Optimal Representation Size: High-Dimensional Analysis of Pretraining and Linear Probing

Valentina Njaradi
Gatsby Unit
University College London
valentina.njaradi.23@ucl.ac.uk

Clémentine Dominé
Harvard University
Institute of Science and Technology Austria

Rachel Swanson
Sainsbury Wellcome Centre
University College London

Marco Mondelli *
Institute of Science and Technology Austria

Andrew Saxe *
Gatsby Unit & Sainsbury Wellcome Centre
University College London

Abstract

Data-efficient generalisation is a central objective in both artificial and biological learning. A common strategy is extracting reusable structure from abundant unlabelled data, enabling efficient adaptation to new tasks using limited supervision. This two-stage paradigm underlies modern training pipelines, in which pretraining is followed by fine-tuning or linear probing. We provide an analytical model of this process, formalising structure extraction as principal component analysis on unlabelled data and downstream learning as linear regression on a separate labelled dataset. In the high-dimensional regime, we derive exact expressions for training and generalisation errors, characterising their dependence on representation dimensionality, unlabelled and labelled sample sizes, and task alignment. Our results show that pretrained representations strongly influence downstream generalisation, and we characterize the optimal representation size across data regimes. With abundant pretraining but scarce downstream data, maximally compressed representations are optimal, whereas with limited pretraining data, higher-dimensional representations generalise better. Furthermore, we establish an exact trade-off between pretraining and supervision, quantifying how much unlabelled data is required to replace a single labelled sample. Beyond our idealised model, we observe similar phenomenology in autoencoders and pretrained LLMs. Altogether, we highlight that optimising representation size is critical for downstream generalisation and identify when compression during pretraining is beneficial.

1 Introduction

Large language models, vision transformers, and related architectures have demonstrated language, reasoning and mathematical capabilities widely considered out of reach just a decade ago [1–4]. This progress has been made possible by multi-stage training pipelines capable of leveraging massive amounts of unlabelled data: models are first pretrained on a large corpus of unlabelled data, then fine-tuned on labelled tasks and finally refined, e.g., via reinforcement learning from human feedback

*Equal contribution.

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

[5, 6], low-rank adaptation [7] or test-time training [8–10]. Such a paradigm unlocks remarkable sample efficiency, as large pretrained models generalise on new tasks where they have access only to thousands (or even hundreds) of samples [11, 12].

At their core, these training pipelines rely on a key fact: pretraining on large datasets yields rich representations that effectively transfer to downstream tasks. This idea is not new, as greedy layer-wise training and stacked autoencoders were designed to learn reusable representations in pretraining a decade before modern transformers [13–15]. However, different pretraining objectives emphasize different aspects of the data, and it remains unclear which features of a representation are most useful for downstream learning.

Characterizing these representations, how they support computation, transfer to new tasks and enable adaptive behaviour in artificial and biological systems, has been a longstanding focus in machine learning, as well as cognitive science and neuroscience [16–18]. In machine learning, structured representations and prior knowledge can accelerate learning and enhance generalisation, but may also hinder adaptation, as exemplified by catastrophic interference in continual and reversal learning [19–22]. Similar trade-offs arise in biological systems, where representations must support flexible and efficient adaptation [23]. A crucial property underlying these effects is the size (or dimensionality) of representations. Both high and low-dimensional representations have been observed across a range of brain regions and tasks [24–27] as well as in machine learning models [11, 28], yet a principled account of how dimensionality interacts with task and environment structure to shape learning performance and sample efficiency remains missing.

Here, we study a two-stage learning setting in which representations are first learned via principal component analysis (PCA) from unlabelled samples drawn from a spiked covariance model, and then reused by linear regression on an independent labelled dataset. The covariance structure plays a central role: it produces both a tractable model for analysis and controlled alignment between the pretraining and supervised training stages, thus enabling us to isolate how pretrained representations influence downstream generalisation. More precisely, our contributions are as follows:

- We derive exact asymptotic expressions for estimation, generalisation and training error with distinct pretraining and downstream spikes (Theorem 1, 2 and 3, respectively). From this, we establish the improvement in generalisation coming from tuning the representation size, as well as the improvement due to adding more pretraining and finetuning data. For optimally-tuned errors, we identify regions where additional unlabelled samples reduce generalisation error more than additional labelled samples (Figure 2).
- We characterize when either *(i)* low-dimensional representations are useful for generalisation on downstream tasks, or *(ii)* tasks benefit from detailed high-dimensional representations (Figure 3). In particular, we identify phase transitions where compression becomes useful (Corollaries 1, 2), and describe their behaviour for an infinite number of pretraining samples (Corollary 4).
- We finally show that the theoretical predictions derived in our idealized setting extend to trained autoencoders when the number of samples is larger than the number of features, and further demonstrate that the insights unveiled by our theory persist in more realistic settings through linear probing of pretrained transformers (Figure 4).

