Understanding axial attention in TSFMs through single location regression

Andrei Tiberiu Pantea*
CMAP, École polytechnique, IP Paris
Electricité de France R&D
Palaiseau, France

Aymeric Dieuleveut
CMAP, École polytechnique, IP Paris
Palaiseau, France

Abstract

Recent advances in Time Series Foundation Models (TSFMs) increasingly rely on axial attention mechanisms, but despite their empirical success, the theoretical properties of axial attention remain largely unexplored. To bridge this gap, we extend the Single Location Regression framework introduced in Marion et al. [2025] to 2D structured data, in order to study sparse information settings where a target depends on a small number of tokens. We formulate an analytically tractable axial attention-type predictor and rigorously analyze its statistical properties and training dynamics. We prove that under a specific asymptotic regime and with knowledge of the underlying oracle parameters, our predictor achieves Bayes optimality, whereas a class of semi-linear simplifications strictly fails to do so, underlining the role of the inner non-linearity. We prove that for a specific temperature scaling, Projected Gradient Descent converges globally to the optimal oracle parameters. Extensive numerical experiments validate our theoretical findings and highlight the critical role of inverse temperature scheduling in empirical convergence.

1 Introduction

Recent years have witnessed a surge in the development of Time Series Foundation Models (TSFMs) leveraging transformer architectures [Vaswani et al., 2017] (e.g., Chronos-2 [Ansari et al., 2025], Moirai 2.0 [Liu et al., 2025], TimesFM [Das et al., 2024], TabPFN-TS [Hoo et al., 2025], Toto [Cohen et al., 2025], PatchTST-FM [Wen et al., 2026]). However, applying standard global attention to high-dimensional multivariate time series presents a significant computational bottleneck. Standard attention treats inputs as a flattened 1D sequence, which scales quadratically with the total number of tokens and fails to explicitly leverage the natural 2D grid structure (time steps $\times$ channels) of multivariate data. To address this, axial attention introduced in Ho et al. [2019], and specifically for multivariate time series in Zhang and Yan [2023], has emerged as a powerful alternative already implemented in some of the state-of-the-art TSFMs, e.g., Chronos-2, TabPFN-TS and Toto, but also in task specific models such as CEReBrO [Dimofte et al., 2025], an EEG foundation model that utilizes alternating intra-channel temporal and inter-channel spatial attention and STAIT (Spatio-Temporal Alternating Iterative Transformer) [Cui et al., 2026] for remote sensing.

The axial-attention layer combines two token-mixing half-layers, as described in Figure 1. Formally, we consider our input time series as $M \times N$ matrices of tokens $\mathbb{X} = (X_{ij})_{1\leq i\leq M, 1\leq j\leq N}$, taking values in the embedding (or latent) space $\mathbb{R}^d$, where M can be seen as the number of channels and N as the number of time series patches. Generically, given as hyperparameters two real numbers λ_r, λ_c, referred to as *inverse temperatures*, and an activation function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$, the standard axial self-attention layer is parametrized by weight matrices $Q_r, K_r \in \mathbb{R}^{d\times p_r}$, $V_r \in \mathbb{R}^{d\times d}$,

*Corresponding author: andrei-tiberiu.pantea@polytechnique.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

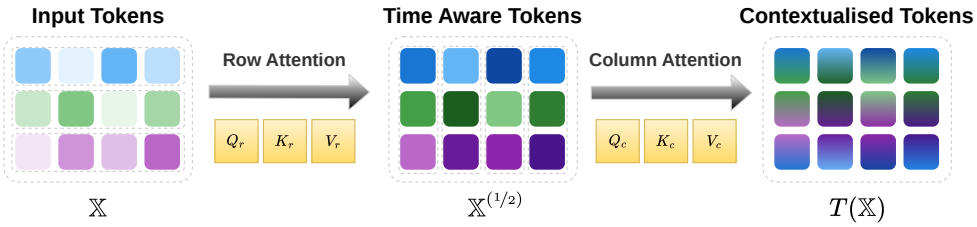


Figure 1: Schematic view of an axial attention layer. Raw input tokens are first processed by a time-wise self-attention layer that isolates features along the rows. A subsequent channel-wise attention layer vertically blends these extracted features, effectively capturing global cross-channel and temporal context.

$Q_c, K_c, V_c \in \mathbb{R}^{d \times p_c}$, and $O_c \in \mathbb{R}^{p_c \times o_c}$. The subscripts r and c represent row-wise and, respectively, column-wise operations. The mapping is defined as, for any $j \in [N]$:

$$T^{(Q_{r,c},K_{r,c},V_{r,c},O_c)}_{\lambda_r,\lambda_c,\sigma}(\mathbb{X})_{\cdot j} = \sigma\Big(\lambda_c \underbrace{\mathbb{X}^{(1/2)}_{\cdot j} Q_c}_{M\times p_c} \underbrace{K_c^T \left(\mathbb{X}^{(1/2)}_{\cdot j}\right)^T}_{p_c \times M}\Big) \underbrace{\mathbb{X}^{(1/2)}_{\cdot j} V_c}_{M\times p_c} \underbrace{O_c}_{p_c \times o_c}, \tag{1}$$

where the initial row-wise computation is, for any $i \in [M]$:

$$\mathbb{X}^{(1/2)}_{i\cdot} = \sigma(\lambda_r \underbrace{\mathbb{X}_{i\cdot} Q_r}_{N\times p_r} \underbrace{K_r^T \mathbb{X}_{i\cdot}^T}_{p_r \times N}) \underbrace{\mathbb{X}_{i\cdot} V_r}_{N\times d}, \tag{2}$$

where we use the notations $[n] = \{1, 2, \ldots, n\}$ for any positive integer n, and $W_{i\cdot}$ (resp. $W_{\cdot j}$) to denote the i-th line (resp. j-th column) of a matrix W. Note that in (2) (and throughout the paper when applied to a matrix) the activation function is used row-wise. When σ is the softmax function, this structure is one of the building blocks used in Chronos-2, TabPFN-TS and Toto, stripped of residual connections and normalization layers. Note that the order in which the row attention and column attention are performed has no impact on the results in our paper; we choose row attention followed by column attention just for presentation purposes.

Although the empirical success of these axial architectures is evident across various domains, their theoretical properties remain largely unexplored. For example, in Berthelier et al. [2026], the authors have shown that pretrained Chronos-2 is not always capable of identifying simple relationships between targets and covariates in the token-wise sparse information setting. Our goal is to improve the theoretical understanding of axial attention by analyzing their ability to approximate and learn the optimal predictor in this setting. To this end, we build on the work of Marion et al. [2025], conducted in the context of classical transformers and motivated by natural language processing (NLP) tasks. In that paper, the authors introduced a new class of statistical models, namely Single Location Regression (SLR) models, through which they aimed to investigate the ability of standard attention layers to recover relevant information in a token-wise sparse information setting, where the target depends only on a relatively small number of tokens. They have provided theoretical guarantees for a modification of the standard transformer architecture, proving that one can recover optimal predictors for SLRs. This naturally raises a fundamental, yet unanswered, question: *Can we achieve the same statistical and optimization performance with axial attention as we do with normal global attention?* The answer is not obvious, since the information flow is much more restricted in an axial transformer, where each token only attends to tokens with which it shares an axis, as opposed to the free flow of information between any two tokens in a standard transformer.

Building upon the foundation established by Marion et al. [2025], we introduce a SLR model and extend the theoretical analysis of transformer architectures to axial attention mechanisms, as a method of assessing the theoretical capabilities of current state-of-the-art TSFMs. Our main contributions are structured as follows:

- In Section 2, we introduce an SLR statistical model for the token-wise sparse information setting adapted to 2D structured data (e.g., multivariate time series, tables). We then explain the connection between such a model and multivariate time series anomaly analysis, and finally, we introduce a trainable predictor and show that it corresponds to a slightly simplified axial attention layer.

- In Section 3, we analyze the performance of the predictor introduced and prove that there exists an asymptotic regime in which it is statistically optimal for the SLR model investigated, when using the oracle parameters. Furthermore, we show that removing the inner non-linearity in our predictor renders it suboptimal.
- In Section 4, we propose an optimization algorithm for our predictor and provide theoretical guarantees for its convergence, despite the non-convexity of the problem. We further conduct extensive numerical studies which reveal a complex optimization landscape that confirm our theoretical findings.

Our structure closely follows that of Marion et al. [2025], but our motivation (the study of TSFMs) and the technical challenges are, nevertheless, very different. Standard global attention benefits from symmetric token coupling, which facilitates the derivation of closed-form analytical solutions for the expected risk. By contrast, axial attention operates in a sequential manner. This decoupled architecture breaks the global symmetry and induces a cascading non-linear dependence between axes. As a result, the optimization landscape exhibits significant complexity, rendering standard closed-form approaches to statistical optimality and convergence of optimization methods intractable and necessitating novel analytical techniques.

2 SLR model and simplified axial attention predictor

In this section, we introduce an SLR model and underline its connection to multivariate time series anomaly analysis. Afterwards, we consider a simplification of axial attention that renders the statistical analysis in Section 3 tractable.

Single location regression model. We focus on the following statistical model, where the output $Y \in \mathbb{R}$ is assumed to be given by

$$Y = X_{I_0 J_0}^T v^* + \xi, \tag{SLR}$$

where $\xi \sim \mathcal{N}(0, \epsilon^2)$ is independent noise and I_0, J_0 are discrete random variables over $[M]$ and $[N]$, respectively. No restriction on the distribution of (I_0, J_0) is imposed unless otherwise specified. Conditionally on (I_0, J_0), for $(\ell, \gamma^2, \epsilon^2, k^*, v^*) \in \mathbb{R}_+ \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$,

$$\begin{cases} X_{I_0 J_0} & \sim & \mathcal{N}\left(\ell k^*, \gamma^2 I_d\right), \\ X_{ij} & \sim & \mathcal{N}\left(0, I_d\right), \text{ for } (i,j) \neq (I_0, J_0), \end{cases}$$

where X_{ij} are independent for all $(i, j) \in [M] \times [N]$.

The parameters of (SLR) are the shift magnitude ℓ, the localization and noise variances γ^2 and ϵ^2, the unobserved projection vector v^* and the localization vector k^*, with both vectors taking values on the d-dimensional unit sphere $\mathbb{S}^{d-1}$. The target variable Y depends exclusively on a single embedded token within the $M \times N$ input matrix, positioned at random discrete coordinates (I_0, J_0). This active token is statistically identifiable because its expected value is shifted along the direction k^*. Consequently, the primary mathematical challenge of this task stems from its dynamic sparsity.

Remark 1. Unlike in Marion et al. [2025], we consider ℓ to be decoupled from d throughout our analysis, but note that for $\ell = \sqrt{d}/2$ and $\gamma^2 = 1/2$, all tokens have the same expected norm, i.e. $\mathbb{E}[\|X_{I_0 J_0}\|_2^2] = \mathbb{E}[\|X_{ij}\|_2^2] = d$, making them indistinguishable just based on their magnitude.

Connection to multivariate time series anomaly analysis. The Single Location Regression (SLR) model detailed above provides a natural theoretical framework for multivariate time series anomaly analysis, particularly when operating on learned representation embeddings $\mathbb{X}$. We consider system-wide anomalous events identifiable through correlated behaviors across multiple sensors, as pictured in Figure 2. We operate under the assumption that the embedding layer has already captured these correlated dynamics, localizing the signature of the system-wide anomaly into the embedding of a single primary driving sensor I_0 at a specific time step J_0. By investigating whether axial attention, which restricts information flow, can successfully recover the true parameters k^* and v^*, we can theoretically assess its ability to isolate and leverage localized anomalies.

Furthermore, using axial attention layers instead of standard global attention is attractive also from a computational point of view. The global attention mechanism treats the input as a flattened 1D

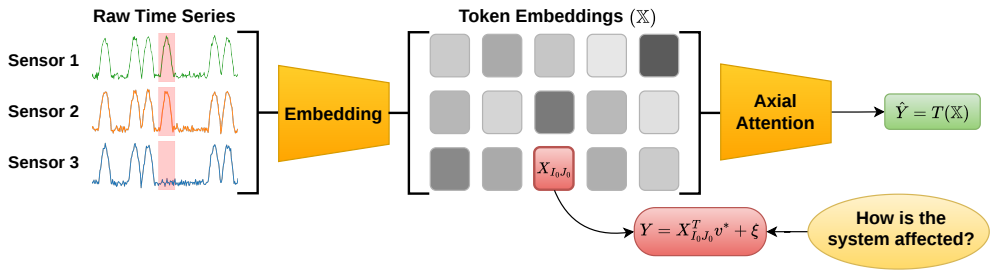


Figure 2: Under normal operating conditions, the background embeddings behave as uninformative noise, or as a standard token, modeled by the standard normal distribution $\mathcal{N}(0, I_d)$. When an anomaly occurs in one sensor, its specific signature is encoded into the token $X_{I_0J_0}$, characterized by a shifted mean ℓk^* and a localization variance $\gamma^2 I_d$. The regression target Y represents a downstream consequence or system-level anomaly score that depends strictly on identifying this specific embedded state via the projection vector v^*. Our goal is to investigate whether we can recover this target through an axial attention layer.

sequence of length $M \times N$ and this coupling requires $\mathcal{O}(M^2N^2)$ operations per layer, while axial attention only requires $\mathcal{O}(MN^2 + M^2N)$ (see Ho et al. [2019] for a derivation).

An analytically tractable simplification of axial attention for SLR. In what follows, we introduce a class of predictors that are direct simplifications of the axial attention layer. For an activation $\sigma\colon \mathbb{R}^d \to \mathbb{R}^d$ and inverse temperatures $\lambda_r, \lambda_c \in (0, \infty)$, let

$$T^{(k,v)}_{\lambda_r,\lambda_c,\sigma}(\mathbb{X}) := \sigma(\lambda_c \underbrace{\mathbb{X}^{(1/2)}k}_{M\times 1})^T \underbrace{\mathbb{X}^{(1/2)}v}_{M\times 1}, \quad \text{where } \mathbb{X}^{(1/2)}_i := \sigma(\lambda_r \underbrace{\mathbb{X}_{i\cdot}k}_{N\times 1})^T \underbrace{\mathbb{X}_{i\cdot}}_{N\times d}, \tag{3}$$

for $i \in [M]$ and learnable parameters $k, v \in \mathbb{S}^{d-1}$. Note that we slightly overload the notation and use $\mathbb{X}^{(1/2)}$ to denote the half step of both our predictor and of axial attention.

In the context of TSFMs, sequences are typically appended with designated `[REG]` tokens to act as readout tokens or attention sinks (see Ansari et al. [2025], for example). Assuming that the input $\mathbb{X}$ is appended with a $(N+1)$th column of row-aggregation `[REG_ROW]` tokens and a column aggregation `[REG_COL]` token at position $(M+1, N+1)$, we have the following lemma

Lemma 1. *For $p_r = p_c = 1$, $o_c = 1$, $O_c = 1$, together with shared keys $K_r = K_c$ and $V_r = I_d$, we have the following equivalence:*

$$T^{(Q_{r,c},K_{r,c},V_{r,c},O_c)}_{\lambda_r,\lambda_c,\sigma}(\mathbb{X})_{M+1,N+1} = T^{(K_c,V_c)}_{\lambda_r a,\lambda_c a^{(1/2)},\sigma}(\mathbb{X}),$$

for some constants a and $a^{(1/2)}$. The proof and further comments on the constraints are in Appendix A.

The main bottleneck in trying to analyze $\sigma =$ softmax in (3) is the softmax operation that induces a global coupling between the tokens, which acts as a major roadblock to rigorous statistical analysis. To bypass this issue without losing the core behavioral properties of the network, we substitute softmax with the component-wise nonlinear function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}\mathrm{d}u$, following Marion et al. [2025]. This adjustment leads to the definition of the analytically tractable predictor that we investigate throughout the remainder of the paper,

$$T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X}) := T^{(k,v)}_{\lambda_r,\lambda_c,\mathrm{erf}}(\mathbb{X}) = \mathrm{erf}(\lambda_c \mathbb{X}^{(1/2)}k)^T\, \mathbb{X}^{(1/2)}v, \quad \text{where } \mathbb{X}^{(1/2)}_i := \mathrm{erf}(\lambda_r \mathbb{X}_{i\cdot}k)^T\, \mathbb{X}_{i\cdot}\,. \tag{4}$$

3 Statistical analysis

Having established the predictor $T^{(k,v)}_{\lambda_r,\lambda_c}$, we now turn to evaluating its statistical properties.

3.1 Asymptotic optimality of the oracle predictor

In this section, we follow the approach of Marion et al. [2025]: we first observe that our predictor only depends on the alignments between k, v, k^* and v^*, projecting the high-dimensional risk landscape

down to five scalars, which is essential to the analysis of the training dynamics in Section 4. Second, under the assumption of orthogonality between k^* and v^*, we obtain a simplified formula for the risk. Finally, using this formula, we find an asymptotic regime in which our predictor is optimal. Although we follow the same structure and maintain similar notation as Marion et al. [2025], in order to optimize readability, the sequential nature of axial attention makes the results and the proofs more complicated as we lack a closed form formula for the risk. Detailed proofs for all theoretical claims in this section can be found in Appendix B.

We quantify the model's performance via its risk:

$$\mathcal{R}_{\lambda_r,\lambda_c}(k,v) \coloneqq \mathbb{E}\left[\left(Y - T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right)^2\right]. \tag{5}$$

Furthermore, we make the following assumption throughout the rest of the paper.

Assumption 1. *The vectors $k^*, v^* \in \mathbb{S}^{d-1}$ are orthogonal to each other i.e. ${k^*}^T v^* = 0$.*

Note that enforcing orthogonality between k^* and v^* is a mild requirement in high-dimensional embedding spaces, commonly found in TSFMs. The following lemma establishes a reduced representation of the risk of our axial estimator based on the alignments between k, v, k^* and v^*.

Lemma 2. *There exists a function $\mathcal{R}^<_{\lambda_r,\lambda_c} : \mathbb{R}^5 \to \mathbb{R}$ such that for any $(k,v) \in \left(\mathbb{S}^{d-1}\right)^2$,*

$$\mathcal{R}_{\lambda_r,\lambda_c}(k,v) = \mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa,\nu,\theta,\eta,\rho), \tag{6}$$

where $\kappa = k^T k^*$, $\nu = v^T v^*$, $\theta = v^T k^*$, $\eta = k^T v^*$ and $\rho = k^T v$. An expression for $\mathcal{R}^<_{\lambda_r,\lambda_c}$ is given in Appendix B.

