sigma}A^{\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}A^{-\frac{1}{2}}\mathbf{W}\right)}{Td} + \frac{\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}}{Td} + \frac{1}{Td}\operatorname{tr}(\boldsymbol{\Sigma}_N\bar{\mathbf{Q}}_2) + O\left(\frac{1}{\sqrt{d}}\right)
\end{aligned}$$

The test risk can be further simplified as

$$\mathcal{R}_{test}^{\infty} = \operatorname{tr}\left(\boldsymbol{\Sigma}_N\right) + \frac{\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}}{Td} + \frac{\operatorname{tr}\left(\boldsymbol{\Sigma}_N\bar{\mathbf{Q}}_2\right)}{Td} + O\left(\frac{1}{\sqrt{d}}\right)$$

D.2 Train Risk

In this section, we derive the asymptotic risk for the training data.

Theorem 5 (Asymptotic training 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* **??***), it follows that:*

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

Proof. We aim in this setting of regression, to compute the asymptotic theoretical training risk given by:

$$\mathcal{R}_{train}^{\infty} = \frac{1}{Tn}\mathbb{E}\left[\left\|\mathbf{Y} - \frac{\mathbf{Z}^{\top}\mathbf{A}\mathbf{Z}}{Td}\mathbf{Q}\mathbf{Y}\right\|_2^2\right]$$

Using the definition of $\mathbf{Q} = \left(\frac{\mathbf{Z}^{\top}\mathbf{A}\mathbf{Z}}{Td} + \mathbf{I}_{Td}\right)^{-1}$, we have that

$$\frac{\mathbf{Z}^{\top}\mathbf{A}\mathbf{Z}}{Td}\mathbf{Q} = \left(\frac{\mathbf{Z}^{\top}\mathbf{A}\mathbf{Z}}{Td} + \mathbf{I}_{Td} - \mathbf{I}_{Td}\right)\left(\frac{\mathbf{Z}^{\top}\mathbf{A}\mathbf{Z}}{Td} + \mathbf{I}_{Td}\right)^{-1} = \mathbf{I}_{Td} - \mathbf{Q}$$

Plugging back into the expression of the training risk then leads to

$$\mathcal{R}_{train}^{\infty} = \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{Y}^{\top}\mathbf{Q}^2\mathbf{Y}\right)\right]$$

Using the definition of the linear generative model and in particular $\mathbf{Y} = \frac{\mathbf{Z}^{\top}\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon}$, we get

$$\begin{aligned}\mathcal{R}_{train}^{\infty} &= \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{\sqrt{Td}}\mathbf{Z}^{\top}\mathbf{W} + \boldsymbol{\varepsilon}\right)^{\top}\mathbf{Q}^2\left(\frac{1}{\sqrt{Td}}\mathbf{Z}^{\top}\mathbf{W} + \boldsymbol{\varepsilon}\right)\right] \\ &= \frac{1}{Tn}\frac{1}{Td}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{Z}\mathbf{Q}^2\mathbf{Z}^{\top}\mathbf{W}\right)\right] + \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\varepsilon}^{\top}\mathbf{Q}^2\boldsymbol{\varepsilon}\right)\right]\end{aligned}$$

To simplify this expression, we will introduced the so-called "coresolvent" defined as:

$$\tilde{\mathbf{Q}} = \left(\frac{\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{A}^{\frac{1}{2}}}{Td} + \mathbf{I}_{Td}\right)^{-1},$$

Employing the elementary relation $\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Q} = \tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}$, one obtains:

$$\frac{1}{Td}\mathbf{Z}\mathbf{Q}^2\mathbf{Z}^{\top} = \frac{1}{Td}\mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Q}\mathbf{Z}^{\top} = \mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}^2\frac{\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{A}^{\frac{1}{2}}}{Td}\mathbf{A}^{-\frac{1}{2}} = \mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{-\frac{1}{2}} - \mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}^2\mathbf{A}^{-\frac{1}{2}},$$

Therefore we further get

$$\mathcal{R}_{train}^{\infty} = \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{A}^{-1/2}\tilde{\mathbf{Q}}\mathbf{A}^{-1/2}\mathbf{W}\right)\right] - \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{A}^{-1/2}\tilde{\mathbf{Q}}^2\mathbf{A}^{-1/2}\mathbf{W}\right)\right] + \frac{1}{Tn}\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\varepsilon}^{\top}\mathbf{Q}^2\boldsymbol{\varepsilon}\right)\right]$$