2 Related work

High-dimensional regression. High-dimensional regression, in the regime where number of features and samples grow proportionally, has been the subject of intense investigation, leading to the precise characterization of the test error [29–32] and phenomena such as benign overfitting [33] and double descent [34, 35]. This line of work further characterizes the distribution of the empirical risk minimiser [36], the impact of spurious correlations [37], performative learning [38], as well as training on multiple or surrogate data sources [39–42]. Principal component regression (PCR), where regression is performed on principal components (PCs) of the sample covariance, provides a natural way to restrict regression to a low-dimensional subspace. Its performance depends critically on the number of retained components, which governs the bias–variance trade-off [43–46], and has been characterized both non-asymptotically [47] and in proportional asymptotics for population and sample PCs [44, 45, 48]. In contrast, we study pretrained regression (PR), with PCs formed during a pretraining phase and reused for a downstream task, separating the sample sets for PCA

and subsequent regression. As a result, both the asymptotic risk and the optimal number of principal components for PR differ fundamentally from the classical PCR setting.

Pretraining and transfer learning. More broadly, the reuse of pretrained representations across tasks has been extensively studied in transfer learning, primarily in supervised settings where generalisation depends on task similarity or alignment with pretrained features [49–55]. These approaches typically do not model representations learned from unlabelled data. Complementary work establishes benefits of unsupervised and self-supervised pretraining [56–59]. Closest to our paper, recent work analyses PCA-based pretraining in spiked models [60], but considers fixed representations without addressing the role of representation dimensionality in generalisation.

Representation dimensionality. A growing body of work highlights that pretrained representations often lie in low-dimensional subspaces, as evidenced by studies of intrinsic dimensionality [11, 28], low-rank adaptation [7], and linear probing [61]. Recent theoretical work has begun to formalise aspects of this structure [62, 63]. While these papers study the structure of pretrained features and how they support downstream prediction, they do not provide a closed-form characterisation of how the optimal retained dimension depends jointly on pretraining, downstream data, and task structure. Addressing this question through an analytically tractable theory is the focus of our work.

Spiked covariance models. The spiked covariance model [64] provides a tractable framework for studying low-dimensional structure in high-dimensional data. A key feature is the Baik–Ben Arous–Péché (BBP) phase transition [65–67]: above this threshold, empirical principal components align with the population spike, while below it they are asymptotically uninformative [65, 68–70]. This transition makes the spiked models well-suited for analysing how representation quality depends on sample size and signal strength. Accordingly, they have been used to analyse kernel methods, neural networks and feature learning [71–73], as well as the emergence of low-rank structure during training [74–77]. Finally, spiked covariances have been used to analyse PCR, showing how the projection onto empirical PCs has a regularising effect [78]. Our setting differs from these works: the signal subspace is estimated from an independent unlabelled pretraining sample and then used for a separate labelled task, allowing us to characterise the optimal representation dimension.

3 Setup

We study two-stage learning with unlabelled pretraining followed by supervised learning:

1. In the pretraining stage, we observe inputs $\mathbf{X}_u \in \mathbb{R}^{n_u \times p}$ whose rows are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_u)$, forming the *unlabelled* dataset $\mathcal{D}_u$. From this data, the model extracts a representation via PCA: it computes the top m principal components of $\mathbf{X}_u$, stacks them into $\mathbf{U}_m \in \mathbb{R}^{p \times m}$, and defines a projection of retained directions $\mathcal{P}_m = \mathbf{U}_m \mathbf{U}_m^\top$ with a discarded complement $\mathcal{P}_\perp = I - \mathcal{P}_m = \mathbf{U}_\perp \mathbf{U}_\perp^\top$. If $m > n_u$, the PCA subspace is completed with uniformly random orthonormal vectors.
2. In the downstream stage, we observe a *labelled* dataset $\mathcal{D}_l = \{\mathbf{X}_l, \mathbf{y}_l\}$, where $\mathbf{y}_l = \mathbf{X}_l \mathbf{w}^\star + \xi \in \mathbb{R}^{n_l}$, the rows of $\mathbf{X}_l \in \mathbb{R}^{n_l \times p}$ are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_l)$ and independent of $\mathbf{X}_u$, $\mathbf{w}^\star \in \mathbb{R}^p$ is the ground-truth signal and $\xi \in \mathbb{R}^{n_l}$ is Gaussian noise with i.i.d. entries of variance σ^2.