Next, the Bayes risk for the regression task in (SLR) is bounded below by the variance ϵ^2 of the noise ξ. Building on Lemma 2, Proposition 3 demonstrates that our oracle estimator $T^{(k^*,v^*)}_{\lambda_r,\lambda_c}$ recovers this theoretical lower bound under a specific asymptotic regime.

Proposition 3. *Assume a joint asymptotic scaling where $\lambda_r \ell \to \infty$, $\lambda_c \ell \to \infty$ and $\lambda_r = o\left(M^{-\frac{1}{2}} N^{-\frac{1}{2}}\right)$. Then,*

$$\mathcal{R}_{\lambda_r,\lambda_c}(k^*,v^*) \to \epsilon^2.$$

Consequently, the oracle axial predictor $T^{(k^*,v^*)}_{\lambda_r,\lambda_c}$ achieves asymptotic Bayes optimality.

Note that Proposition 3 does not impose any scaling on d, which is expected given Lemma 2. We provide a non-rigorous intuition for the imposed scalings. First, we wish $X_{I_0 J_0}$ to pass through the first layer asymptotically unaltered, that is,

$$X_{I_0 J_0} \approx \mathbb{X}^{(1/2)}_{I_0} = \mathrm{erf}(\underbrace{\lambda_r X^T_{I_0 J_0} k^*}_{\sim \lambda_r \ell}) X_{I_0 J_0} + \sum_{j \neq J_0} \mathrm{erf}(\lambda_r X^T_{I_0 j} k^*) X_{I_0 j}.$$

Therefore, we need $\lambda_r \ell \to \infty$ to ensure that the argument of the first erf goes to ∞ and $\lambda_r \to 0$ to ensure that the others go to 0. In an analogous manner, we wish $X_{I_0 J_0}$ to pass unaltered through the second layer, which leads to $\lambda_c \ell \to \infty$. Lastly, in the expansion of $T^{(k^*,v^*)}_{\lambda_r,\lambda_c}$, there appear $(M-1)N$ noise terms of the form $C \cdot \mathrm{erf}(\lambda_r X^T_{ij} k^*) X^T_{ij} v^*$, with $C < 1$, one for each (i,j) with $i \neq I_0$. Multiple scalings could make the sum of these terms go to 0, one of them being $\lambda_r = o\left(M^{-\frac{1}{2}} N^{-\frac{1}{2}}\right)$, as shown in the proof in Appendix B.

3.2 Removing the inner non-linearity is suboptimal

To fully appreciate the asymptotic optimality of our oracle predictor, it is instructive to compare it against semi-linear models, which replace the inner non-linearity with a fixed linear combination layer governed by $\beta \in \mathbb{R}^N$. This justifies the necessity for the inner non-linearity in the axial attention layer in order to achieve asymptotic optimality. More precisely, consider the *semi-linear* predictor

$$T^{(k,v)}_{\beta,\lambda_c}(\mathbb{X}) \coloneqq \mathrm{erf}(\lambda_c \mathbb{X}^{(1/2)} k)^T \, \mathbb{X}^{(1/2)} v, \quad \text{where } \mathbb{X}^{(1/2)}_i \coloneqq \beta^T \, \mathbb{X}_{i\cdot}\,. \tag{7}$$

Then we have the following sub-optimality statement.

Proposition 4. *Let $p_j = \mathbb{P}(J_0 = j)$ for $j \in [N]$. Denote by $S_i = \beta^T \mathbb{X}_i$. Then, for any function $f\colon \mathbb{R}^{dM} \to \mathbb{R}$ in $L^2(\mathbb{R}^{dM})$, the semi-linear predictor achieves a risk strictly larger than ϵ^2 for any choice of $\beta \neq 0$ and hyperparameters M, N, ℓ, λ_c i.e.*

$$\mathbb{E}\left[\left(Y - f(S_1, S_2, \dots, S_M)\right)^2\right] \geq \epsilon^2 + \gamma^2 \left(1 - \sum_{j_0=1}^{N} p_{j_0} \frac{\gamma^2 \beta_{j_0}^2}{\gamma^2 \beta_{j_0}^2 + \sum_{j \neq j_0} \beta_j^2}\right) > \epsilon^2,$$

if there is no $j \in [N]$ with $p_j = 1$. We provide a proof of this statement in Appendix B.2.

Remark 2. Discarding the outer non-linearity leads to a less restrictive information flow compared to our predictor, and it amounts to a simplification of the standard attention mechanism. Indeed,

$$T^{(k,v)}_{\lambda_r,\beta}(\mathbb{X}) \coloneqq \sum_{i=1}^{M} \beta_i \sum_{j=1}^{N} \operatorname{erf}\left(\lambda_r X_{ij}^T k\right) X_{ij}^T v$$

is a modification of the trainable predictor proposed in Marion et al. [2025], which is asymptotically optimal for $\beta_i = 1$, $i \in [M]$, under the joint scaling $\lambda_r = o\left(M^{-\frac{1}{2}} N^{-\frac{1}{2}}\right)$ and $\lambda_r \ell \to \infty$.

4 Gradient descent dynamics

The last section of our analysis is concerned with the theoretical training dynamics of our predictor (4) for the risk $\mathcal{R}_{\lambda_r,\lambda_c}(k, v)$. While in practice training minimizes an empirical risk evaluated over a finite dataset, we limit our investigation of the empirical risk minimization to numerical simulations, which confirm our findings about the theoretical risk.

4.1 Convergence of gradient descent

Lemma 2 shows that on the manifold $\left(\mathbb{S}^{d-1}\right)^2$ the risk can be described through 5 alignment metrics. We want to use this fact and formula since it renders the analysis tractable after further constraints. Therefore, it is natural to employ Projected Riemannian Gradient Descent (PGD) as an optimization algorithm for two main reasons: it keeps the parameters on the manifold and it guarantees that the Euclidean projection back onto the sphere remains well-defined by preventing the parameters from reaching the origin. We emphasize that this geometric constraint is necessary because our reduced risk formulation (6) is strictly valid only on this manifold.

Algorithm 1 Projected Riemannian Gradient Descent

Input: Initialization $(k_0, v_0) \in (\mathbb{S}^{d-1})^2$, step size $\alpha > 0$, inverse temperatures sequence $(\lambda_r^t, \lambda_c^t)_{t \geq 0}$

for $t \geq 0$ **do**

$k_{t+1} \leftarrow \operatorname{Proj}_{\mathbb{S}^{d-1}}\left(k_t - \alpha(I_d - k_t k_t^T)\nabla_k \mathcal{R}_{\lambda_r^t,\lambda_c^t}(k_t, v_t)\right)$

$v_{t+1} \leftarrow \operatorname{Proj}_{\mathbb{S}^{d-1}}\left(v_t - \alpha(I_d - v_t v_t^T)\nabla_v \mathcal{R}_{\lambda_r^t,\lambda_c^t}(k_t, v_t)\right)$

end for

Analyzing the dynamics of the resulting 5-dimensional system is intractable due to the highly complex nature of $\mathcal{R}^{<}_{\lambda_r,\lambda_c}$. To enable a rigorous analysis, we restrict our focus to scenarios in which the parameters are initialized on the submanifold $\mathcal{M}$ defined by

$$\mathcal{M} \coloneqq \left\{(k, v) \in \left(\mathbb{S}^{d-1}\right)^2 \;\middle|\; k^T v^* = 0,\ v^T k^* = 0,\ k^T v = 0\right\}. \tag{8}$$

Under Assumption 1, the true underlying parameters (k^*, v^*) naturally reside within $\mathcal{M}$ and, crucially, the training parameters remain on the submanifold throughout the optimization process if they are initialized on it, from Lemma 6 in Appendix C. Due to this invariance, we only need to characterize the dynamics of the system restricted to $\mathcal{M}$ when initialized on this manifold. Although in practice this cannot be enforced, we show empirically in Section 4.2 that penalizing a proxy for the distance to the manifold leads to similar dynamics. This lower-dimensional analysis is mathematically tractable and leads to the primary optimization result of this paper, which guarantees asymptotic recovery of the oracle parameters.

Theorem 5. *Consider $\tau > 0$. Then, there exists $\bar{\lambda} > 0$ and $\bar{\alpha} > 0$ such that for constant inverse temperatures $\bar{\lambda} > \lambda_r > 0$ and $\lambda_c = \tau\lambda_r$, for any step size $\alpha \leq \bar{\alpha}$ and for a generic initialization $(k_0, v_0) \in \mathcal{M}$,*

$$(k_t, v_t) \to \pm(k^*, v^*) \quad \textit{as} \quad t \to \infty.$$

The convergence to $(-k^*, -v^*)$ is not a problem, as by symmetry around 0, $T^{(k^*,v^*)}_{\lambda_r,\lambda_c} = T^{(-k^*,-v^*)}_{\lambda_r,\lambda_c}$.

Remark 3. Although some assumptions made throughout the paper are identical to the ones in Marion et al. [2025], our main result Theorem 5 is qualitatively different from theirs. In their simplified standard (single-axis) attention layer case, convergence is always reached, regardless of the inverse temperatures. On the contrary, here, as shown analytically in the proof of Theorem 5 and numerically in Section 4.2 and Appendix E.4, the values of the inverse temperatures greatly impact the optimization landscape and convergence is not attainable in general. In particular, global convergence is only achieved below a specific threshold $\bar{\lambda}$.

Remark 4 (Training and inference temperatures). The inverse temperatures used during optimization and inference serve different purposes and are therefore allowed to differ. During training, we use inverse temperatures $(\lambda_r^{\text{train}}, \lambda_c^{\text{train}})$ satisfying the conditions of Theorem 5. In particular, they are sufficiently small and satisfy $\lambda_c^{\text{train}} = \tau\lambda_r^{\text{train}}$, which guarantees recovery of (k^*, v^*) up to their joint sign for generic initializations on $\mathcal{M}$. After training, the recovered parameters are frozen and the inverse temperatures are reset to inference values $(\lambda_r^{\text{inf}}, \lambda_c^{\text{inf}})$ satisfying the scaling conditions of Proposition 3. The resulting predictor is then asymptotically Bayes optimal.

Thus, the two results provide a sequential two-stage guarantee: parameter recovery is performed in an optimization-friendly temperature regime, followed by prediction in a statistically optimal inference regime. This separation is also used in our experiments, where the evaluation risk is computed using fixed inverse temperatures independently of the schedule used for optimization.

4.2 Understanding the dynamics - numerical experiments

In practice, one cannot use the theoretical risk for optimization using PGD, since there is no closed form formula for it. Furthermore, one also does not have knowledge of what values for λ_r and λ_c are small enough. In our experiments, we bypass these issues by using batch stochastic PGD for the optimization process and implementing a decreasing schedule for the inverse temperatures.

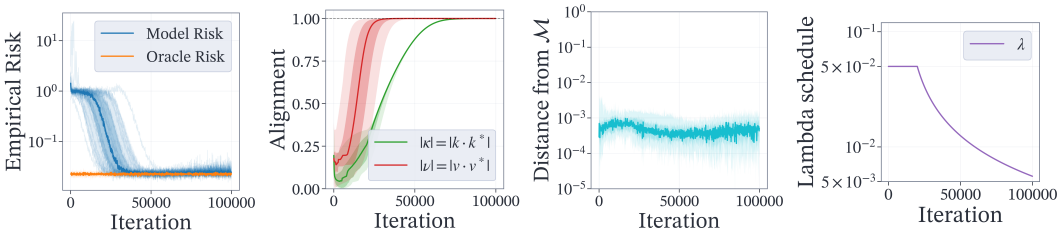
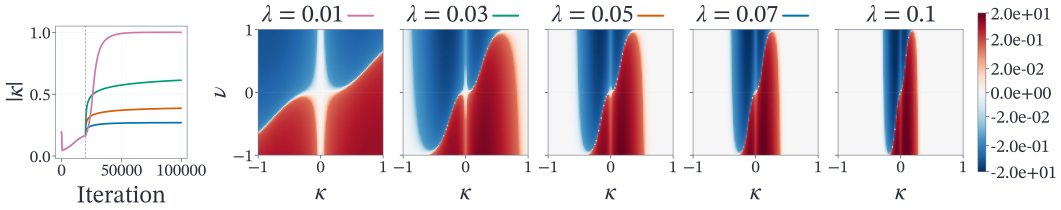


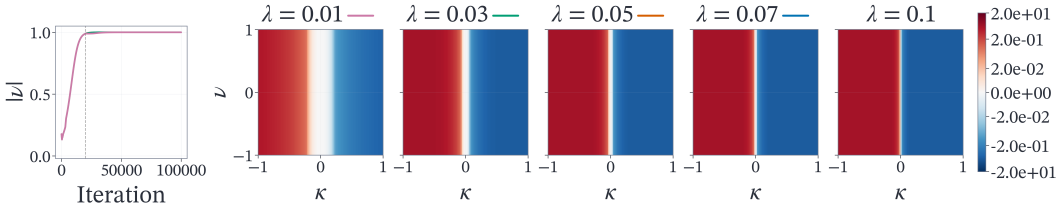
Figure 3: Dynamics of stochastic PGD to the oracle parameters over 20 independent runs. The solid curve represents the empirical median across the runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). **Left:** The empirical risk obtained by our predictor on the training batches of data using both the training parameters k, v and the oracle parameters k^*, v^*. For computing the oracle and model losses we use the initial inverse temperatures. **Middle-left:** The alignment between the training and oracle parameters $\kappa = k^T k^*$, $\nu = v^T v^*$. **Middle-right:** The distance from the manifold $\mathcal{M}$ of the training parameters (k, v). **Right:** The schedule used for $\lambda = \lambda_r = \lambda_c$. The hyperparameters of the experiment are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$ and $\lambda_t = 5 \cdot 10^{-2}/(1 + \max(t - 2 \cdot 10^4, 0) \cdot 10^{-4})$.

Figure 3 presents numerical results of the optimization obtained for initializations on the manifold using stochastic PGD. We perfectly recover the oracle parameters, up to numerical error, and there is no departure from the manifold, again, up to numerical error and noise, as was predicted by theory.

We emphasize the fact that the schedule for the inverse temperatures is crucial for the dynamics. To understand this, in Figure 4 we run an experiment in which we train 4 predictors using different schedulers: 20k warm-up steps with $\lambda_c = \lambda_r = 0.1$, followed by different values of constant inverse temperatures $\lambda = \lambda_c = \lambda_r$. The leftmost plots show two different behaviors for ν and κ: ν converges fast, usually before the warm up period ends, while the convergence of κ is highly dependent on the value of λ. The heat maps present the respective partial derivatives of $\mathcal{R}^{<}_{\lambda_r,\lambda_c}$ for the different values of λ. They show that for λ small enough, minus the gradient always points towards the global minima when far from the origin. For large values of λ, the dynamics push away from the origin much faster, but the gradients vanish and have the wrong sign close to the boundaries $\kappa = \pm 1$. As a consequence, we see that for those values, κ does not converge to one, while it does for $\lambda = 0.01$. The warm start is needed, as for small λ, the dynamics would remain stuck around the axis $\kappa = 0$ where the values of the partial derivatives are extremely small, as visible on the plot.



(a) Alignment κ and the partial derivative $\partial_\kappa \mathcal{R}^{<}_{\lambda,\lambda}$.



(b) Alignment ν and the partial derivative $\partial_\nu \mathcal{R}^{<}_{\lambda,\lambda}$.

Figure 4: **Leftmost:** Alignment dynamics during the training of our predictor with an inverse temperature warm-up period of 20k steps using $\lambda_c = \lambda_r = 0.1$, followed by constant $\lambda_c = \lambda_r = \lambda$ for $\lambda \in \{0.01, 0.03, 0.05, 0.07\}$. We plot the median across 20 runs for each value of λ. The dashed vertical line represents the warm up period. **Center-left to right:** The numerically computed gradients of $\mathcal{R}^{<}_{\lambda_r,\lambda_r}(\kappa, \nu)$ for $\lambda_r = \lambda_c = \lambda \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$, $N = 5$ and $\ell = 100$. The gradient, given by (32), is approximated using the Monte Carlo method with a sample size of 10,000.

One last and very important problem remains: in practice we cannot enforce the initialization to be on the manifold $\mathcal{M}$, since that would require knowing the oracle parameters k^* and v^*. We can ensure that $k^T v = 0$ at initialization and penalize their inner product throughout the training. Furthermore, a noisy proxy for $v^T k^*$ is $\frac{1}{\ell} v^T \sum_{i,j} X_{ij}$. Therefore, in the case of initialization outside the manifold, we impose $k^T v = 0$ at initialization and optimize for the *penalized* risk

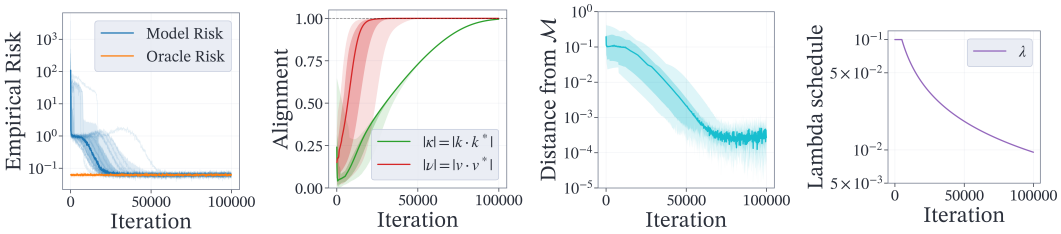
$$\bar{\mathcal{R}}_{\lambda_r,\lambda_c}(k, v) = \mathcal{R}_{\lambda_r,\lambda_c}(k, v) + \mu_1 (k^T v)^2 + \mu_2 \mathbb{E}\big[\big(v^T \textstyle\sum_{i,j} X_{ij}\big)^2\big], \tag{9}$$

where $\mu_1, \mu_2 \in (0, \infty)$ are fixed penalization coefficients. A similar use of regularization to encourage the orthogonality conditions defining an invariant manifold is explored by Maulen Soto et al. [2025] for attention-based clustering, using a different penalty to handle initializations outside the manifold.