Using deterministic equivalents in Lemma 1, the training risk then leads to

$$\mathcal{R}_{train}^{\infty} = \frac{1}{Tn}\operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{A}^{-1/2}\bar{\tilde{\mathbf{Q}}}\mathbf{A}^{-1/2}\mathbf{W}\right) - \frac{1}{Tn}\operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{A}^{-1/2}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{I}_{Td})\mathbf{A}^{-1/2}\mathbf{W}\right) + \frac{1}{Tn}\operatorname{tr}\left(\mathbf{\Sigma}_N\bar{\mathbf{Q}}_2\right) + O\left(\frac{1}{\sqrt{d}}\right)$$

□

E Interpretation and insights of the theoretical analysis

E.1 Analysis of the test risk

We recall the test risk as

$$\mathcal{R}_{test}^{\infty} = \operatorname{tr}\left(\mathbf{\Sigma}_N\right) + \frac{\mathbf{W}^{\top}\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}}{Td} + \frac{\operatorname{tr}\left(\mathbf{\Sigma}_N\bar{\mathbf{Q}}_2\right)}{Td} + O\left(\frac{1}{\sqrt{d}}\right)$$

The test risk is composed of a signal term of a signal term $\mathcal{S} = \frac{\mathbf{W}^{\top}\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}}{Td}$ and a noise term $\mathcal{N} = \frac{\operatorname{tr}\left(\mathbf{\Sigma}_N\bar{\mathbf{Q}}_2\right)}{Td}$.

E.2 Interpretation of the signal term

Let's denote by $\bar{\mathbf{\Sigma}} = \sum_{t=1}^{T} \frac{n_t d_t}{Td(1+\delta_t)^2}\mathbf{\Sigma}^{(t)}$ and $\tilde{\mathbf{\Sigma}} = \sum_{t=1}^{T} \frac{c_0}{1+\delta_t}\mathbf{\Sigma}^{(t)}$. The signal term reads as

$$\mathcal{S} = \mathbf{W}^\top \mathbf{A}^{-\frac{1}{2}} \bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}}\mathbf{W}.$$

Using the following identity,

$$\begin{aligned} \mathbf{A}^{-\frac{1}{2}} \bar{\tilde{\mathbf{Q}}}_2(\mathbf{A})\mathbf{A}^{-\frac{1}{2}} &= \mathbf{A}^{-\frac{1}{2}}\bar{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\left(\mathbf{I}+\bar{\mathbf{\Sigma}}\right)\mathbf{A}^{\frac{1}{2}}\bar{\mathbf{Q}}\mathbf{A}^{-\frac{1}{2}} \\ &= \left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1}\left(\mathbf{I}+\bar{\mathbf{\Sigma}}\right)\left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1} \end{aligned}$$

This finally leads to

$$\mathcal{S} = \mathbf{W}^\top \left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1}\left(\mathbf{I}+\bar{\mathbf{\Sigma}}\right)\left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1}\mathbf{W}$$

The matrix $\mathcal{H} = \left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1}\left(\mathbf{I}+\bar{\mathbf{\Sigma}}\right)\left(\mathbf{A}\tilde{\mathbf{\Sigma}}+\mathbf{I}\right)^{-1}$ is responsible to amplifying the signal $\mathbf{W}^\top\mathbf{W}$ in order to let the test risk to decrease more or less. It is is decreasing as function of the number of samples in the tasks n_t. Furthermore it is composed of two terms (from the independent training $\mathbf{W}_t^\top\mathbf{W}$) and the cross term $\mathbf{W}_t^\top\mathbf{W}_v$ for $t \neq v$. Both terms decreases as function of the number of samples n_t, smaller values of γ_t and increasing value of λ. The cross term depends on the matrix $\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_v$ which materializes the covariate shift between the tasks. More specifically, if the features are aligned $\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_v = I$ and the cross term is maximal while for bigger Fisher distance between the covariance of the tasks, the correlation is not favorable for multi task learning. To be more specific the off-diagonal term of $\mathcal{H}$ are responsible for the cross term therefore for the multi tasks and the diagonal elements are responsible for the independent terms.