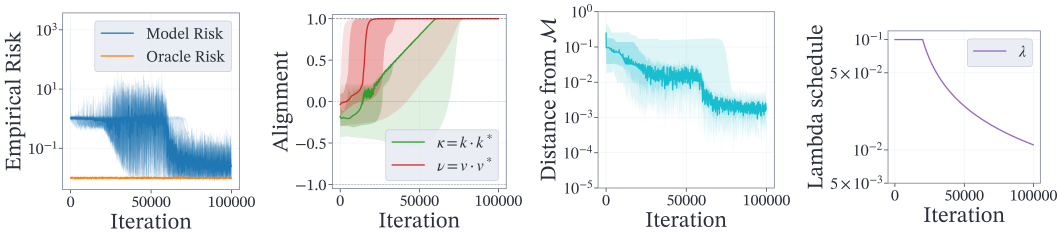
Indeed, Figure 5 presents the results obtained when optimizing for this penalized empirical risk from a generic random initialization on $(\mathbb{S}^{d-1})^2$. The top row (Figure 5a) displays the training dynamics for our trainable predictor (4), employing the erf non-linearity. We observe that the penalty terms effectively guide the parameters towards the invariant manifold $\mathcal{M}$ throughout the training and the convergence is slower, even though we have used a higher initial inverse temperature. The lower row (Figure 5b) presents the identical experimental setup applied to the predictor (3) with $\sigma = \text{softmax}$. Although the training dynamics are not identical, both models converge on the manifold $\mathcal{M}$ and there are phase transitions in the empirical loss curves for both models. In the case of the erf model, the phase transition appears to be correlated with ν reaching near-perfect alignment, while in the

case of the softmax model, the phase transition presents itself when both ν and κ reach near-perfect alignment.

As we have shown in the present experiments, the inverse temperature schedule is of the utmost importance to the performance of the training procedure, and finding a suitable one that performs well across all initializations adds a layer of complexity to the optimization process. A more in-depth analysis of and further details related to the numerical experiments can be found in Appendix E.



(a) Training dynamics of the trainable predictor (4) (i.e. (3) with $\sigma = \text{erf}$).



(b) Training dynamics of the softmax predictor (3) with $\sigma = \text{softmax}$.

Figure 5: Convergence of stochastic PGD over 20 runs with random initializations on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, optimizing for the penalized empirical risk (9). The solid curve represents the empirical median across the runs. The hyperparameters of the experiment are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$, $\mu_1 = 10$, $\mu_2 = 1$. The inverse temperature schedules are $\lambda_t = 10^{-1}/(1 + \max(t - 5 \cdot 10^3, 0) \cdot 10^{-4})$, for the erf predictor and $\lambda_t = 10^{-1}/(1 + \max(t - 2 \cdot 10^4, 0) \cdot 10^{-4})$ for the softmax predictor. For computing the oracle and model losses we use the initial inverse temperatures.

5 Conclusion and discussion

We presented a rigorous theoretical framework for understanding axial attention in TSFMs by extending the SLR model to 2D grids. We proved that an analytically tractable simplification of axial attention achieves asymptotic Bayes optimality for the oracle parameters, whereas semi-linear models in which the inner non-linearity is removed strictly fail to do so. Furthermore, we established that Projected Gradient Descent converges to these optimal parameters for small enough inverse temperatures. Extensive numerical experiments validate our theoretical findings and highlight the critical role of inverse temperature scheduling in converging to the global minima within the non-convex optimization landscape.

Our statistical framework assumes fixed temporal and spatial dimensions, independent background tokens, and a target that depends on a single localized token. Relaxing these assumptions provides several natural directions for future work. In particular, incorporating autoregressive temporal dynamics and dense cross-channel correlations would better reflect the structure of real multivariate time series, while allowing the dimensions or number of informative tokens to grow would broaden the range of sparse information settings covered by the model. These extensions would introduce additional dependencies into both the statistical model and the optimization landscape and would therefore require analytical tools beyond those used in the present work.

It is also necessary to explicitly reiterate the architectural constraints that make the analysis possible. Our results are heavily dependent on a rank-1 constraint applied to the query and key projection

matrices, as well as the substitution of the standard softmax activation with the erf function and the shared-key constraint $K_r = K_c$. The theoretical results in this paper apply to this simplified erf predictor, while our experiments with the corresponding softmax predictor provide encouraging evidence that similar optimization behavior persists beyond the activation function used in the analysis. Full TSFMs additionally contain realistic projection matrices, residual connections, normalization layers, and multiple axial-attention blocks; the implications of our results for these architectures are therefore qualitative. Nevertheless, our Single Location Regression model preserves the core structural challenge of identifying and exploiting a sparsely located informative token, making it highly relevant for the theoretical study of axial attention in TSFMs.

Finally, while our current focus is restricted to a single informative token, the same framework suggests a natural extension to multiple-token regression through multi-head attention. If the target depends on several informative locations, different heads could learn distinct localization directions and specialize in selecting different tokens, while their outputs could be combined to recover the target. Under suitable identifiability and separation assumptions, parts of the statistical analysis could then be applied head-wise, following an approach similar to the one developed here. The optimization problem would, however, become substantially richer: the heads would be trained jointly, their roles would be interchangeable, and several heads could compete for or collapse onto the same informative location. We leave this extension, together with the study of the corresponding statistical and optimization guarantees, for future work.

AI use statement. We have used generative AI tools to assist with writing and cleaning code, and with editing the manuscript. The authors reviewed all AI-assisted outputs and take responsibility for the final content of this work.

Acknowledgments. We thank Pierre Marion, whose insightful comments helped us improve the final version of this paper. The work of Andrei Tiberiu Pantea and Aymeric Dieuleveut is supported by French State aid managed by the Agence Nationale de la Recherche (ANR) under the France 2030 program ANR-23-IACL-0005 (Hi! PARIS).

References

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Gaspard Berthelier, Mariia Baranova, Andrei-Tiberiu Pantea, Etienne Le Naour, Adrien Petralia, Tahar Nabil, and Themis Palpanas. Investigating simple target-covariate relationships for Chronos-2 and TabPFN-TS. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models, 2025.

Yukun Cui, Jiangshe Zhang, Haowen Bai, Zixiang Zhao, Lilun Deng, Shuang Xu, and Chunxia Zhang. STAIT: A spatio-temporal alternating iterative transformer for multi-temporal remote sensing image cloud removal. *Remote Sensing*, 18(4), 2026. ISSN 2072-4292. doi: 10.3390/rs18040596.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR, 21–27 Jul 2024.

Alexandru Dimofte, Glenn Anta Bucagu, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Luca Benini, and Yawei Li. Cerebro: Compact encoder for representations of brain oscillations using efficient alternating attention. *arXiv preprint arXiv:2501.10885*, 2025.

O Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *The Thirteenth International Conference on Learning Representations*, 2025.

Rodrigo Maulen Soto, Pierre Marion, and Claire Boyer. Attention-based clustering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 66455–66506. Curran Associates, Inc., 2025. doi: 10.52202/085713-2231.

Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yunshi Wen, Wesley M Gifford, Chandra Reddy, Lam M Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models. *arXiv preprint arXiv:2602.06909*, 2026.

A. Yu, D.C. Maddix, B. Han, X. Zhang, A.F. Ansari, O. Shchur, C. Faloutsos, A.G. Wilson, M.W. Mahoney, and Y. Wang. Understanding transformers for time series: Rank structure, flow-of-ranks, and compressibility. In *International Conference on Learning Representations*, 2026.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

A Proof of Lemma 1 and further discussion about architectural constraints

We recall the definition of the standard axial self-attention layer.

$$T^{(Q_{r,c},K_{r,c},V_{r,c},O_c)}_{\lambda_r,\lambda_c,\sigma}(\mathbb{X})_{\cdot j} = \sigma\left(\lambda_c \underbrace{\mathbb{X}^{(1/2)}_{\cdot j} Q_c}_{M\times p_c} \underbrace{K_c^T \left(\mathbb{X}^{(1/2)}_{\cdot j}\right)^T}_{p_c\times M}\right) \underbrace{\mathbb{X}^{(1/2)}_{\cdot j} V_c}_{M\times p_c} \underbrace{O_c}_{p_c\times o_c},$$

where the initial row-wise computation is:

$$\mathbb{X}^{(1/2)}_{i\cdot} = \sigma(\lambda_r \underbrace{\mathbb{X}_{i\cdot} Q_r}_{N\times p_r} \underbrace{K_r^T \mathbb{X}_{i\cdot}^T}_{p_r\times N}) \underbrace{\mathbb{X}_{i\cdot} V_r}_{N\times d}$$

In order to justify our simplified estimator (3), we explicitly map it to the mechanism of an actual axial attention block. As mentioned in the main paper, sequences are typically appended with designated `[REG]` tokens to act as either readout tokens or attention sinks in the context of TSFMs. Following this convention, we assume that the input's final column consists of specialized `[REG_ROW]` tokens. To collapse the 2D grid into a single regression scalar, we then incorporate an additional `[REG_COL]` token at the specific index $(M+1, N+1)$, which acts as a global aggregator during the subsequent

column-wise operation. We further assume the two [REG] tokens do not attend to themselves, which is a very light assumption, since a priori these tokens contain no information. Isolating the output for this specific token yields the following

$$\mathbb{X}^{(1/2)}_{i,N+1} = \sum_{l=1}^{N} \sigma(\lambda_r X^T_{[\text{REG_ROW}]} Q_r K_r^T \mathbb{X}_{i\cdot}^T)_l X_{il}^T V_r = \sigma(\lambda_r a K_r^T \mathbb{X}_{i\cdot}^T)\mathbb{X}_{i\cdot} \in \mathbb{R}^d$$

if we assume that the projection $V_r \in \mathbb{R}^{d\times d}$ is the identity matrix and encapsulate the query interaction as a vector $a \coloneqq X^T_{[\text{REG_ROW}]} Q_r \in \mathbb{R}^{1\times p_r}$. By applying an analogous definition for the column-wise query $a^{(1/2)} \coloneqq X^T_{[\text{REG_COL}]} Q_c \in \mathbb{R}^{1\times p_c}$, the output of the final attention block becomes:

$$T^{(Q_{r,c},K_{r,c},V_{r,c},O_c)}_{\lambda_r,\lambda_c,\sigma}(\mathbb{X})_{M+1,N+1} = \sum_{i=1}^{M} \sigma\left(\lambda_c a^{(1/2)} K_c^T \mathbb{X}^{(1/2)}_{i,N+1}\right)_i \left(\mathbb{X}^{(1/2)}_{i,N+1}\right)^T V_c O_c$$

To perfectly align this with our theoretical model, we enforce rank-1 restrictions on the projections, setting $p_r = p_c = 1$. With these constraints in place, along with $o_c = 1$, $O_c = 1$, and shared keys $K_r = K_c$, the full axial mechanism cleanly reduces to our formulated predictor:

$$T^{(Q_{r,c},K_{r,c},V_{r,c},O_c)}_{\lambda_r,\lambda_c}(\mathbb{X})_{M+1,N+1} = T^{(K_c,V_c)}_{\lambda_r a,\lambda_c a^{(1/2)},\sigma}(\mathbb{X})$$

Discussion. In this paragraph, we address the main architectural constraints and modifications imposed in our analysis.

- REG **tokens.** The addition of REG tokens is not a relevant modification, since any transformer layer whose output must be a single value, be it scalar of vector, will implement a global aggregation mechanism for the input tokens and will be trained for it. REG tokens remove the inductive biases that come with pooling methods or positional decoding and allow for a learned aggregation mechanism.
- **Rank-1 restriction on projection matrices.** This constraint is not at all harsh in the context of SLR, where there exists a single informative direction in the latent space, namely k^*. Moreover, Yu et al. [2026] has shown that patch-based TSFMs admit low-rank approximations of the projection matrices.
- **Shared keys.** $K_r = K_c$ is indeed a strong constraint in practice. However, in the particular case of the studied SLR, there is a single special direction in the embedding space, k^*, and, heuristically, token selection should be made according to this direction.
- **Non-linearity.** For the theoretical analysis, we have restricted the value of σ to erf, maintaining the same convention as in Marion et al. [2025]. However, only a couple of properties of the erf function are essential to the present proofs, such as global boundedness of erf and of its derivatives, the asymptotics $\lim_{x\to\infty} \text{erf}(x) = 1$ and $\lim_{x\to-\infty} \text{erf}(x) = -1$, $\text{erf}(x) = 0$ etc. Extending the proofs and the numerical analysis to other activation functions seems feasible and would be an interesting further research direction. Furthermore, the statistical role of the activation function in single-location regression is investigated by Duranthon et al. [2026], who establish advantages of softmax over alternative activations in a single-axis setting. Understanding how these comparisons extend to sequential axial attention is an interesting direction for future work.

B Proofs for Section 3

In what follows, we shall denote by

$$\mathcal{R}^<(\kappa,\nu) \coloneqq \mathcal{R}^<(\kappa,\nu,0,0,0),$$

the value of the risk when restricted to $\mathcal{M}$.

B.1 Proofs of Lemma 2 and of Proposition 3

We recall the formula for the risk

$$\mathcal{R}_{\lambda_r,\lambda_c}(k,v) = \mathbb{E}\left[\left(Y - \sum_{i=1}^{M} \text{erf}\left(\lambda_c \sum_{j=1}^{N} \text{erf}(\lambda_r X_{ij}^T k) X_{ij}^T k\right) \sum_{j=1}^{N} \text{erf}(\lambda_r X_{ij}^T k) X_{ij}^T v\right)^2\right] \quad (10)$$

Let us now denote by $G^u_{ij} = X^T_{ij}u$, for $u \in \mathbb{R}^d$. Note that we consider $X_{I_0J_0} = \ell k^* + \gamma \tilde{X}_{I_0J_0}$ and $G^u_{I_0J_0} = \tilde{X}^T_{I_0J_0}u$. Then, our predictor only depends on κ, θ and the alignment random variables $G^v_{ij}, G^{v^*}_{ij}, G^k_{ij}$ which are i.i.d. and follow the distribution

$$\begin{pmatrix} G^k_{ij} \\ G^v_{ij} \\ G^{v^*}_{ij} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho & \eta \\ \rho & 1 & \nu \\ \eta & \nu & 1 \end{pmatrix}\right).$$

We further assume, without loss of generality, that $I_0 = J_0 = 1$. Then $Y = \gamma G^{v^*}_{11} + \xi$ and, dropping the subscripts of $\mathcal{R}_{\lambda_r,\lambda_c}$, one indeed has

$$\begin{aligned} \mathcal{R}(k,v) &= \mathbb{E}\left[\left(\gamma G^{v^*}_{11} + \xi - T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right)^2\right] \\ &= \epsilon^2 + \mathbb{E}\left[\left(\gamma G^{v^*}_{11} - T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right)^2\right] \\ &= \epsilon^2 + \gamma^2 - 2\gamma\,\mathbb{E}\left[G^{v^*}_{11} T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right] + \mathbb{E}\left[T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})^2\right]. \end{aligned} \tag{11}$$

where

$$\begin{aligned} T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X}) = \sum_{i=2}^{M} &\operatorname{erf}\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^k_{ij}\right) \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^v_{ij} \\ &+ \operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r\kappa\ell + \lambda_r\gamma G^k_{11}\right)\left(\kappa\ell + \gamma G^k_{11}\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G^k_{1j}\right) G^k_{1j}\right) \cdot \\ &\cdot \left(\operatorname{erf}\left(\lambda_r\kappa\ell + \lambda_r\gamma G^k_{11}\right)\left(\theta\ell + \gamma G^v_{11}\right) + \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G^k_{1j}\right) G^v_{1j}\right) \end{aligned} \tag{12}$$

Since Equation (11) and Equation (12) only depend on the 5 parameters $\kappa, \nu, \rho, \eta, \theta$, this ends the proof of Lemma 2. On the manifold $\mathcal{M}$, where $\rho = \theta = \eta = 0$, we can simplify the expression by dropping the $\theta\ell$ term in the last row of Equation (12). We continue our proof by analyzing each term in Equation (11) individually.

The first term can be rewritten as

$$\begin{aligned} &\mathbb{E}\left[G^{v^*}_{11} T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right] = \\ &= \sum_{i=2}^{M} \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^k_{ij}\right) \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^v_{ij} G^{v^*}_{11}\right] \\ &\quad + \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r\kappa\ell + \lambda_r\gamma G^k_{11}\right)\left(\kappa\ell + \gamma G^k_{11}\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G^k_{1j}\right) G^k_{1j}\right) \cdot \right. \\ &\quad \left. \cdot \left(\operatorname{erf}\left(\lambda_r\kappa\ell + \lambda_r\gamma G^k_{11}\right)\left(\theta\ell + \gamma G^v_{11}\right) G^{v^*}_{11} + \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G^k_{1j}\right) G^v_{1j} G^{v^*}_{11}\right)\right] \end{aligned} \tag{13}$$

On the manifold $\mathcal{M}$, this becomes:

$$\begin{aligned} &\mathbb{E}\left[G^{v^*}_{11} T^{(k,v)}_{\lambda_r,\lambda_c}(\mathbb{X})\right] = \\ &= \sum_{i=2}^{M} \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^k_{ij}\right) \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G^k_{ij}) G^v_{ij} G^{v^*}_{11}\right] \\ &\quad + \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r\kappa\ell + \lambda_r\gamma G^k_{11}\right)\left(\kappa\ell + \gamma G^k_{11}\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G^k_{1j}\right) G^k_{1j}\right) \cdot \right. \end{aligned}$$

$$
\begin{aligned}
&\cdot \left(\operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \gamma G_{11}^v G_{11}^{v^*} + \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^v G_{11}^{v^*} \right) \Bigg] \\
&= \sum_{i=2}^{M} \sum_{j=1}^{N} \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}(\lambda_r G_{ij}^k) G_{ij}^k \right) \operatorname{erf}(\lambda_r G_{ij}^k) G_{ij}^v \right] \underbrace{\mathbb{E}\left[G_{11}^{v^*}\right]}_{=0} \\
&\quad + \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \left(\kappa \ell + \gamma G_{11}^k\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^k \right) \cdot \right. \\
&\qquad \left. \cdot \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \right] \gamma \underbrace{\mathbb{E}\left[G_{11}^v G_{11}^{v^*}\right]}_{=\nu} \\
&\quad + \sum_{j=2}^{N} \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \left(\kappa \ell + \gamma G_{11}^k\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^k \right) \cdot \right. \\
&\qquad \left. \cdot \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^v \right] \underbrace{\mathbb{E}\left[G_{11}^{v^*}\right]}_{=0} \\
&= \gamma \nu \mathbb{E}\left[\operatorname{erf}\left(\lambda_c \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \left(\kappa \ell + \gamma G_{11}^k\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^k \right) \cdot \right. \\
&\qquad \left. \cdot \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \right], \qquad (14)
\end{aligned}
$$

where we have used the independence between $G_{11}^{v^*}$ and G_{ij}^v, G_{ij}^k.