To analyze more the element of $\mathcal{H}$, let's consider the case where $\Sigma^{(t)} = \mathbf{I}$ and $\gamma_t = \gamma$. In this case the diagonal and non diagonal elements $\mathbf{D}_{IL}$ and $\mathbf{C}_{MTL}$ are respectively given by

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

Both function are decreasing function of λ, $1/\gamma$ and c_0.

E.3 Interpretation and insights of the noise terms

We recall the definition of the noise term $\mathcal{N}$ as

$$\mathcal{N} = \operatorname{tr}\left(\mathbf{\Sigma}_N\left(\mathbf{A}^{-1}+\mathbf{\Sigma}\right)^{-1}\right)$$

Now at the difference of the signal term there are no cross terms due to the independence between the noise of the different tasks. In this case on the diagonal elements of $\left(\mathbf{A}^{-1}+\mathbf{\Sigma}\right)^{-1}$ matters. This diagonal term is increasing for an increasing value of the sample size, the value of λ. Therefore this term is responsible for the negative transfer. In the specific case where $\mathbf{\Sigma}^{(t)} = \mathbf{I}_d$ and $\gamma_t = \gamma$ for all task t, the diagonal terms read as

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

E.4 Simplified Model Insights

For $T = 2$ tasks with identical covariance and $\gamma_1 = \gamma_2 = \gamma$, the test risk simplifies to:

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

where $\mathbf{D}_{IL}$, $\mathbf{C}_{MTL}$, and $\mathbf{N}_{NT}$ represent independent learning, multi-task learning, and noise contributions, respectively, depending on γ and λ.

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

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

The interplay between signal and noise terms shows that task similarity enhances the signal, while noise induces negative transfer. Optimal λ balances these effects, given by $\lambda^\star = \frac{n}{d}\mathrm{SNR} - \frac{\gamma}{2}$, where SNR depends on task weights and noise.

E.5 Optimal Lambda

Deriving $\mathcal{R}_{test}^{\infty}$ with respect to λ leads after some algebraic calculus to

$$\lambda^\star = \frac{n}{d}\mathit{SNR} - \frac{\gamma}{2}$$

where the signal noise ratio is composed of the independent signal to noise ratio and the cross signal to noise ratio $\mathit{SNR} = \frac{\|\mathbf{W}_1\|_2^2 + \mathbf{W}_2\|_2^2}{\mathrm{tr}\boldsymbol{\Sigma}_n} + \frac{\mathbf{W}_1^\top \mathbf{W}_2}{\mathrm{tr}\boldsymbol{\Sigma}_n}$

F Theoretical Estimations

F.1 Estimation of the training and test risk

The different theorems depends on the ground truth $\mathbf{W}$ that needs to be estimated through $\hat{\mathbf{W}}$.

To estimate the test risk, one needs to estimate functionals of the form $\mathbf{W}^\top\mathbf{M}\hat{\mathbf{W}}$ and $\boldsymbol{\varepsilon}^\top\mathbf{M}\boldsymbol{\varepsilon}$ for any matrix $\mathbf{M}$. Using the expression of $\mathbf{W} = \mathbf{AZQY}$, we start computing $\hat{\mathbf{W}}^\top\mathbf{M}\hat{\mathbf{W}}$

$$\hat{\mathbf{W}}^\top\mathbf{M}\mathbf{W} = \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}^\top\mathbf{AMAZQY}$$

Using the generative model for $\mathbf{Y} = \frac{\mathbf{Z}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{W}}^\top\mathbf{M}\mathbf{W}\right] &= \mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon}\right)^\top \mathbf{Q}\mathbf{Z}^\top\mathbf{AMAZQ}\left(\frac{\mathbf{Z}^\top\mathbf{W}}{\sqrt{Td}} + \boldsymbol{\varepsilon}\right)\right] \\ &= \frac{1}{Td}\mathbb{E}\left[\mathbf{W}^\top\mathbf{ZQZ}^\top\mathbf{AMAZQZ}^\top\mathbf{W}\right] + \mathbb{E}\left[\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{AMAZQ}\boldsymbol{\varepsilon}\right]\end{aligned}$$

Employing the elementary relation $\mathbf{A}^{\frac{1}{2}}\mathbf{ZQ} = \tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}$, one obtains:

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{W}}^\top\mathbf{M}\mathbf{W}\right] &= \frac{1}{Td}\mathbb{E}\left[\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}^\top\mathbf{Z}\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{Z}\mathbf{Z}^\top\mathbf{W}\right] + \mathbb{E}\left[\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{AMAZQ}\boldsymbol{\varepsilon}\right] \\ &= \mathbb{E}\left[\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{I}-\tilde{\mathbf{Q}}\right)\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}\left(\mathbf{I}-\tilde{\mathbf{Q}}\right)\mathbf{A}^{-\frac{1}{2}}\mathbf{W}\right] + \mathbb{E}\left[\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{AMAZQ}\boldsymbol{\varepsilon}\right] \\ &= \mathbb{E}\left[\mathbf{W}^\top\mathbf{M}\mathbf{W}\right] - 2\mathbb{E}\left[\mathbf{W}^\top\mathbf{M}\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{-\frac{1}{2}}\mathbf{W}\right] + \mathbb{E}\left[\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}\tilde{\mathbf{Q}}\mathbf{A}^{-\frac{1}{2}}\mathbf{W}\right] \\ &\quad + \mathbb{E}\left[\boldsymbol{\varepsilon}^\top\mathbf{QZ}^\top\mathbf{AMAZQ}\boldsymbol{\varepsilon}\right]\end{aligned}$$

Using the deterministic equivalent of Lemma 1, we obtain

$$\begin{aligned}\hat{\mathbf{W}}^\top\mathbf{M}\hat{\mathbf{W}} &\leftrightarrow \mathbf{W}^\top\mathbf{M}\mathbf{W} - 2\mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{W} + \mathrm{tr}\boldsymbol{\Sigma}_n\mathbf{M}(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}) + \mathbf{W}^\top\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{-\frac{1}{2}}\mathbf{W} \\ &\leftrightarrow \mathbf{W}^\top\left(\mathbf{M} - 2\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}\mathbf{A}^{\frac{1}{2}}\mathbf{M} + \mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{-\frac{1}{2}}\right)\mathbf{W} + \mathrm{tr}\boldsymbol{\Sigma}_n\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}) \\ &\leftrightarrow \mathbf{W}^\top\kappa(\mathbf{M})\mathbf{W} + \mathrm{tr}\boldsymbol{\Sigma}_n\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})\end{aligned}$$

where We define the mapping $\kappa : \mathbb{R}^{Td\times Td} \to \mathbb{R}^{q\times q}$ as follows

$$\kappa(\mathbf{M}) = \mathbf{M} - 2\mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}\mathbf{A}^{\frac{1}{2}}\mathbf{M} + \mathbf{A}^{-\frac{1}{2}}\bar{\tilde{\mathbf{Q}}}_2(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{-\frac{1}{2}}.$$

F.2 Estimation of the noise covariance

The estimation of the noise covariance remains a technical challenge in this process. However, when the noise covariance is isotropic, it is sufficient to estimate only the noise variance. By observing that

$$\lim_{\lambda \to 0, \gamma \to \infty} \mathcal{R}^{\infty}_{train} = \sigma^2 \frac{\mathrm{tr}\mathbf{Q}_2}{kn},$$

we can estimate the noise level from the training risk evaluated at large γ and $\lambda = 0$.

G Multi-Task Regression

G.1 Related Work

High-Dimensional Regression Analysis. High-dimensional regression has been extensively studied in single-task settings using RMT [9] and other statistical methods [10]. These works typically focus on linear signal-plus-noise models to derive test risk based on signal parameters and noise covariance. Our research extends these concepts to MTL, providing unique insights into the effects of shared and task-specific learning. Unlike previous studies, we offer a practical approach to estimate asymptotic test risks and optimize hyperparameters, making our theoretical findings actionable within the MTL framework for multivariate forecasting.

G.2 Empirical vs. Theoretical Comparison

We validate our theory using a two-task setting with adjustable similarity α. Figure 1 compares empirical and theoretical errors across λ, showing strong alignment and accurately predicting the optimal λ.

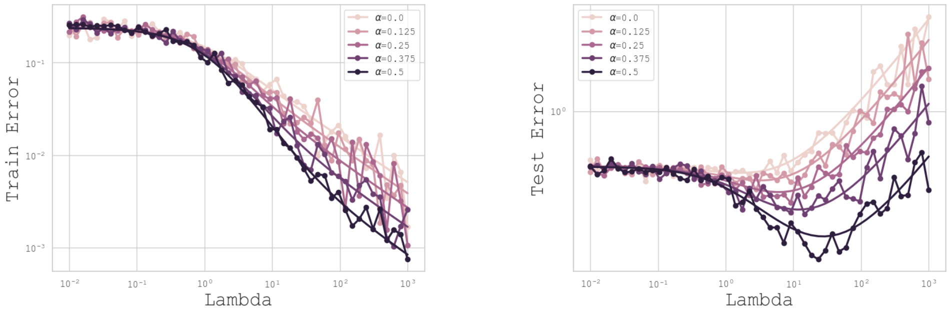


Figure 1: Empirical and theoretical MSE as functions of λ for different α. The close match highlights the theory's precision, even with moderate sample sizes.

H Multivariate Time Series Forecasting

A more general version of this work can be found at [13].

H.1 Architecture and Training Parameters

Architectures without MTL regularization. We follow Chen et al. [6], Nie et al. [22], and to ensure a fair comparison of baselines, we apply the reversible instance normalization (`RevIN`) of Kim et al. [14]. All the baselines presented here are univariate i.e. no operation is performed by the network along the channel dimension. For the `PatchTST` baseline [22], we used the official implementation than can be found on Github. The network used in `Transformer` follows the one in [12], using a single layer `Transformer` with one head of attention, while `RevIN` normalization and denormalization are applied respectively before and after the neural network function. The dimension of the model is $d_\mathrm{m} = 16$ and remains the same in all our experiments. `DLinearU` is a single linear layer applied for each channel to directly project the subsequence of historical length into the forecasted subsequence of prediction length. It is the univariate extension of the multivariate

DLinear, DLinearM, used in Zeng et al. [36]. The implementation of SAMformer can be found here, and for the iTransformer architecture here. These two multivariate models serve as baseline comparisons. We reported the results found in [12] and [17]. For all of our experiments, we train our baselines PatchTST, DLinearU and Transformer with the Adam optimizer [15], a batch size of 32 for the ETT datasets and 256 for the Weather dataset , and the learning rates summarized in Table 2.

Architectures with MTL Regularization. We implemented the univariate PatchTST, DLinearU, and Transformer baselines with MTL regularization. Initially, we scale the inputs twice using RevIN normalization. The first scaling is applied to the univariate components, and the second scaling is applied to the multivariate components. For each channel, we then apply our model without MTL regularization. The outputs are concatenated along the channel dimension, and this concatenation is flattened to form a matrix of shape (batch size, $q \times T$), where q is the prediction horizon and T is the number of channels. We then learn a square matrix W of shape $(q \times T) \times (q \times T)$ for projection and reshape the result to obtain an output of shape (batch size, q, T). This method can be applied on top of any univariate model. Our regularized loss has been introduced in 5.

Training parameters. The training/validation/test split is $12/4/4$ months on the ETT datasets and $70\%/20\%/10\%$ on the Weather dataset. We use a look-back window $d = 336$ for PatchTST and $d = 512$ for DLinearU and Transformer, using a sliding window with stride 1 to create the sequences. The training loss is the MSE. Training is performed during 100 epochs and we use early stopping with a patience of 5 epochs. For each dataset, baselines, and prediction horizon $H \in \{96, 192, 336, 720\}$, each experiment is run 3 times with different seeds, and we display the average of the test MSE over the 3 trials in Table 1.

Table 2: Learning rates used in our experiments.