Due to the independence between $G_{ij}^{\cdot}$ and $G_{rs}^{\cdot}$ for $(i,j) \neq (r,s)$, there appear no cross-terms when expanding the expectation of $T_{\lambda_r,\lambda_c}^{(k,v)}(\mathbb{X})^2$. Therefore, under $\rho = 0$, the second term simplifies as

$$
\begin{aligned}
&\mathbb{E}\left[T_{\lambda_r,\lambda_c}^{(k,v)}(\mathbb{X})^2\right] \\
&= \sum_{i=2}^{M} \sum_{j=1}^{N} \mathbb{E}\left[\operatorname{erf}^2\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}\left(\lambda_r G_{ij}^k\right) G_{ij}^k \right) \cdot \operatorname{erf}^2\left(\lambda_r G_{ij}^k\right) \left(G_{ij}^v\right)^2 \right] \\
&\quad + \mathbb{E}\left[\operatorname{erf}^2\left(\lambda_c \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \left(\kappa \ell + \gamma G_{11}^k\right) + \lambda_c \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^k \right) \right. \\
&\qquad \cdot \left(\operatorname{erf}^2\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \left(\theta \ell + \gamma G_{11}^v\right)^2 + \sum_{j=2}^{N} \operatorname{erf}^2\left(\lambda_r G_{1j}^k\right) \left(G_{1j}^v\right)^2 \right. \\
&\qquad \left. \left. + 2\theta \ell \operatorname{erf}\left(\lambda_r \kappa \ell + \lambda_r \gamma G_{11}^k\right) \sum_{j=2}^{N} \operatorname{erf}\left(\lambda_r G_{1j}^k\right) G_{1j}^v \right) \right] \qquad (15)
\end{aligned}
$$

On the manifold $\mathcal{M}$ ($\rho = \theta = \eta = 0$), this becomes

$$
\begin{aligned}
&\mathbb{E}\left[T_{\lambda_r,\lambda_c}^{(k,v)}(\mathbb{X})^2\right] \\
&= \sum_{i=2}^{M} \sum_{j=1}^{N} \mathbb{E}\left[\operatorname{erf}^2\left(\lambda_c \sum_{j=1}^{N} \operatorname{erf}\left(\lambda_r G_{ij}^k\right) G_{ij}^k \right) \cdot \operatorname{erf}^2\left(\lambda_r G_{ij}^k\right) \right]
\end{aligned}
$$

$$+\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_c\mathrm{erf}\left(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\right)\left(\kappa\ell+\gamma G_{11}^k\right)+\lambda_c\sum_{j=2}^{N}\mathrm{erf}\left(\lambda_r G_{1j}^k\right)G_{1j}^k\right)\right.$$
$$\left.\cdot\left(\gamma^2\mathrm{erf}^2\left(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\right)+\sum_{j=2}^{N}\mathrm{erf}^2\left(\lambda_r G_{1j}^k\right)\right)\right]$$

In what follows, we shall prove that under the joint scaling $\lambda_r\ell\to\infty$, $\lambda_c\ell\to\infty$ and $\lambda_r = o\left(M^{-\frac{1}{2}}N^{-\frac{1}{2}}\right)$, the risk converges to ϵ^2. For the oracle parameters, $\kappa=\nu=1$. Moreover, $\mathrm{erf}(\lambda_r x)x\geq 0$ for all $x\in\mathbb{R}$ and $|\mathrm{erf}(x)|\leq 1$. Therefore, Equation (14) implies

$$\begin{aligned}\mathbb{E}\left[G_{11}^{v^*}T_{\lambda_r,\lambda_c}^{(k,v)}(\mathbb{X})\right] &\geq \gamma\mathbb{E}\left[\mathrm{erf}\left(\lambda_c\mathrm{erf}\left(\lambda_r\ell+\lambda_r\gamma G_{11}^k\right)\left(\ell+\gamma G_{11}^k\right)\right)\mathrm{erf}\left(\lambda_r\ell+\lambda_r\gamma G_{11}^k\right)\right]\\ &\quad-\gamma\mathbb{P}\left(\gamma G_{11}^k\leq-\ell\right)\\ &=\gamma\mathbb{E}\left[\mathrm{erf}\left(\frac{\lambda_c}{\lambda_r}\mathrm{erf}(G)G\right)\mathrm{erf}(G)\right]-\gamma\Phi\left(\frac{-\ell}{\gamma}\right),\end{aligned}$$

where $G\sim\mathcal{N}\left(\lambda_r\ell,\lambda_r^2\gamma^2\right)$ and Φ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Next, since $|\mathrm{erf}(x)|\leq 1$ and $\mathrm{erf}^2(x)\leq\frac{4}{\pi}x^2$ from Lemma 12 for all $x\in\mathbb{R}$, one has

$$\begin{aligned}\sum_{i=2}^{M}\sum_{j=1}^{N}\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_c\sum_{j=1}^{N}\mathrm{erf}\left(\lambda_r G_{ij}^k\right)G_{ij}^k\right)\cdot\mathrm{erf}^2\left(\lambda_r G_{ij}^k\right)\right] &\leq \sum_{i=2}^{M}\sum_{j=1}^{N}\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_r G_{ij}^k\right)\right]\\ &\leq\sum_{i=2}^{M}\sum_{j=1}^{N}\frac{4\lambda_r^2}{\pi}\mathbb{E}\left[\left(G_{ij}^k\right)^2\right]\\ &=\frac{4}{\pi}\lambda_r^2(M-1)N\to 0,\end{aligned}$$

due to the scaling $\lambda_r = o\left(M^{-\frac{1}{2}}N^{-\frac{1}{2}}\right)$.

Finally, substituting $\kappa=1$ and using $|\mathrm{erf}(x)|\leq 1$ and $\mathrm{erf}^2(x)\leq\frac{4}{\pi}x^2$ again from Lemma 12, one has that

$$\begin{aligned}&\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_c\mathrm{erf}\left(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\right)\left(\kappa\ell+\gamma G_{11}^k\right)+\lambda_c\sum_{j=2}^{N}\mathrm{erf}\left(\lambda_r G_{1j}^k\right)G_{1j}^k\right)\right.\\ &\left.\cdot\left(\gamma^2\mathrm{erf}^2\left(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\right)+\sum_{j=2}^{N}\mathrm{erf}^2\left(\lambda_r G_{1j}^k\right)\right)\right]\\ &\leq\gamma^2\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_c\mathrm{erf}\left(\lambda_r\ell+\lambda_r\gamma G_{11}^k\right)\left(\ell+\gamma G_{11}^k\right)+\lambda_c\sum_{j=2}^{N}\mathrm{erf}\left(\lambda_r G_{1j}^k\right)G_{1j}^k\right)\right.\\ &\qquad\left.\mathrm{erf}^2\left(\lambda_r\ell+\lambda_r\gamma G_{11}^k\right)\right]+\sum_{j=2}^{N}\mathbb{E}\left[\mathrm{erf}^2\left(\lambda_r G_{1j}^k\right)\right]\\ &\leq\gamma^2\mathbb{E}\left[\mathrm{erf}^2\left(\frac{\lambda_c}{\lambda_r}\mathrm{erf}(G)G+\lambda_c\sum_{j=2}^{N}\mathrm{erf}\left(\lambda_r G_{1j}^k\right)G_{1j}^k\right)\mathrm{erf}^2(G)\right]+\underbrace{\frac{4}{\pi}\lambda_r^2(N-1)}_{\to 0},\end{aligned}$$

where $G\sim\mathcal{N}\left(\lambda_r\ell,\lambda_r^2\gamma^2\right)$.

Aggregating everything and considering the oracle parameters k^*, v^*, one now has that

$$\mathcal{R}(k^*,v^*)\leq\epsilon^2+\gamma^2-2\gamma^2\mathbb{E}\left[\mathrm{erf}\left(\frac{\lambda_c}{\lambda_r}\mathrm{erf}(G)G\right)\mathrm{erf}(G)\right]+2\gamma^2\Phi\left(\frac{-\ell}{\gamma}\right)$$

$$+\gamma^2\mathbb{E}\left[\text{erf}^2\left(\frac{\lambda_c}{\lambda_r}\text{erf}(G)G+\lambda_c\sum_{j=2}^{N}\underbrace{\text{erf}\left(\lambda_r G_{1j}^k\right)G_{1j}^k}_{\geq 0}\right)\text{erf}^2(G)\right]+o(1),$$

where $G\sim\mathcal{N}(\lambda_r\ell,\lambda_r^2\gamma^2)$.

To conclude the proof of Proposition 3, it is enough to prove

$$\mathbb{E}\left[\text{erf}\left(\frac{\lambda_c}{\lambda_r}\text{erf}(G)G\right)\text{erf}(G)\right]\to 1,$$

as $\ell\to\infty$ implies $\Phi\left(\frac{-\ell}{\gamma}\right)\to 0$. Since $x\mapsto\text{erf}\left(\frac{\lambda_c}{\lambda_r}\text{erf}(x)x\right)\text{erf}(x)$ is increasing on $(0,\infty)$ and less than 1 in absolute value, it is enough to consider

$$\begin{aligned}\mathbb{E}\left[\text{erf}\left(\frac{\lambda_c}{\lambda_r}\text{erf}(G)G\right)\text{erf}(G)\right]&\geq\text{erf}\left(\frac{\lambda_c}{\lambda_r}\text{erf}\left(\frac{\lambda_r\ell}{2}\right)\frac{\lambda_r\ell}{2}\right)\text{erf}\left(\frac{\lambda_r\ell}{2}\right)\mathbb{P}\left(G\geq\frac{\lambda_r\ell}{2}\right)\\&\quad-\mathbb{P}(G<0)\\&=\text{erf}\left(\frac{\lambda_c\ell}{2}\text{erf}\left(\frac{\lambda_r\ell}{2}\right)\right)\text{erf}\left(\frac{\lambda_r\ell}{2}\right)\left(1-\Phi\left(-\frac{\ell}{2\gamma}\right)\right)\\&\quad-\Phi\left(-\frac{\ell}{\gamma}\right).\end{aligned}$$

The conclusion now follows since $\lambda_r\ell\to\infty$, $\ell\to\infty$, $\lambda_c\ell\to\infty$.

B.2 Proof of Proposition 4

The proof aims to make explicit the loss of information from $\mathbb{X}$ to $(S_1,S_2,\ldots,S_M)$. To this end, one has

$$\begin{aligned}\mathbb{E}\left[\left(Y-f(S_1,S_2,\ldots,S_M)\right)^2\right]&\geq\mathbb{E}\left[\text{Var}\left(Y|S_1,S_2,\ldots,S_M\right)\right]\\&\geq\mathbb{E}\left[\text{Var}\left(Y|S_1,S_2,\ldots,S_M,I_0,J_0\right)\right].\end{aligned}$$

Now, consider $i_0\in[M]$ and $j_0\in[N]$. Conditional on $(I_0,J_0)=(i_0,j_0)$, S_i is independent of (Y,S_{i_0}), for $i\neq i_0$. Then, one has

$$\text{Var}\left(Y|S_1,S_2,\ldots,S_N,I_0=i_0,J_0=j_0\right)=\text{Var}\left(Y|S_{i_0},I_0=i_0,J_0=j_0\right).\tag{16}$$

For multivariate Gaussian random variables (U,V), one has

$$\text{Var}(U|V)=\Sigma_{UU}-\Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU},\tag{17}$$

where Σ_{AB} stands for $\text{Cov}(A,B)$, for random vectors A and B.

Conditioned on $I_0=i_0$ and $J_0=j_0$, one has:

- $X_{i_0j_0}=\ell k^*+\gamma Z_{j_0}$ and $X_{i_0j}=Z_j$ for $j\in[N]\setminus\{j_0\}$ with standard normal random variables $Z_1,\ldots,Z_N$;
- $S_{i_0}=\beta_{j_0}\ell k^*+\gamma\beta_{j_0}Z_{j_0}+\sum_{j\neq j_0}\beta_jZ_j$;
- $Y=\gamma Z_{j_0}^Tv^*+\xi$.

One then has the following:

- $\text{Var}(Y)=\gamma^2+\epsilon^2$;
- $\text{Cov}(S_{i_0},Y)=\gamma^2\beta_{j_0}v^*$
- $\text{Cov}(S_{i_0},S_{i_0})=\left(\gamma^2\beta_{j_0}^2+\sum_{j\neq j_0}\beta_j^2\right)I$

Introducing in (17), one obtains

$$\text{Var}\left(Y|S_{i_0},I_0=i_0,J_0=j_0\right)=\gamma^2+\epsilon^2-\frac{\gamma^4\beta_{j_0}^2}{\gamma^2\beta_{j_0}^2+\sum_{j\neq j_0}\beta_j^2}\|v^*\|^2=\epsilon^2+\gamma^2\left(1-\frac{\gamma^2\beta_{j_0}^2}{\gamma^2\beta_{j_0}^2+\sum_{j\neq j_0}\beta_j^2}\right).\tag{18}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Var}\left(Y|S_1, S_2, \ldots, S_M, I_0, J_0\right)\right] &= \sum_{j_0=1}^{N}\sum_{i_0=1}^{M} \mathbb{P}(I_0 = i_0|J_0 = j_0)p_{j_0}\mathrm{Var}\left(Y|S_{i_0}, I_0 = i_0, J_0 = j_0\right) \\
&= \sum_{j_0=1}^{N} p_{j_0}\left(\epsilon^2 + \gamma^2\left(1 - \frac{\gamma^2\beta_{j_0}^2}{\gamma^2\beta_{j_0}^2 + \sum_{j\neq j_0}\beta_j^2}\right)\right) \\
&= \epsilon^2 + \gamma^2\left(1 - \sum_{j_0=1}^{N} p_{j_0}\frac{\gamma^2\beta_{j_0}^2}{\gamma^2\beta_{j_0}^2 + \sum_{j\neq j_0}\beta_j^2}\right) \\
&\geq \epsilon^2,
\end{aligned}$$

with equality in the last step only if there exists j_0 such that $p_{j_0} = 1$ and $\beta_j = 0$ for $j \neq j_0$.

C Proof of Theorem 5

In this section, we establish the global convergence of the Projected Gradient Descent (PGD) dynamics to the true oracle parameters. We follow the proof of Theorem 5 in Marion et al. [2025] as closely as possible while adapting it to our model $T^{(k,v)}_{\lambda_r,\lambda_c}$. The only result which requires a completely different and much more involved proof is that of Proposition 9. Throughout these proofs, $C > 0$ denotes a generic constant that may vary from line to line but is strictly independent of the step size α and the iteration step t. Moreover, we shall drop the subscript λ_r, λ_c and assume an implicit dependence on the inverse temperatures.

The theoretical guarantees of our optimization trajectory are established through the following five primary results. We first establish the invariance of the manifold to PGD steps.

Lemma 6. *The manifold $\mathcal{M}$ is invariant under the PGD dynamics (Algorithm 1), in the sense that if $(k_t, v_t) \in \mathcal{M}$, then $(k_{t+1}, v_{t+1}) \in \mathcal{M}$.*

By restricting our analysis to the manifold $\mathcal{M}$, we can project the high-dimensional dynamics down to the scalar alignment variables $\kappa_t = k_t^T k^*$ and $\nu_t = v_t^T v^*$. We then map the PGD updates to an autonomous dynamical system:

Lemma 7. *When initialized on the manifold $\mathcal{M}$, the PGD iterations (Algorithm 1) can be reformulated as the autonomous discrete dynamical system $(\kappa_{t+1}, \nu_{t+1}) = g(\kappa_t, \nu_t)$, where $g\colon [-1,1]^2 \to [-1,1]^2$ is given by*

$$g(\kappa,\nu) = \left(\frac{\kappa - \alpha(\partial_\kappa\mathcal{R}^<(\kappa,\nu))(1-\kappa^2)}{\sqrt{1+\alpha^2(\partial_\kappa\mathcal{R}^<(\kappa,\nu))^2(1-\kappa^2)}}, \frac{\nu - \alpha(\partial_\nu\mathcal{R}^<(\kappa,\nu))(1-\nu^2)}{\sqrt{1+\alpha^2(\partial_\nu\mathcal{R}^<(\kappa,\nu))^2(1-\nu^2)}}\right). \tag{19}$$

With the dynamics formalized, we establish that the risk strictly decreases, forcing convergence to the stationary points:

Proposition 8. *For a sufficiently small step size α, the risk $\mathcal{R}^<$ is strictly decreasing along the PGD iterates unless it is at a fixed point. Furthermore, if (κ, ν) is an accumulation point of the sequence $(\kappa_t, \nu_t)_{t\geq 0}$, it must satisfy:*

$$(1-\kappa^2)\partial_\kappa\mathcal{R}^<(\kappa,\nu) = 0 \quad \textit{and} \quad (1-\nu^2)\partial_\nu\mathcal{R}^<(\kappa,\nu) = 0. \tag{20}$$

We then rigorously classify the geometry of these accumulation points, isolating the oracle states as the unique global minima:

Proposition 9. *For λ_c and λ_r small enough, the fixed points of the dynamics, given by (20), on $[-1,1]^2$ are $(\pm 1, \pm 1)$ and $(0,0)$. Specifically:*

- $(\pm 1, \mp 1)$ *are global maxima.*
- $(\pm 1, \pm 1)$ *are global minima.*
- $(0,0)$ *is a saddle point.*

Finally, we characterize the local geometry around the saddle point to guarantee that random initializations escape it:

Proposition 10. *For a sufficiently small step size α, the mapping g is a local diffeomorphism around $(0,0)$, whose Jacobian matrix has one eigenvalue in $(0,1)$ and one in $(1,\infty)$. Furthermore, g is injective on $[-1,1]^2$.*

The global convergence to the oracle states then follows directly from the exclusion of the saddle point and maxima. The proofs for all stated results are provided in the subsections below.