Dataset	ETTh1/ETTh2	Weather
Learning rate	0.001	0.0001

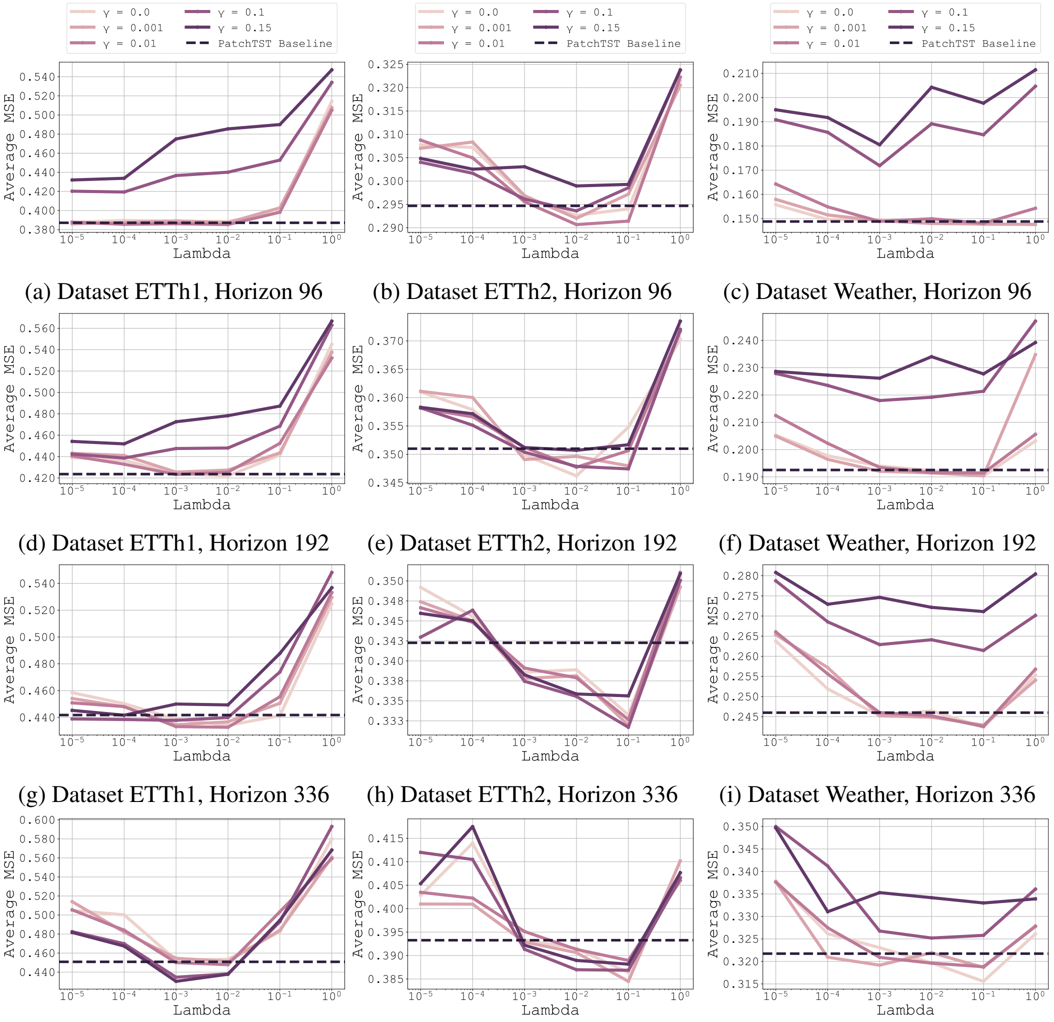
H.2 Datasets

We conduct our experiments on 3 publicly available datasets of real-world time series, widely used for multivariate long-term forecasting [6, 22, 34]. The 2 Electricity Transformer Temperature datasets ETTh1, and ETTh2 [37] contain the time series collected by electricity transformers from July 2016 to July 2018. Whenever possible, we refer to this set of 2 datasets as ETT. Weather [19] contains the time series of meteorological information recorded by 21 weather indicators in 2020. It should be noted Weather is large-scale datasets. The ETT datasets can be downloaded here while the Weather dataset can be downloaded here. Table 3 sums up the characteristics of the datasets used in our experiments.

Table 3: Characteristics of the multivariate time series datasets used in our experiments.

Dataset	ETTh1/ETTh2	Weather
# features	7	21
# time steps	17420	52696
Granularity	1 hour	10 minutes

H.3 Additional Experiments.



(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

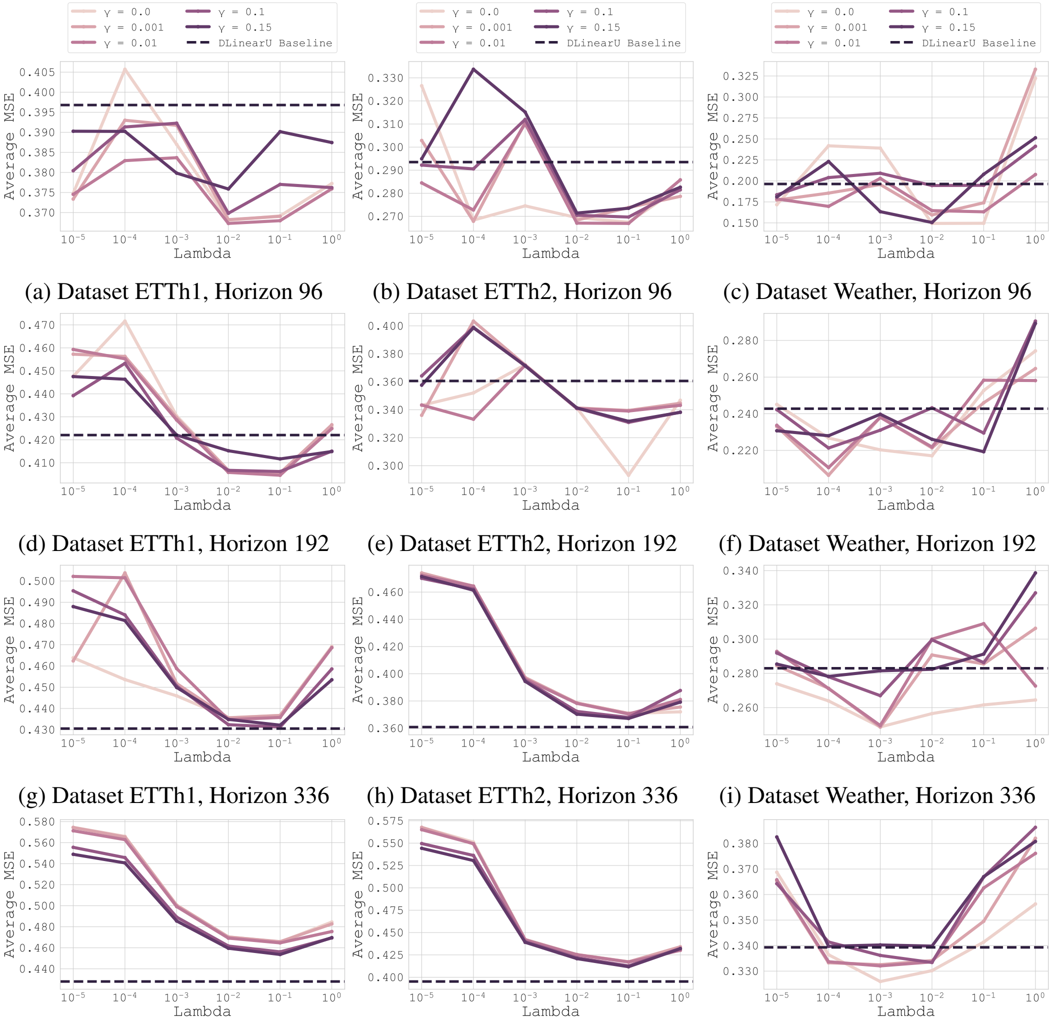
(i) Dataset Weather, Horizon 336

(j) Dataset ETTh1, Horizon 720

(k) Dataset ETTh2, Horizon 720

(l) Dataset Weather, Horizon 720

Figure 2: 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



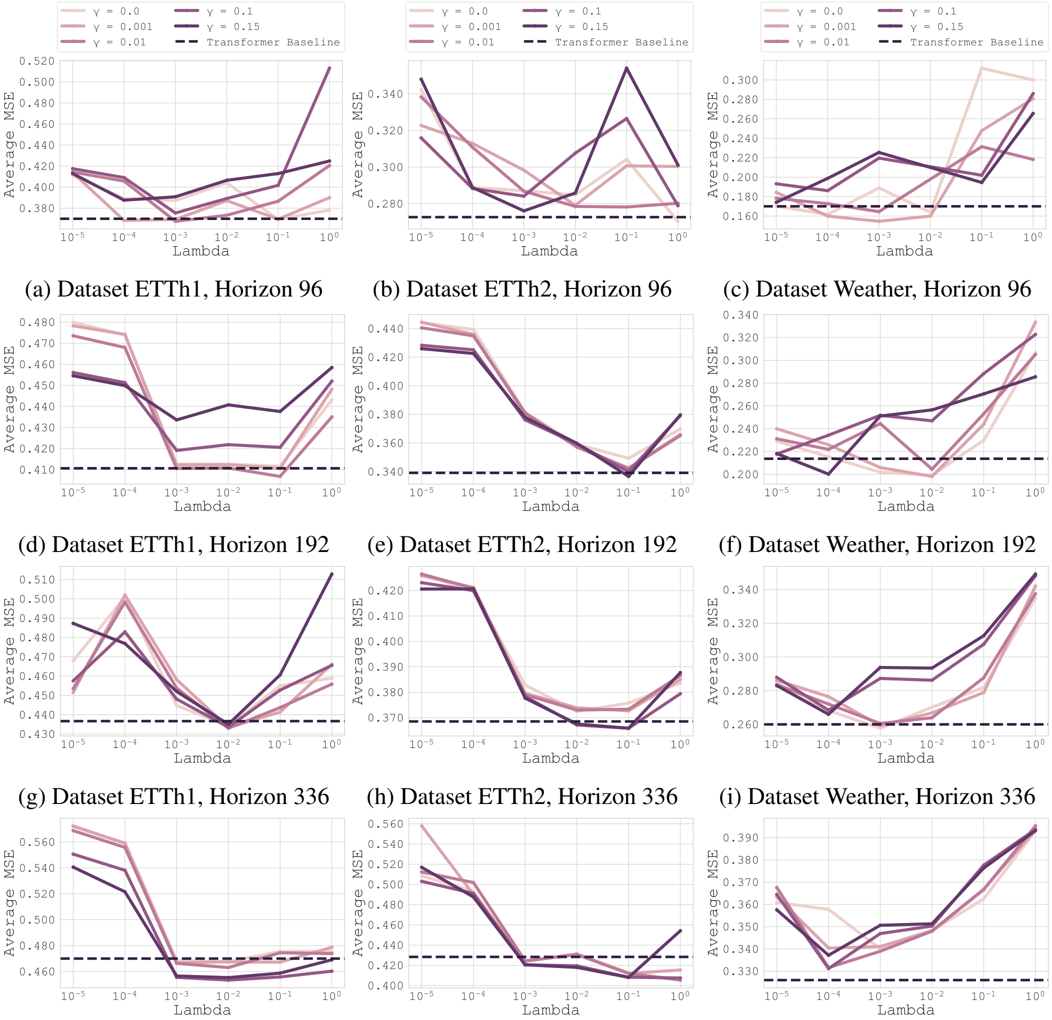
(a) Dataset ETTh1, Horizon 96 (b) Dataset ETTh2, Horizon 96 (c) Dataset Weather, Horizon 96

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



(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 `Transformer` baseline

I Limitations

While the study provides valuable insights through its theoretical analysis within a linear framework, it is important to acknowledge its limitations. The linear approach serves as a solid foundation for understanding more complex models, but its practical applications may be constrained. Linear models, though mathematically tractable and often easier to interpret, might not fully capture the intricacies and nonlinear relationships present in real-world data, especially in the context of multivariate time series forecasting.

To address this limitation, we decided to extend our algorithm's application to more complex models, specifically within the nonlinear setting of neural networks. This transition aims to evaluate whether the theoretical insights derived from the linear framework hold true empirically when applied to neural networks. As part of this endeavor, an optimal parameter lambda was selected by an oracle, leading to promising results, as detailed in Section 5. This oracle-based selection underscores the potential efficacy of our approach when appropriately tuned, even in more complex, nonlinear contexts.

It is important to note that the limitations are not related to the real-world data itself, as our setting performs well in the context of multi-task regression for real-world data, as shown in Section **??**. The difficulty arises from transitioning from a linear to a nonlinear model. The results in Section 5 are particularly encouraging, demonstrating that our method can improve upon univariate baselines by regularizing with an optimal lambda, as indicated by our oracle. While the oracle provides an upper bound on performance, actual implementation would require robust methods for hyperparameter optimization in non-linear scenarios, which remains an open area for further research.

By expanding the scope of our theoretical framework to encompass nonlinear models, we pave the way for future work that could focus on the theoretical analysis of increasingly complex architectures

Acknowledgements

Work partially funded by EU project AI4Europe (101070000).