C.1 Proof of Lemma 6

To prove that the manifold $\mathcal{M}$ is invariant under the Projected Gradient Descent (PGD) dynamics, we must first establish that the gradient of the risk does not "push" the parameters outside of $\mathcal{M}$. We do this by proving an auxiliary symmetry result regarding the cross-term derivatives.

Lemma 11. *At any point $(\kappa,\nu,\theta,\eta,\rho)$ on the manifold $\mathcal{M}$ (i.e., where $\theta=\eta=\rho=0$), the partial derivatives of the risk with respect to the cross-terms vanish:*

$$\partial_\theta \mathcal{R}^< = \partial_\eta \mathcal{R}^< = \partial_\rho \mathcal{R}^< = 0.$$

The proof follows by recalling that the risk can be reduced to a function $\mathcal{R}^<$ depending only on the covariance structure of the Gaussian projections $G^k_{ij} = X^T_{ij}k$, $G^v_{ij} = X^T_{ij}v$, and $G^{v^*}_{ij} = X^T_{ij}v^*$. The relevant covariance matrix is:

$$\begin{pmatrix} G^k_{ij} \\ G^v_{ij} \\ G^{v^*}_{ij} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho & \eta \\ \rho & 1 & \nu \\ \eta & \nu & 1 \end{pmatrix}\right)$$

and the alignment between v and k^* is given by $\theta = v^T k^*$.

It is clear from Equation 10 that the risk is invariant to the sign inversion of (v, v^*) to $(-v, -v^*)$. This amounts to

$$\mathcal{R}^<(\kappa,\nu,\theta,\eta,\rho) = \mathcal{R}^<(\kappa,\nu,-\theta,-\eta,-\rho).$$

Taking the partial derivative with respect to θ now yields the following result:

$$\partial_\theta \mathcal{R}^<(\kappa,\nu,\theta,\eta,\rho) = -\partial_\theta \mathcal{R}^<(\kappa,\nu,-\theta,-\eta,-\rho).$$

Evaluating this at the manifold $\mathcal{M}$ where $\theta = \eta = \rho = 0$, we immediately obtain $\partial_\theta \mathcal{R}^<(\kappa,\nu,0,0,0) = 0$. Identical arguments hold for $\partial_\eta \mathcal{R}^<$ and $\partial_\rho \mathcal{R}^<$, concluding the proof of Lemma 11.

We now return to the proof of Lemma 6. By the chain rule, the Euclidean gradients of the risk with respect to k and v can be expanded in the basis of the parameter vectors:

$$\begin{aligned} \nabla_k \mathcal{R} &= (\partial_\kappa \mathcal{R}^<)k^* + (\partial_\eta \mathcal{R}^<)v^* + (\partial_\rho \mathcal{R}^<)v \\ \nabla_v \mathcal{R} &= (\partial_\nu \mathcal{R}^<)v^* + (\partial_\theta \mathcal{R}^<)k^* + (\partial_\rho \mathcal{R}^<)k \end{aligned}$$

Applying Lemma 11, on the manifold $\mathcal{M}$ (where $\theta=\eta=\rho=0$), the gradients simplify strictly to:

$$\begin{aligned} \nabla_k \mathcal{R}(k_t, v_t) &= \partial_\kappa \mathcal{R}^<(\kappa_t,\nu_t)k^* \\ \nabla_v \mathcal{R}(k_t, v_t) &= \partial_\nu \mathcal{R}^<(\kappa_t,\nu_t)v^* \end{aligned} \tag{21}$$

Recall the PGD updates for (k_{t+1}, v_{t+1}):

$$\begin{aligned} k_{t+1} &= \frac{k_t - \alpha(I_d - k_t k_t^T)\nabla_k \mathcal{R}}{\|k_t - \alpha(I_d - k_t k_t^T)\nabla_k \mathcal{R}\|_2} := \frac{\tilde{k}_{t+1}}{c_k} \\ v_{t+1} &= \frac{v_t - \alpha(I_d - v_t v_t^T)\nabla_v \mathcal{R}}{\|v_t - \alpha(I_d - v_t v_t^T)\nabla_v \mathcal{R}\|_2} := \frac{\tilde{v}_{t+1}}{c_v} \end{aligned}$$

where $c_k, c_v > 0$ are the respective normalization constants. To prove $(k_{t+1}, v_{t+1}) \in \mathcal{M}$, we must verify the three orthogonality conditions: $(v^*)^T k_{t+1} = 0$, $(k^*)^T v_{t+1} = 0$, and $v_{t+1}^T k_{t+1} = 0$.

Condition 1: $(v^*)^T k_{t+1} = 0$**.** Using the simplified gradient (21), we project $\tilde{k}_{t+1}$ onto v^*:

$$(v^*)^T \tilde{k}_{t+1} = (v^*)^T k_t - \alpha \partial_\kappa \mathcal{R}^< \left((v^*)^T k^* - (v^*)^T k_t (k_t^T k^*) \right).$$

Because $(k_t, v_t) \in \mathcal{M}$, we have $(v^*)^T k_t = 0$. By Assumption 1, the oracle parameters are orthogonal, so $(v^*)^T k^* = 0$. Thus:

$$(v^*)^T \tilde{k}_{t+1} = 0 - \alpha \partial_\kappa \mathcal{R}^<(0 - 0) = 0 \implies (v^*)^T k_{t+1} = 0.$$

Condition 2: $(k^*)^T v_{t+1} = 0$**.** By an exactly symmetric argument using $\tilde{v}_{t+1}$ and the fact that $(k^*)^T v_t = 0$, we obtain $(k^*)^T v_{t+1} = 0$.

Condition 3: $v_{t+1}^T k_{t+1} = 0$**.** We compute the inner product of the unnormalized updates $\tilde{v}_{t+1}^T \tilde{k}_{t+1}$:

$$\begin{aligned}
\tilde{v}_{t+1}^T \tilde{k}_{t+1} &= \left(v_t - \alpha \partial_\nu \mathcal{R}^<(v^* - \nu_t v_t) \right)^T \left(k_t - \alpha \partial_\kappa \mathcal{R}^<(k^* - \kappa_t k_t) \right) \\
&= v_t^T k_t - \alpha \partial_\kappa \mathcal{R}^< \left(v_t^T k^* - \kappa_t v_t^T k_t \right) - \alpha \partial_\nu \mathcal{R}^< \left((v^*)^T k_t - \nu_t v_t^T k_t \right) \\
&\quad + \alpha^2 \partial_\kappa \mathcal{R}^< \partial_\nu \mathcal{R}^< \left((v^* - \nu_t v_t)^T (k^* - \kappa_t k_t) \right).
\end{aligned}$$

Since $(k_t, v_t) \in \mathcal{M}$, we already know $v_t^T k_t = 0$, $v_t^T k^* = 0$, and $(v^*)^T k_t = 0$. Thus, all the linear terms vanish. For the quadratic term, we expand:

$$(v^* - \nu_t v_t)^T (k^* - \kappa_t k_t) = (v^*)^T k^* - \kappa_t (v^*)^T k_t - \nu_t v_t^T k^* + \nu_t \kappa_t v_t^T k_t.$$

Substituting the orthogonalities $(v^*)^T k^* = 0$, $(v^*)^T k_t = 0$, $v_t^T k^* = 0$, and $v_t^T k_t = 0$ yields the desired orthogonality.

Therefore, $\tilde{v}_{t+1}^T \tilde{k}_{t+1} = 0$, meaning $v_{t+1}^T k_{t+1} = 0$. Having verified all three conditions, we conclude that $(k_{t+1}, v_{t+1}) \in \mathcal{M}$. q.e.d.

C.2 Proof of Lemma 7

Using the simplified gradient (21) valid on the manifold $\mathcal{M}$, we compute the unnormalized update for k:

$$\tilde{k}_{t+1} = k_t - \alpha (I_d - k_t k_t^T)(\partial_\kappa \mathcal{R}^<) k^* = k_t - \alpha (\partial_\kappa \mathcal{R}^<)(k^* - \kappa_t k_t).$$

Projecting this unnormalized update onto k^* yields the numerator for κ_{t+1}:

$$(k^*)^T \tilde{k}_{t+1} = (k^*)^T k_t - \alpha (\partial_\kappa \mathcal{R}^<)((k^*)^T k^* - \kappa_t (k^*)^T k_t) = \kappa_t - \alpha (\partial_\kappa \mathcal{R}^<)(1 - \kappa_t^2).$$

To find the normalization constant $c_k = \|\tilde{k}_{t+1}\|_2$, we apply the Pythagorean theorem. Since $(k^* - \kappa_t k_t)$ is the projection of k^* onto the orthogonal complement of k_t, it is strictly orthogonal to k_t. Thus:

$$\begin{aligned}
\|\tilde{k}_{t+1}\|_2^2 &= \|k_t\|_2^2 + \alpha^2 (\partial_\kappa \mathcal{R}^<)^2 \|k^* - \kappa_t k_t\|_2^2 \\
&= 1 + \alpha^2 (\partial_\kappa \mathcal{R}^<)^2 (1 - \kappa_t^2).
\end{aligned}$$

Dividing the numerator by $\|\tilde{k}_{t+1}\|_2$ directly yields $g_1(\kappa_t, \nu_t)$. Analogously, one obtains the result for $g_2(\kappa_t, \nu_t)$.

C.3 Proof of Proposition 8

The theoretical risk $\mathcal{R}^<$ is composed of smooth error functions evaluated over Gaussian expectations, ensuring it is $\mathcal{C}^\infty$ on the compact square $[-1, 1]^2$. Consequently, it is Λ-smooth, meaning we can bound its variation:

$$\begin{aligned}
\mathcal{R}^<(\kappa_{t+1}, \nu_{t+1}) - \mathcal{R}^<(\kappa_t, \nu_t) \leq (\partial_\kappa \mathcal{R}^<)(\kappa_{t+1} - \kappa_t) + (\partial_\nu \mathcal{R}^<)(\nu_{t+1} - \nu_t) \\
+ \frac{\Lambda}{2} \left[(\kappa_{t+1} - \kappa_t)^2 + (\nu_{t+1} - \nu_t)^2 \right].
\end{aligned} \tag{22}$$

Using the discrete mapping g from Lemma 7, we rewrite the step difference for κ:

$$\kappa_{t+1} - \kappa_t = \kappa_t \left(\frac{1}{\sqrt{1 + \alpha^2 (\partial_\kappa \mathcal{R}^<)^2 (1 - \kappa_t^2)}} - 1 \right) - \frac{\alpha (\partial_\kappa \mathcal{R}^<)(1 - \kappa_t^2)}{\sqrt{1 + \alpha^2 (\partial_\kappa \mathcal{R}^<)^2 (1 - \kappa_t^2)}}$$

$$= -\alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2) + \left(\frac{1}{\sqrt{1+\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2)}} - 1\right)(\kappa_t - \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)). \tag{23}$$

As a consequence,

$$\begin{aligned}|\kappa_{t+1} - \kappa_t + \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)| &\leq \left|\frac{1}{\sqrt{1+\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2)}} - 1\right| \left|\kappa_t - \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)\right| \\ &\leq \alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2)\left|\kappa_t - \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)\right| \\ &\leq C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2), \end{aligned} \tag{24}$$

where the second inequality holds by Lemma 14 and the last bound holds since the function $(\kappa, \nu) \mapsto \left|\kappa - \alpha(\partial_\kappa \mathcal{R}^<(\kappa,\nu))(1-\kappa^2)\right|$ is uniformly bounded for all $\alpha \leq 1$. This bound has two implications. First,

$$\begin{aligned}(\partial_\kappa \mathcal{R}^<)(\kappa_{t+1} - \kappa_t) + \alpha(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) &= (\partial_\kappa \mathcal{R}^<)\left((\kappa_{t+1} - \kappa_t) + \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)\right) \\ &\leq |(\partial_\kappa \mathcal{R}^<)|\left|(\kappa_{t+1} - \kappa_t) + \alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)\right| \\ &\leq C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2), \end{aligned} \tag{25}$$

where we use the fact that $|\partial_\kappa \mathcal{R}^<|$ is bounded and the bound (24). Second, since the square function is Lipschitz on compact sets, we have

$$\left|(\kappa_{t+1} - \kappa_t)^2 - \left(\alpha(\partial_\kappa \mathcal{R}^<)(1-\kappa_t^2)\right)^2\right| \leq C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2).$$

Thus,

$$\begin{aligned}(\kappa_{t+1} - \kappa_t)^2 &\leq \alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2)^2 + C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) \\ &\leq C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2). \end{aligned} \tag{26}$$

We obtain analogous bounds for ν, namely

$$(\partial_\nu \mathcal{R}^<)(\nu_{t+1} - \nu_t) + \alpha(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2) \leq C\alpha^2(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2) \tag{27}$$

and

$$(\nu_{t+1} - \nu_t)^2 \leq C\alpha^2(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2). \tag{28}$$

Plugging the bounds (25), (26), (27), (28) into Equation (22), we obtain the desired inequality

$$\begin{aligned}\mathcal{R}^<(\kappa_{t+1}, \nu_{t+1}) - \mathcal{R}^<(\kappa_t, \nu_t) \leq &-\alpha(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) - \alpha(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2) \\ &+ C\alpha^2(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) + C\alpha^2(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2).\end{aligned}$$

By choosing the step size $\alpha \leq \frac{1}{2C}$, this ensures that

$$\mathcal{R}^<(\kappa_{t+1}, \nu_{t+1}) - \mathcal{R}^<(\kappa_t, \nu_t) \leq -\frac{\alpha}{2}(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) - \frac{\alpha}{2}(\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2).$$

This shows that the risk is decreasing along the PGD iterates. Next, introducing $\mathcal{R}^<_{\min} = \min_{(\kappa,\nu)\in[-1,1]^2} \mathcal{R}^<(\kappa,\nu)$ and using a telescopic sum, we have, for all $T \geq 0$,

$$\begin{aligned}\mathcal{R}^<(\kappa_0, \nu_0) - \mathcal{R}^<_{\min} &\geq \mathcal{R}^<(\kappa_0, \nu_0) - \mathcal{R}^<(\kappa_T, \nu_T) \\ &\geq \frac{\alpha}{2}\sum_{t=0}^{T-1}(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) + (\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2).\end{aligned}$$

Since the left-hand side is finite, and the terms of the sum are nonnegative, we conclude that the series converges as $T \to \infty$. In particular, the generic term $(\partial_\kappa \mathcal{R}^<)^2(1-\kappa_t^2) + (\partial_\nu \mathcal{R}^<)^2(1-\nu_t^2)$ of the series converges to 0 as $t \to \infty$. Therefore, the accumulation points $(\kappa_\infty, \nu_\infty)$ satisfy

$$\begin{cases}\partial_\kappa \mathcal{R}^<(\kappa_\infty, \nu_\infty) = 0 \quad \text{or} \quad \kappa_\infty^2 = 1 \\ \partial_\nu \mathcal{R}^<(\kappa_\infty, \nu_\infty) = 0 \quad \text{or} \quad \nu_\infty^2 = 1.\end{cases}$$

Inspecting identity (23), we observe that the convergence of the general term also implies $\kappa_{t+1} - \kappa_t \to 0$. Similarly, $\nu_{t+1} - \nu_t \to 0$. q.e.d.

C.4 Proof of Proposition 9

First we recall the formula of the risk

$$\begin{aligned}
\mathcal{R}^<(\kappa,\nu) =& \\
&= -2\gamma^2\nu\mathbb{E}\Bigg[\operatorname{erf}\Bigg(\lambda_c\operatorname{erf}\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\,(\kappa\ell+\gamma G_{11}^k)+\lambda_c\sum_{j=2}^N\operatorname{erf}\big(\lambda_r G_{1j}^k\big)\,G_{1j}^k\Bigg)\cdot \\
&\qquad\cdot\operatorname{erf}\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\Bigg] \\
&\quad+\gamma^2\mathbb{E}\Bigg[\operatorname{erf}^2\Bigg(\lambda_c\operatorname{erf}\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\big(\kappa\ell+\gamma G_{11}^k\big)+\lambda_c\sum_{j=2}^N\operatorname{erf}\big(\lambda_r G_{1j}^k\big)\,G_{1j}^k\Bigg) \\
&\qquad\cdot\operatorname{erf}^2\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\Bigg] \\
&\quad+\mathbb{E}\Bigg[\operatorname{erf}^2\Bigg(\lambda_c\operatorname{erf}\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\big(\kappa\ell+\gamma G_{11}^k\big)+\lambda_c\sum_{j=2}^N\operatorname{erf}\big(\lambda_r G_{1j}^k\big)\,G_{1j}^k\Bigg) \\
&\qquad\cdot\sum_{j=2}^N\operatorname{erf}^2\big(\lambda_r G_{1j}^k\big)\Bigg]+C,
\end{aligned}\tag{29}$$

where C does not depend on κ and ν.

$$V(x_1,x_2,\dots,x_N)\coloneqq\operatorname{erf}\Bigg(\frac{\lambda_c}{\lambda_r}\sum_{j=1}^N\operatorname{erf}(x_j)\,x_j\Bigg)\tag{30}$$

and by

$$U(x_1,x_2,\dots,x_N)\coloneqq\ V(x_1,\dots,x_N)\operatorname{erf}(x_1)\tag{31}$$

Furthermore, we shall make a slight abuse of notation and denote by

$$\begin{aligned}
V(\kappa):=\ & V\left(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k,\lambda_r G_{12}^k,\dots,\lambda_r G_{1N}^k\right)\\
&=\operatorname{erf}\Bigg(\lambda_c\operatorname{erf}\big(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k\big)\,\big(\kappa\ell+\gamma G_{11}^k\big)+\lambda_c\sum_{j=2}^N\operatorname{erf}\big(\lambda_r G_{1j}^k\big)\,G_{1j}^k\Bigg)
\end{aligned}$$

and by

$$U(\kappa)\coloneqq\ V(\kappa)\cdot\operatorname{erf}(\lambda_r\kappa\ell+\lambda_r\gamma G_{11}^k).$$

One can easily observe that $V(\kappa)$ is symmetric in all G_{1j}^k and in κ since $x\operatorname{erf}(x)$ is symmetric around 0 and that $|V(\kappa)|\le 1$. Moreover, this implies that $U(\kappa)$ is strictly increasing in κ and that $|U(\kappa)|\le 1$. Finally, the risk can be rewritten as

$$\mathcal{R}^<(\kappa,\nu)=-2\gamma^2\nu\mathbb{E}\left[U(\kappa)\right]+\gamma^2\mathbb{E}\left[U(\kappa)^2\right]+\mathbb{E}\left[V(\kappa)^2\sum_{j=2}^N\operatorname{erf}^2\big(\lambda_r G_{1j}^k\big)\right]+C.$$

Now, (κ,ν) is a stationary point of the PGD dynamics if

$$\begin{cases}\partial_\kappa\mathcal{R}^<(\kappa,\nu)=0 \quad\text{or}\quad \kappa^2=1\\ \partial_\nu\mathcal{R}^<(\kappa,\nu)=0 \quad\text{or}\quad \nu^2=1.\end{cases}$$

The partial derivatives of the risk are

$$\begin{cases}\partial_\kappa\mathcal{R}^<(\kappa,\nu)=2\gamma^2\mathbb{E}\left[\partial_\kappa U(\kappa)(U(\kappa)-\nu)\right]+2\mathbb{E}\left[\partial_\kappa V(\kappa)V(\kappa)\sum_{j=2}^N\operatorname{erf}^2\big(\lambda_r G_{1j}^k\big)\right]\\ \partial_\nu\mathcal{R}^<(\kappa,\nu)=-2\gamma^2\mathbb{E}\left[U(\kappa)\right].\end{cases}\tag{32}$$

We split our analysis in different cases.

Case 1. $\partial_\nu \mathcal{R}^<(\kappa,\nu)=0$ This implies that $\mathbb{E}[U(\kappa)]=0$. It is clear that $\mathbb{E}[U(0)]=0$ and due to the fact that U is strictly increasing, $\kappa=0$ is also the unique such point.

Now,

$$S(\kappa) \coloneqq \gamma^2 \mathbb{E}\left[U(\kappa)^2\right] + \mathbb{E}\left[V(\kappa)^2 \sum_{j=2}^{N} \text{erf}^2\left(\lambda_r G_{1j}^k\right)\right] \tag{33}$$

is jointly symmetric in $\left(\kappa, \left(G_{1j}^k\right)_{j\in[N]}\right)$. Taking the expectation over $\left(G_{1j}^k\right)_{j\in[N]}$, we obtain that the term in (33) is symmetric in κ, so its derivative must satisfy $\partial_\kappa S(0)=0$. This implies that

$$0=\partial_\kappa \mathcal{R}^<(0,\nu) = -2\gamma^2\nu \mathbb{E}\left[\partial_\kappa U(0)\right]$$

But U is strictly increasing, which implies that $\nu=0$.

Therefore, the only critical point of $\mathcal{R}^<$ inside of $[-1,1]^2$ is $(0,0)$. Moreover, since the Hessian

$$\mathcal{H}_{\mathcal{R}^<}(0,0) = \begin{pmatrix} \partial_{\kappa\kappa}\mathcal{R}^< & -2\gamma^2\mathbb{E}\left[\partial_\kappa U(0)\right] \\ -2\gamma^2\mathbb{E}\left[\partial_\kappa U(0)\right] & 0 \end{pmatrix},$$

we obtain that $(0,0)$ is a saddle point.

Case 2. $\nu=-1$ It is clear that $\partial_\kappa \mathcal{R}^<(0,-1)>0$, since $\mathbb{E}[U(\kappa)^2]$ and $\mathbb{E}[V(\kappa)^2]$ are even functions and $\mathbb{E}[U(\kappa)]$ is an increasing function in κ. Moreover, on $\kappa\in[0,1]$, the function $\mathcal{R}^<(\kappa,-1)$ is clearly increasing in κ since $\mathbb{E}[U(\kappa)]$ and $\mathbb{E}[V^2(\kappa)]$ are.

Let us now denote by $Z_1 \coloneqq \kappa\ell + \gamma G_{11}^k$ and by $Z_j \coloneqq G_{1j}^k$, for $j\in\{2,3,\dots,N\}$. Then, one has

$$\begin{cases} \partial_\kappa U(\kappa) = \partial_\kappa U\left(\lambda_r\kappa\ell + \lambda_r\gamma G_{11}^k, \lambda_r G_{12}^k, \dots, \lambda_r G_{1N}^k\right) = \lambda_r \ell\, \partial_1 U(\lambda_r Z_1, \dots, \lambda_r Z_N), \\ \partial_\kappa V(\kappa) = \partial_\kappa V\left(\lambda_r\kappa\ell + \lambda_r\gamma G_{11}^k, \lambda_r G_{12}^k, \dots, \lambda_r G_{1N}^k\right) = \lambda_r \ell\, \partial_1 V(\lambda_r Z_1, \dots, \lambda_r Z_N). \end{cases} \tag{34}$$

By directly applying the chain rule to the expectation, the formula for the partial derivative of the risk becomes

$$\begin{aligned} \partial_\kappa \mathcal{R}^<(\kappa,-1) &= \partial_\kappa \mathbb{E}\left[\gamma^2\left(U(\kappa)^2+2U(\kappa)\right) + V(\kappa)^2\sum_{j=2}^{N}\text{erf}^2(\lambda_r Z_j)\right] \\ &= \lambda_r\ell\,\mathbb{E}\left[\gamma^2\partial_1\left(U(\lambda_r Z)^2+2U(\lambda_r Z)\right) + \partial_1\left(V(\lambda_r Z)^2\right)\sum_{j=2}^{N}\text{erf}^2(\lambda_r Z_j)\right] \\ &= 2\lambda_r\ell\,\mathbb{E}\left[\gamma^2\partial_1 U(\lambda_r Z)\left(1+U(\lambda_r Z)\right) + V(\lambda_r Z)\partial_1 V(\lambda_r Z)\sum_{j=2}^{N}\text{erf}^2(\lambda_r Z_j)\right], \end{aligned} \tag{35}$$

where $U(\lambda_r Z)$ stands for $U(\lambda_r Z_1,\dots,\lambda_r Z_N)$ and the same for the other functions.

We will now show that if $\lambda_c=\tau\lambda_r$, for λ_r small enough and the other hyperparameters fixed, the expectation above is strictly positive for all $\kappa\in[-1,1]$. We shall make use of the Taylor expansions

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}x + \frac{\text{erf}'''(\xi_x)}{6}x^3 = \frac{2}{\sqrt{\pi}}x + \mathcal{O}(x^3),$$

for some $\xi_x\in[0,x]$, and

$$e^{-x} = 1 - e^{-\zeta_x}x = 1+\mathcal{O}(x),$$

for some $\zeta_x\in[0,x]$. Note that $|\text{erf}'''(x)|<2.5$, for any $x\in\mathbb{R}$ and $|e^{-\zeta_x}|\le 1$ for any $x\in[0,\infty)$. This in turn implies the following asymptotic scalings

$$V(\lambda_r Z) = \text{erf}\left(\tau\lambda_r\sum_{i=1}^{N} Z_i \text{erf}(\lambda_r Z_i)\right) = \frac{4\tau}{\pi}\lambda_r^2\sum_{i=1}^{N} Z_i^2 + \mathcal{O}(\lambda_r^3),$$

$$
\begin{aligned}
U(\lambda_r Z) = V(\lambda_r Z)\text{erf}(\lambda_r Z_1) &= \frac{8\tau}{\pi\sqrt{\pi}}\lambda_r^3 Z_1 \sum_{i=1}^{N} Z_i^2 + \mathcal{O}(\lambda_r^4), \\
\partial_1 V(\lambda_r Z) &= \frac{2\tau}{\sqrt{\pi}} e^{-\tau^2\left(\sum_{i=1}^{N}\lambda_r Z_i \text{erf}(\lambda_r Z_i)\right)^2}\left(\frac{2}{\sqrt{\pi}}\lambda_r Z_1 e^{-\lambda_r^2 Z_1^2} + \text{erf}(\lambda_r Z_1)\right) \\
&= \frac{2\tau}{\sqrt{\pi}}\left(1-\mathcal{O}(\lambda_r^2)\right)\left(\frac{2}{\sqrt{\pi}}\lambda_r Z_1(1-\mathcal{O}(\lambda_r^2)) + \frac{2}{\sqrt{\pi}}\lambda_r Z_1\right) \\
&= \frac{8\tau}{\pi}\lambda_r Z_1 + \mathcal{O}(\lambda_r^3), \\
\partial_1 U(\lambda_r Z) &= \frac{2}{\sqrt{\pi}} e^{-\lambda_r^2 Z_1^2} V(\lambda_r Z) + \partial_1 V(\lambda_r Z)\text{erf}(\lambda_r Z_1) \\
&= \frac{8\tau}{\pi\sqrt{\pi}}\lambda_r^2 \sum_{i=1}^{N} Z_i^2 + \frac{16\tau}{\pi\sqrt{\pi}}\lambda_r^2 Z_1^2 + \mathcal{O}(\lambda_r^3) \\
&= \frac{8\tau}{\pi\sqrt{\pi}}\lambda_r^2\left(\sum_{i=1}^{N} Z_i^2 + 2Z_1^2\right) + \mathcal{O}(\lambda_r^3). \qquad (36)
\end{aligned}
$$

From these computations, it is also clear that every term $\mathcal{O}(\lambda_r^n) = C(Z_1, Z_2, \ldots, Z_N)\lambda_r^n$, where $\mathbb{E}[|C(Z_1, \ldots, Z_N)|] < \infty$.

Finally, (35) becomes

$$
\partial_\kappa \mathcal{R}^<(\kappa, -1) = \frac{16\tau\ell\gamma^2}{\pi\sqrt{\pi}}\lambda_r^3 \mathbb{E}\left[\sum_{i=1}^{N} Z_i^2 + 2Z_1^2\right] + \mathcal{O}(\lambda_r^4).
$$

Since the coefficient of the leading term is positive, this concludes the proof that there exists λ_r small enough such that $\partial_\kappa \mathcal{R}^<(\kappa, -1) > 0$ for all $\kappa \in [-1, 1]$.

Case 3. $\nu = 1$ By leveraging the global symmetry of the theoretical risk function, specifically $\mathcal{R}^<(-\kappa, -\nu) = \mathcal{R}^<(\kappa, \nu)$, an analogous argument applies along the opposite boundary $\nu = 1$. The identical asymptotic smoothness guarantees $\partial_\kappa \mathcal{R}^<(\kappa, 1) < 0$ for all $\kappa \in [-1, 1]$

Finally combining the results of the three cases, one obtains the conclusion.

C.5 Proof of Proposition 10

We verify the claims of the proposition sequentially.

Local diffeomorphism and eigenvalues at the origin. Let $x = (\kappa, \nu)$. Because the gradient of the risk vanishes at the origin ($\partial_\kappa \mathcal{R}^<(0, 0) = 0$) and $\mathcal{R}^<$ is smooth, the squared partial derivative satisfies $(\partial_\kappa \mathcal{R}^<(\kappa, \nu))^2 = \mathcal{O}(\|x\|_2^2)$. This allows us to write the Taylor expansion for the first component of g, denoted $g_1(\kappa, \nu)$:

$$
\begin{aligned}
g_1(\kappa, \nu) &= \frac{\kappa - \alpha(\partial_\kappa \mathcal{R}^<(\kappa, \nu))(1-\kappa^2)}{\sqrt{1 + \alpha^2(\partial_\kappa \mathcal{R}^<(\kappa, \nu))^2(1-\kappa^2)}} \\
&= \frac{\kappa - \alpha(\partial_\kappa \mathcal{R}^<(\kappa, \nu))(1-\kappa^2)}{\sqrt{1 + \mathcal{O}(\|x\|_2^2)}} \\
&= (\kappa - \alpha(\partial_\kappa \mathcal{R}^<(\kappa, \nu))(1-\kappa^2))(1 + \mathcal{O}(\|x\|_2^2)) \\
&= \kappa - \alpha\partial_\kappa \mathcal{R}^<(\kappa, \nu) + \mathcal{O}(\|x\|_2^2).
\end{aligned}
$$

Applying the same expansion to the second component $g_2(\kappa, \nu)$ yields the Jacobian matrix of g evaluated at the origin, $J_g(0, 0) = I_2 - \alpha H_{\mathcal{R}^<}(0, 0)$. Because $\det(H_{\mathcal{R}^<}(0, 0)) = -4\gamma^4 \left(\mathbb{E}[\partial_\kappa U(0)]\right)^2$, from the proof of Proposition 9, the eigenvalues of $H_{\mathcal{R}^<}(0, 0)$ have strictly opposite signs. Consequently, for a sufficiently small step size $\alpha > 0$, $J_g(0, 0)$ possesses one eigenvalue strictly inside the interval $(0, 1)$ and another strictly greater than 1. This confirms that $J_g(0, 0)$ is invertible, which by the Inverse Function Theorem ensures g is a local diffeomorphism around $(0, 0)$.

Differentiability and non-degenerate Jacobian on $[-1,1]^2$. The mapping g is globally differentiable on $[-1,1]^2$ as it is a composition of smooth elementary functions. To prove its Jacobian is nowhere degenerate, we express the mapping as $g(x) = x + \alpha h(x)$. The first coordinate of h takes the form:

$$h_1(\kappa,\nu) = \frac{1}{\alpha}(g_1(\kappa,\nu) - \kappa)$$
$$= \underbrace{\frac{\kappa}{\alpha}\left(\frac{1}{\sqrt{1+\alpha^2(\partial_\kappa \mathcal{R}^<(\kappa,\nu))^2(1-\kappa^2)}} - 1\right)}_{\text{Term 1}} - \underbrace{\frac{\partial_\kappa \mathcal{R}^<(\kappa,\nu)(1-\kappa^2)}{\sqrt{1+\alpha^2(\partial_\kappa \mathcal{R}^<(\kappa,\nu))^2(1-\kappa^2)}}}_{\text{Term 2}}.$$

We must prove that the gradient of $h_1(\kappa,\nu)$ remains uniformly bounded for all $\alpha \le 1$. Term 2 is a composition of functions with uniformly bounded gradients on the compact domain, so its gradient is trivially bounded. For Term 1, we introduce the auxiliary function $q(a,b) = \frac{a}{\alpha}\left(\frac{1}{\sqrt{1+\alpha^2 b}} - 1\right)$ and define the supremum $B = \sup_{[-1,1]^2}(\partial_\kappa \mathcal{R}^<(\kappa,\nu))^2(1-\kappa^2)$. Term 1 can be expressed as $q(\kappa, (\partial_\kappa \mathcal{R}^<)^2(1-\kappa^2))$. Since the inner function is smooth and independent of α, we only need to bound the partial derivatives of q. By Lemma 14, for any $\alpha \le 1$:

$$|\partial_a q(a,b)| = \frac{1}{\alpha}\left|\frac{1}{\sqrt{1+\alpha^2 b}} - 1\right| \le \alpha b \le B,$$
$$|\partial_b q(a,b)| = \left|-\frac{\alpha a}{2(1+\alpha^2 b)^{3/2}}\right| \le \frac{\alpha}{2} \le \frac{1}{2}.$$

Since both partial derivatives are uniformly bounded, the gradient of h_1 is uniformly bounded. A symmetric argument applies to h_2, confirming that the Jacobian matrix $J_h(\kappa,\nu)$ is uniformly bounded over $\alpha \le 1$. Because $J_g(\kappa,\nu) = I_2 + \alpha J_h(\kappa,\nu)$, selecting a sufficiently small α ensures the eigenvalues of J_g remain strictly positive, guaranteeing the Jacobian is nowhere degenerate.

Injectivity. The uniform bound on J_h implies that h is B_L-Lipschitz continuous on $[-1,1]^2$ for some constant $B_L > 0$ independent of α (for small α). Choose $\alpha < 1/B_L$. Suppose there exist two distinct points $x, y \in [-1,1]^2$ such that $g(x) = g(y)$. This implies:

$$x + \alpha h(x) = y + \alpha h(y) \implies x - y = \alpha(h(y) - h(x)).$$

Taking the Euclidean norm of both sides yields:

$$\|x - y\|_2 = \alpha \|h(y) - h(x)\|_2 \le \alpha B_L \|x - y\|_2.$$

Because $\alpha B_L < 1$, this inequality forces $\|x-y\|_2 < \|x-y\|_2$, which is a strict contradiction. Therefore, g must be injective on the domain $[-1,1]^2$.

C.6 Final details.

From Proposition 8, the distance between successive iterates $\|(\kappa_{t+1},\nu_{t+1}) - (\kappa_t,\nu_t)\|_2$ tends to zero. Because the sequence $(\kappa_t,\nu_t)_{t\ge 0}$ lies within the compact set $[-1,1]^2$, the set of its accumulation points must be connected. Given that the set of available fixed points is finite and isolated $\{(0,0),(\pm 1,\pm 1),(\pm 1,\mp 1)\}$, the sequence cannot oscillate and must uniquely converge to exactly one of them.

It remains to be shown that the probability of converging to the saddle point $(0,0)$ or the maxima $(\pm 1,\mp 1)$ is zero. Because the risk strictly decreases at every step (Proposition 8), convergence to the global maxima is impossible unless initialized exactly upon them.

To systematically rule out the saddle point $(0,0)$, we apply the Center-Stable Manifold Theorem [Shub, 2013, Theorem III.7]. Because $J_g(0,0)$ has exactly one eigenvalue with magnitude strictly greater than 1 (Proposition 10), the local center-stable manifold W^{cs}_{loc} associated with the origin has codimension 1. In the 2-dimensional space $[-1,1]^2$, a codimension-1 manifold is a curve, and its Lebesgue measure is strictly zero.

The stable set W^s, consisting of all initializations that converge to $(0,0)$ under iterations of g, is contained within the countable union of pre-images of W^{cs}_{loc}:

$$W^s \subset \bigcup_{T\ge 0} g^{-T}(W^{cs}_{loc}). \tag{37}$$

Because g is injective and differentiable, its inverse preserves sets of measure zero. The countable union of measure-zero sets is also of measure zero.

Therefore, the set of all initializations on $\mathcal{M}$ that converge to the saddle point $(0,0)$ has Lebesgue measure zero. We conclude that for a generic random initialization $(k_0, v_0) \in \mathcal{M}$, the PGD sequence almost surely bypasses the saddle point and the maxima, converging strictly to the true oracle structure $(\pm 1, \pm 1)$.

D Technical lemmas

In this section we list the technical lemmas used throughout the proofs in the Appendix.

Lemma 12. *For $u \in \mathbb{R}$,*

$$\mathit{erf}^2(u) \leq \frac{4}{\pi} u^2.$$

Proof. The proof relies on the concavity of erf on $(0, \infty)$. Since erf is an odd function, it is enough to prove the statement for $u \geq 0$. Since erf is concave,

$$\text{erf}(u) \leq \text{erf}(0) + \text{erf}'(0)u = \frac{2}{\sqrt{\pi}} u,$$

which implies the conclusion.

Lemma 13. *For $G \sim \mathcal{N}(0, \sigma^2)$ and $t \in \mathbb{R}$,*

$$\mathbb{E}[\text{erf}(t + G)] = \text{erf}\left(\frac{t}{\sqrt{1 + 2\sigma^2}}\right).$$

Proof. One has

$$\begin{aligned}
\mathbb{E}[\text{erf}(t+G)] &= \int_0^t \mathbb{E}[\text{erf}'(s+G)]\mathrm{d}s \\
&= \int_0^t \mathbb{E}\left[\frac{2}{\sqrt{\pi}} e^{-(s+G)^2}\right] \mathrm{d}s \\
&= \int_0^t \int_{-\infty}^{\infty} \frac{\sqrt{2}}{\pi\sigma} e^{-(s+g)^2} e^{-\frac{g^2}{2\sigma^2}} \mathrm{d}g \, \mathrm{d}s \\
&= \frac{\sqrt{2}}{\pi\sigma} \int_0^t \int_{-\infty}^{\infty} e^{-g^2 \frac{1+2\sigma^2}{2\sigma^2} - 2gs - s^2} \mathrm{d}g \, \mathrm{d}s \\
&= \frac{\sqrt{2}}{\pi\sigma} \int_0^t \int_{-\infty}^{\infty} e^{-\left(g\sqrt{\frac{1+2\sigma^2}{2\sigma^2}} + s\sqrt{\frac{2\sigma^2}{1+2\sigma^2}}\right)^2} e^{-\frac{1}{1+2\sigma^2} s^2} \mathrm{d}g \, \mathrm{d}s \\
&= \frac{\sqrt{2}}{\pi\sigma} \int_0^t \sqrt{\pi} \sqrt{\frac{2\sigma^2}{1+2\sigma^2}} e^{-\frac{1}{1+2\sigma^2} s^2} \mathrm{d}s \\
&= \frac{2}{\sqrt{\pi}} \int_0^{\frac{t}{\sqrt{1+2\sigma^2}}} e^{-u^2} \mathrm{d}u \\
&= \text{erf}\left(\frac{t}{\sqrt{1+2\sigma^2}}\right).
\end{aligned}$$

Lemma 14. *For $u \geq 0$,*

$$\left| \frac{1}{\sqrt{1+u}} - 1 \right| \leq u.$$

Proof. For $u \geq 0$, the term inside the absolute value is at most 0, so we need to prove that

$$u + \frac{1}{\sqrt{1+u}} \geq 1.$$

This statement is obvious for $u \geq 1$. For $u \in [0, 1)$, it is equivalent to

$$\frac{1}{\sqrt{1+u}} \geq 1 - u \iff \frac{1}{1+u} \geq (1-u)^2 \iff 1 \geq (1-u^2)(1-u),$$

which is clearly true since $u \in [0, 1)$. One has equality only when $u = 0$.

E Additional Experimental Details and Results

In this section, we provide supplementary empirical results that expand upon the optimization dynamics discussed in Section 4. Specifically, we analyze the training dynamics of the predictors without manifold penalties in Section E.2, provide the full results for the experiments in Figure 4 in Section E.3 and further investigate the optimization landscape for decoupled λ_c and λ_r in Section E.4, and for different values of ℓ and N in Section E.5.

E.1 Implementation details and hyperparameters

Code availability. The complete codebase used to run all numerical simulations and generate the figures presented in this work can be found at `https://github.com/anditp/Single-Location-Regression.git`. All the numerical experiments were run on a MacBook Pro M1 on CPU. Each individual experiment (20 runs) in the paper takes less than an hour to run.

Distance to the invariant manifold. Throughout our experiments, we track the optimization trajectories to determine if they remain on, or converge to, the invariant manifold $\mathcal{M}$. Recall that $\mathcal{M}$ is defined by the strict orthogonality constraints:

$$\mathcal{M} = \left\{ (k, v) \in \left(\mathbb{S}^{d-1}\right)^2 \;\middle|\; k^T v^* = 0,\ v^T k^* = 0,\ k^T v = 0 \right\}.$$

Consequently, for any parameter pair $(k, v) \in \left(\mathbb{S}^{d-1}\right)^2$, we compute its distance to $\mathcal{M}$ continuously during training using the root sum of squares of the corresponding inner products:

$$d_{\mathcal{M}}(k, v) := \sqrt{(k^T v^*)^2 + (v^T k^*)^2 + (k^T v)^2}. \tag{38}$$

Hyperparameters. Table 1 summarizes the specific values and schedules used across all numerical experiments presented in Section 4 and Appendix E. For the penalized empirical risk experiments, we introduced the weighting coefficients μ_1 and μ_2 to enforce the structural constraints.

Table 1: Summary of hyperparameters used across the numerical experiments.

Parameter	**Fig. 3** ($\mathcal{M}$ initialization)	**Fig. 4**, **Fig. 13** (Constant λs)	**Fig. 5** (Penalized)	**Fig. 6** (Unpenalized)
Grid height (M)	5	5	5	5
Grid width (N)	5	5	5	5
Embedding dim. (d)	20	20	20	20
Shift magnitude (ℓ)	100	100	100	100
Special token variance (γ^2)	1	1	1	1
Noise variance (ϵ^2)	0.1	0.1	0.1	0.1
Penalty weight (μ_1)	0	0	10	0
Penalty weight (μ_2)	0	0	1	0
Learning rate (α)	1e-4	1e-4	1e-4 (a) / 1e-3 (b)	1e-4 (a) / 1e-3 (b)
Number of iterations	100,000	100,000	100,000	100,000
Batch size	256	256	256	256
Independent runs	20	20 / λ value	20	20

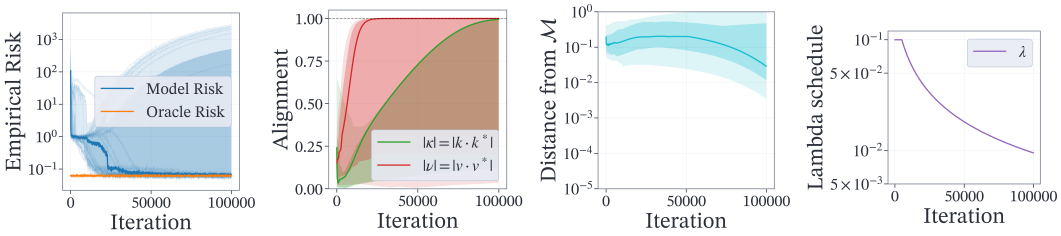
The batches for Stochastic PGD are generated dynamically and resampled at each step, since we are aiming to mimic PGD on the theoretical risk.

E.2 Optimization without Manifold Penalties

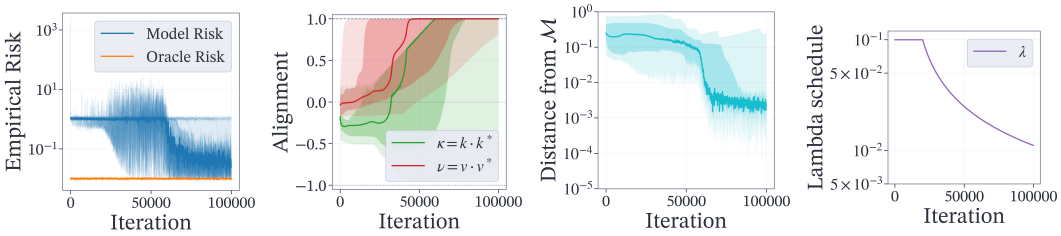
In Section 4, we demonstrated that optimizing the penalized empirical risk allows both the erf and softmax predictors to converge to the oracle parameters from generic random initializations. To isolate the effect of the penalty terms, we conduct an ablation study where both predictors are trained using the strictly unpenalized risk from generic initializations outside of $\mathcal{M}$.

Figure 6 presents the optimization trajectories for both $T^{(k,v)}_{\lambda_r,\lambda_c}$ (erf predictor) and $T^{(k,v)}_{\lambda_r,\lambda_c,\text{softmax}}$ (softmax predictor) under these unconstrained conditions. Without the explicit geometric regularization provided by the penalties, the trajectories fail to always converge. For the erf predictor there

are runs which get trapped in high loss regions. On the other hand, the optimization of the softmax predictor appears better behaved, with most of the runs converging in a similar manner as for the penalized version; however, there are some runs that get trapped in regions with high positive ν and high negative κ, far from $\mathcal{M}$. This empirical finding underscores the practical necessity of the regularization terms when operating outside the theoretical guarantees of Theorem 5.



(a) Dynamics for the erf predictor.



(b) Dynamics for the softmax predictor.

Figure 6: Optimization dynamics from 20 random initializations without the alignment penalties ($\mu_1 = \mu_2 = 0$). The solid curve represents the empirical median across the runs. The bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals. The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$ and $\lambda_t = 10^{-1}/(1 + \max(t - 5 \cdot 10^4, 0) \cdot 10^{-4})$, for the erf predictor and $\lambda_t = 10^{-1}/(1 + \max(t - 2 \cdot 10^4, 0) \cdot 10^{-4})$ for the softmax predictor. For computing the oracle loss we use the initial inverse temperatures.

E.3 Full results for the experiment in Figure 4

In Section 4, we discussed the optimization dynamics under various constant inverse temperatures following an initial warm-up phase. Here, we present the full empirical results for this experiment.

As illustrated in Figures 7, 8, 9, 10, the training procedure begins with a 20,000 step warm up using $\lambda_c = \lambda_r = 0.1$. After this period, the inverse temperatures are held constant at values $\lambda \in \{0.01, 0.03, 0.05, 0.07\}$. We observe that without a meticulously chosen lambda schedule, not all the runs converged in the simulated 100,000 steps. For lower values of λ, more runs got stuck around the origin compared to the more consistent convergence obtained for larger λs.

These full results empirically validate the need to employ a decreasing schedule for the inverse temperatures. A static λ forces a compromise between quickly escaping the saddle point at the origin and converging precisely at the global minima. A schedule dynamically adapts the landscape to first push the parameters away from the saddle point, and subsequently allows for fine-tuned convergence to the exact oracle parameters.

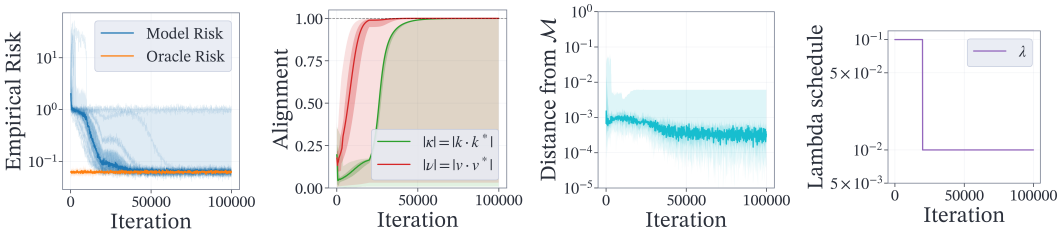


Figure 7: Training dynamics of the predictor with a 20k warm-up phase $\lambda_c = \lambda_r = \lambda = 0.1$ followed by constant inverse temperatures $\lambda_c = \lambda_r = 0.01$. The solid curve represents the empirical median across 20 runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$.

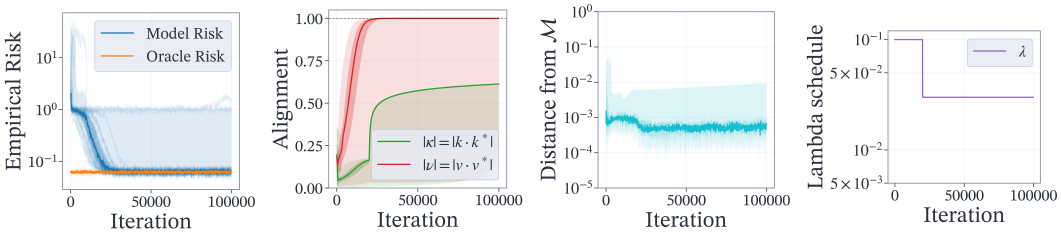


Figure 8: Training dynamics of the predictor with a 20k warm-up phase $\lambda_c = \lambda_r = \lambda = 0.1$ followed by constant inverse temperatures $\lambda_c = \lambda_r = 0.03$. The solid curve represents the empirical median across 20 runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$.

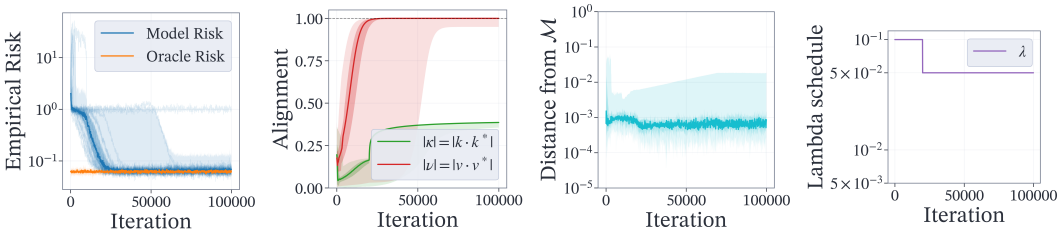


Figure 9: Training dynamics of the predictor with a 20k warm-up phase $\lambda_c = \lambda_r = \lambda = 0.1$ followed by constant inverse temperatures $\lambda_c = \lambda_r = 0.05$. The solid curve represents the empirical median across 20 runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$.

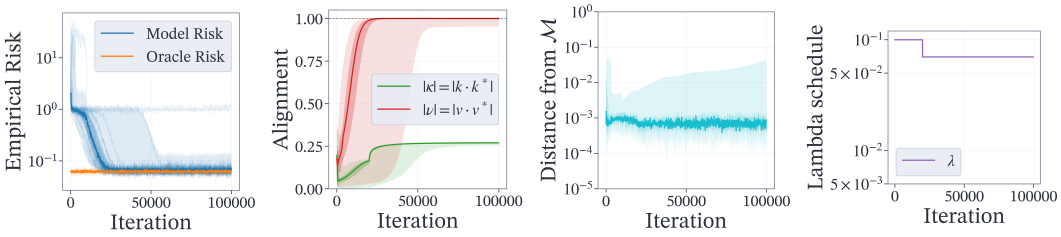


Figure 10: Training dynamics of the predictor with a 20k warm-up phase $\lambda_c = \lambda_r = \lambda = 0.1$ followed by constant inverse temperatures $\lambda_c = \lambda_r = 0.07$. The solid curve represents the empirical median across 20 runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$.

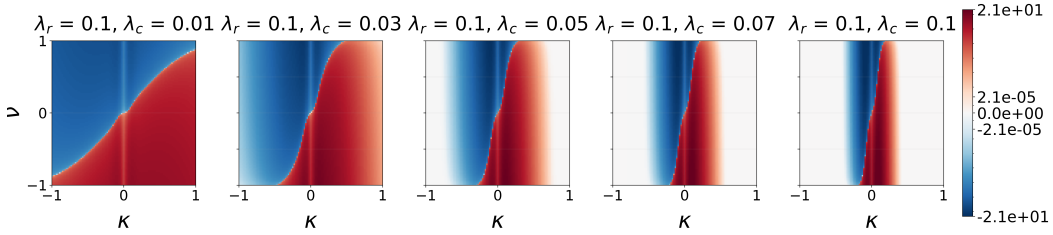
E.4 Gradient analysis for decoupled inverse temperatures

This subsection isolates the effects of decoupling the two inverse temperatures. We visualize the gradient vector field of the reduced risk $\mathcal{R}^{<}_{\lambda_r,\lambda_c}(\kappa,\nu)$ evaluated on the invariant manifold $\mathcal{M}$ for multiple configurations of (λ_r, λ_c).

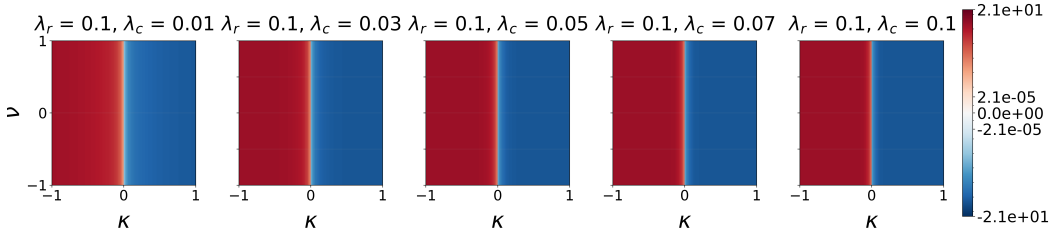
Figures 11, 12 present the numerically computed gradients for various mixed combinations of λ_c and λ_r. By breaking the symmetry, we observe how the magnitude and direction of the gradients shift across the landscape. Specifically, two interesting phenomena appear:

- $\partial_\nu \mathcal{R}^{<}_{\lambda_r,\lambda_c}$ behaves symmetrically with respect to both λ_r and λ_c;
- The same is not true for $\partial_\kappa \mathcal{R}^{<}_{\lambda_r,\lambda_c}$. The shape of the gradient and most importantly of the 0 level curve of $\partial_\kappa \mathcal{R}^{<}$ varies significantly with the pairs (λ_r, λ_c).

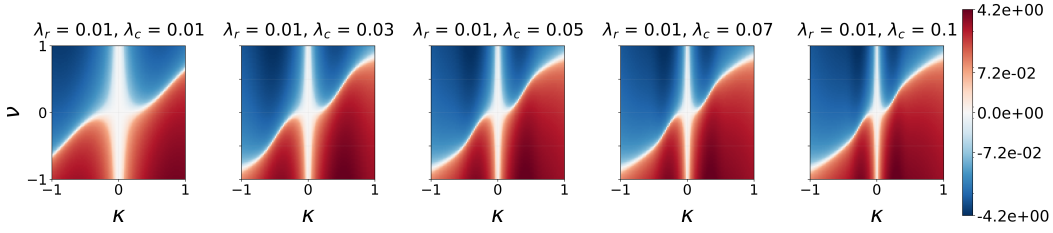
This analysis leads to another method of controlling the training dynamics, namely through the ratio $\tau = \lambda_c/\lambda_r$. Indeed, Figure 13 presents the optimization dynamics obtained for constant $\lambda_c = 0.01$ and $\lambda_r = 0.1$. While 100,000 steps are not enough for convergence at the oracle parameters, the runs consistently move towards convergence. Moreover, the training dynamics are different from the ones studied in the main paper, as can be seen from the much faster increase of $|\kappa|$, closely mirroring the behavior of $|\nu|$.



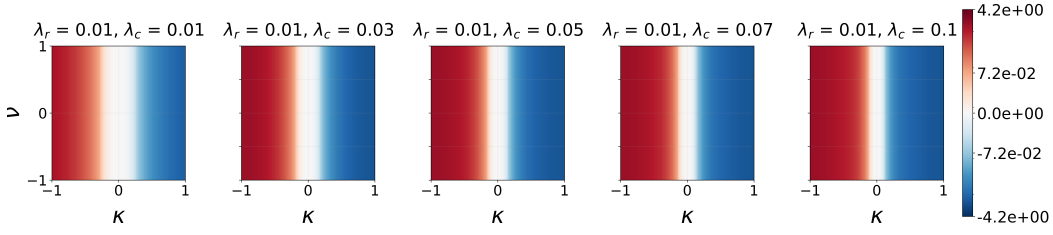
(a) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = 0.1$ and $\lambda_c \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.



(b) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = 0.1$ and $\lambda_c \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.

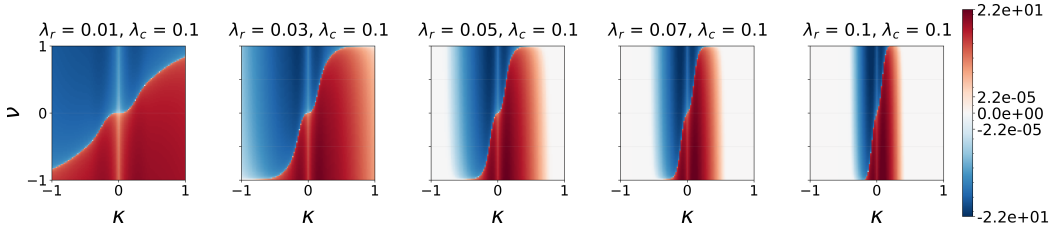


(c) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = 0.01$ and $\lambda_c \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.

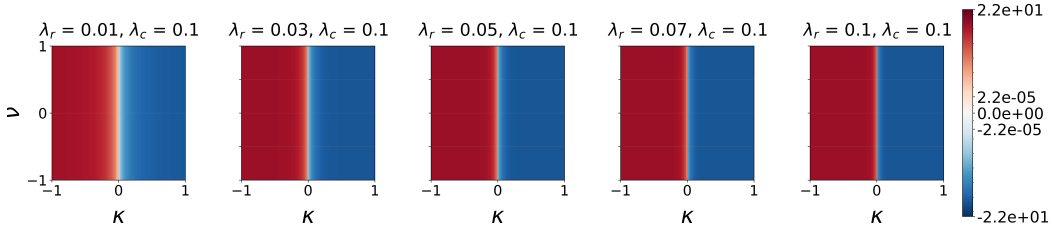


(d) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = 0.01$ and $\lambda_c \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.

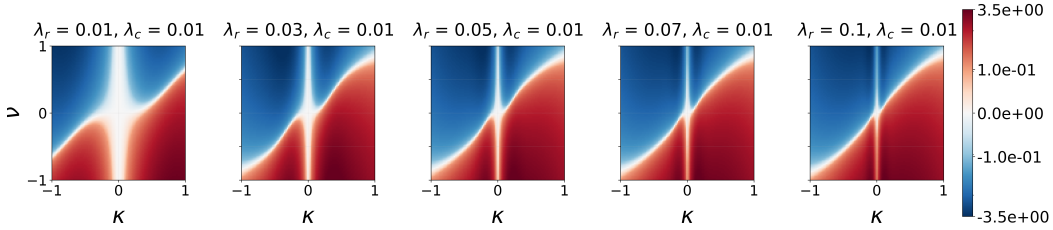
Figure 11: Numerically computed gradients of the theoretical risk $\mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa, \nu)$ for $\lambda_r \in \{0.1, 0.01\}$ and varying values of λ_c. The parameters used are $M = N = 5$ and $\ell = 100$. The gradients, given by Equation (32), are approximated using the Monte Carlo method with a sample size of 10,000 on a 100×100 discretization grid.



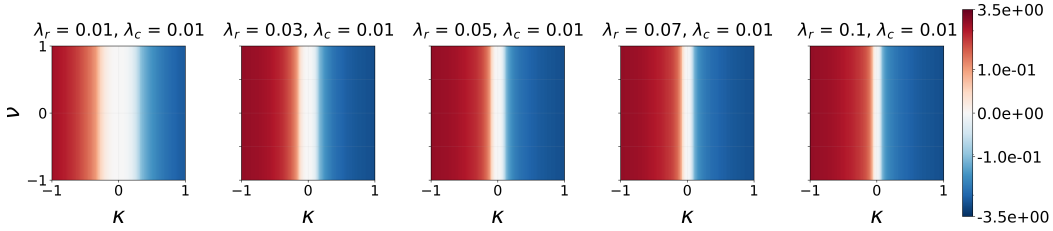
(a) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_c = 0.1$ and $\lambda_r \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.



(b) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_c = 0.1$ and $\lambda_r \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.



(c) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_c = 0.01$ and $\lambda_r \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.



(d) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_c = 0.01$ and $\lambda_r \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$.

Figure 12: Numerically computed gradients of the theoretical risk $\mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa, \nu)$ for $\lambda_c \in \{0.1, 0.01\}$ and varying values of λ_r. The parameters used are $M = N = 5$ and $\ell = 100$. The gradients, given by Equation (32), are approximated using the Monte Carlo method with a sample size of 10,000 on a 100×100 discretization grid.

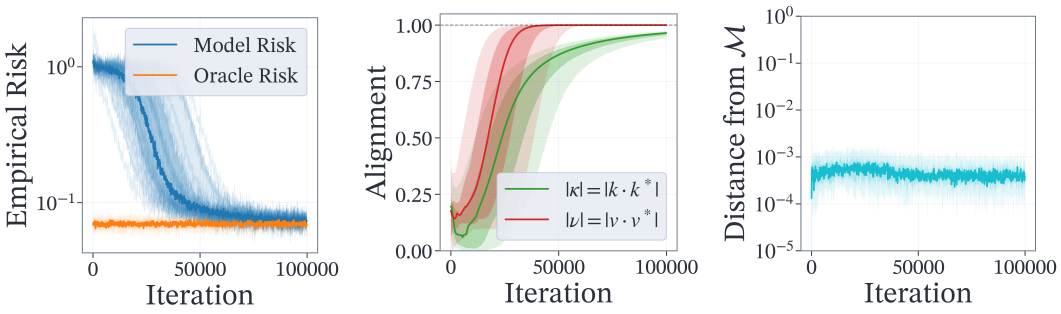
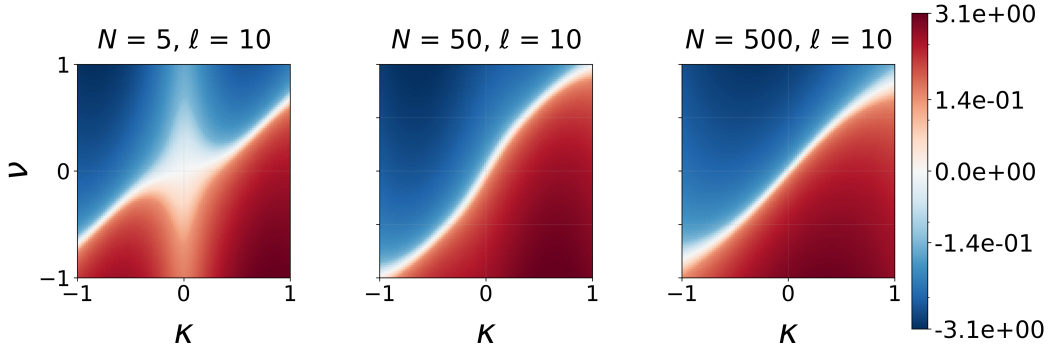


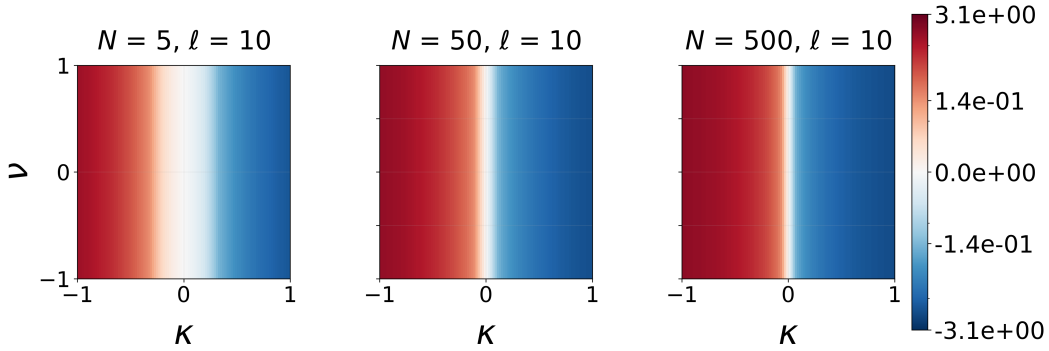
Figure 13: Training dynamics of the predictor with constant $\lambda_c = 0.01$ and $\lambda_r = 0.1$. The solid curve represents the empirical median across the runs and the bands represent the 0.05-0.95 and, respectively, the 0.25-0.75 quantile intervals (although not always visible). The parameters used are $M = N = 5$, $d = 20$, $\gamma = 1$, $\ell = 100$.

E.5 Gradient analysis for other hyperparameters N and ℓ

In this section, we investigate the behavior of the partial derivatives of $\mathcal{R}^<$ when varying N and ℓ, for fixed $\lambda_r = \lambda_c = 0.1$.



(a) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 10$ and $N \in \{5, 50, 500\}$.

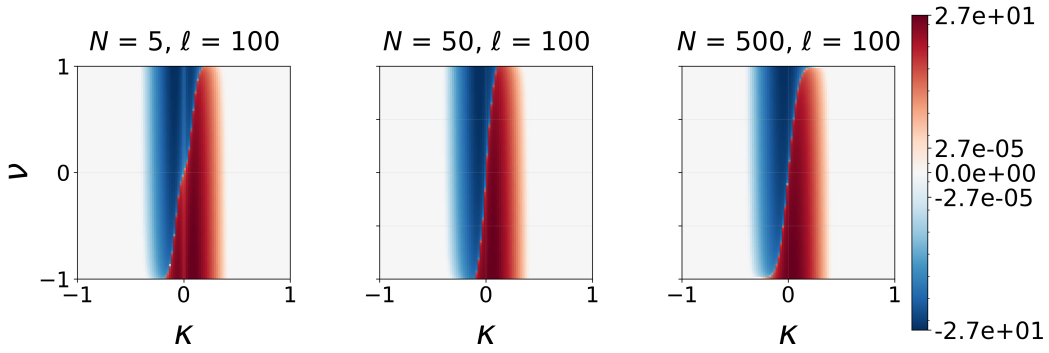


(b) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 10$ and $N \in \{5, 50, 500\}$.

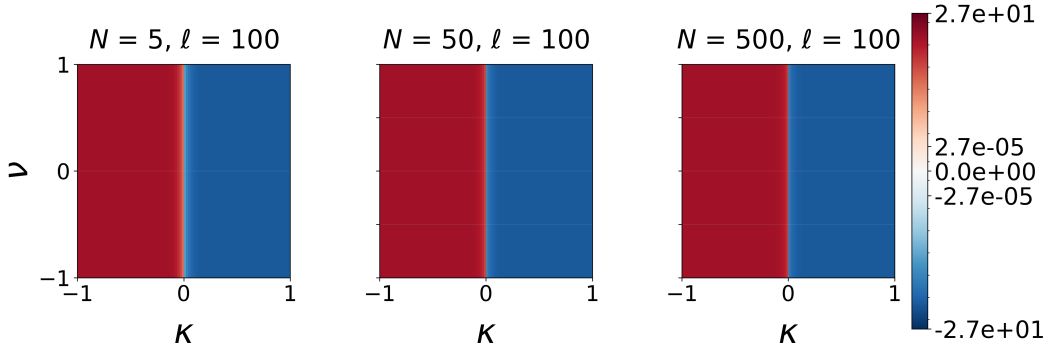
Figure 14: Numerically computed gradients of the theoretical risk $\mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa, \nu)$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 10$ and varying values of N. The gradients, given by Equation (32), are approximated using the Monte Carlo method with a sample size of 10,000 on a 100×100 discretization grid.

The values considered are not following the asymptotics required for Bayes optimality of the predictor, but better reflect real-world scenarios in which TSFMs are trained to take tens of channels and hundreds of time patches as input, while an individual token's shift in the mean cannot be too large

due to normalization layers. We restrict our analysis to exploring the gradient across multiple pairs of values (N, ℓ), as conducting full trainings for each pair would be computationally expensive.



(a) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 100$ and $N \in \{5, 50, 500\}$.

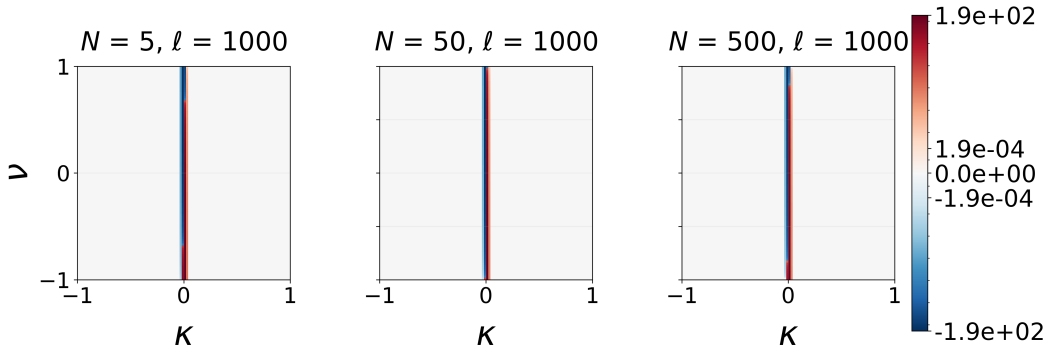


(b) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 100$ and $N \in \{5, 50, 500\}$.

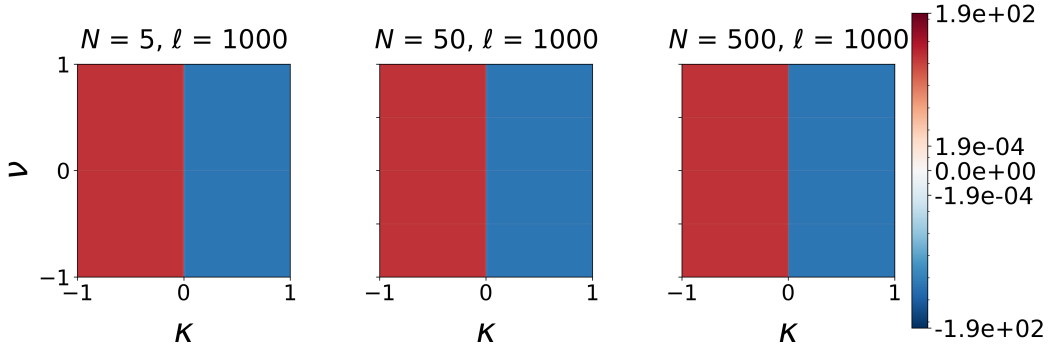
Figure 15: Numerically computed gradients of the theoretical risk $\mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa, \nu)$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 100$ and varying values of N. The gradients, given by Equation (32), are approximated using the Monte Carlo method with a sample size of 10,000 on a 100×100 discretization grid.

As can be seen in Figures 14, 15, 16, the effect of ℓ overshadows that of N. Although for small ℓ, a higher N appears to eliminate the problem of escaping the origin for κ, this disappears for larger ℓ values.

Furthermore, the higher ℓ is, the higher the values of $\partial_\nu \mathcal{R}^<$ and the steeper its sign change around $\kappa = 0$.



(a) Partial derivative $\partial_\kappa \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 1000$ and $N \in \{5, 50, 500\}$.



(b) Partial derivative $\partial_\nu \mathcal{R}^<_{\lambda_r,\lambda_c}$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 1000$ and $N \in \{5, 50, 500\}$.

Figure 16: Numerically computed gradients of the theoretical risk $\mathcal{R}^<_{\lambda_r,\lambda_c}(\kappa, \nu)$ for $\lambda_r = \lambda_c = 0.1$, $\ell = 1000$ and varying values of N. The gradients, given by Equation (32), are approximated using the Monte Carlo method with a sample size of 10,000 on a 100×100 discretization grid.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide proofs for all the theoretical claims made and extensive numerical experiments both in the main text and in the supplemental material for all the empirically observed claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: While not having a separate "Limitations" section in the paper, we discuss limitations of our work in the "Conclusions and limitations" section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the theoretical results and claims are accompanied by full proofs in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the full code used for the numerical experiments in the paper. Furthermore, we provide the complete set of parameters for each experiment in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is given as a resource in the supplemental materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The hyperparameters used in all the experiments are listed in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Empirical distributions are always plotted alongside the main curves.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details regarding the compute resource are given in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the experiments are conducted using synthetically generated data. The authors don't see any possible harm arising from the present research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work in this paper is purely theory-oriented, aiming to better understand existing architectures. The authors see no societal impact related to the present paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The authors don't see any risk in the models used for this research.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: No existing asset is used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only asset released is code for the experiments in the main paper. The code is well documented in the opinion of the authors.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing, editing, or formatting purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.