The pretrained regression model uses the representation learned from $\mathcal{D}_u$ to fit the downstream task by regressing on the projected inputs (see Figure 1**(a)** for a schematic representation):

$$\hat{\mathbf{w}} = (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{y}_l = (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star + (\mathbf{X}_l \mathcal{P}_m)^\dagger \xi, \tag{1}$$

where $(\cdot)^\dagger$ denotes the pseudoinverse. Unlike classical PCR [45], $\mathcal{P}_m$ is learned from independent unlabelled data, so $(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \neq \mathcal{P}_m$ in general, yielding distinct dependence on pretraining and downstream sample sizes (see Appendix G.1 for a longer discussion).

The quantities of interest are estimation, generalisation and training error:

$$E^{\text{est}} = \mathbb{E}_\xi \left[\|\mathbf{w}^\star - \hat{\mathbf{w}}\|^2 \right], \quad E^{\text{gen}} = \mathbb{E}_{\mathbf{x},y,\xi} \left[(y - \mathbf{x}^\top \hat{\mathbf{w}})^2 \right], \quad E^{\text{train}} = \frac{1}{n_l} \mathbb{E}_\xi \left[\|\mathbf{y}_l - \mathbf{X}_l \hat{\mathbf{w}}\|^2 \right]$$

where the test point $(\mathbf{x}, y)$ in E^{gen} is drawn from the labelled dataset $\mathcal{N}(0, \mathbf{\Sigma}_l)$. We work in the high-dimensional regime where $p/n_l \to \gamma_l \in (0, \infty)$, $p/n_u \to \gamma_u \in (0, \infty)$, $m/p \to \alpha \in [0, 1]$ with α the fraction of dimensions to keep, and γ_u, γ_l the aspect ratios for the datasets $\mathcal{D}_u, \mathcal{D}_l$. Due to the projection of the downstream inputs, we define the effective aspect ratio $\gamma_{\text{eff}} := \frac{m}{n_l} \to \alpha \gamma_l$ and

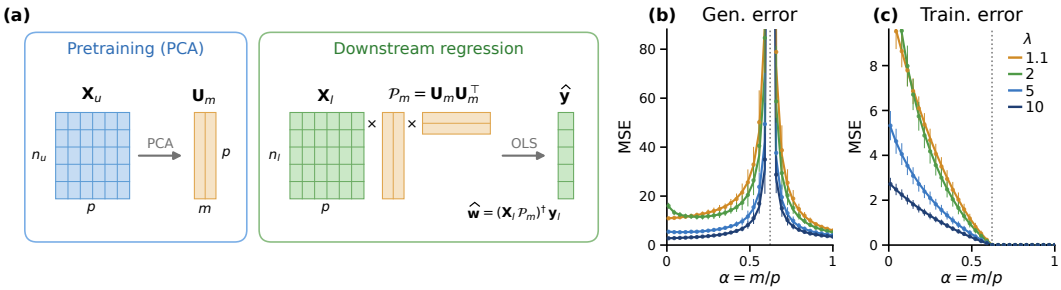


Figure 1: **Analytically tractable model of two-stage learning**. **(a)** In the pretraining stage, PCA extracts the top-m principal components from n_u unlabelled samples. The downstream task with n_l labelled samples is learned via regression on inputs projected through $\mathcal{P}_m = \mathbf{U}_m\mathbf{U}_m^\top$. Theoretically derived generalisation **(b)** and training errors **(c)** match numerical simulations. Solid lines show the theory; dots and error bars show the mean $\pm$ one standard deviation over 50 simulations. Colours indicate the spike strength λ; the dotted line marks the interpolation threshold $m = n_l$.

assume $\gamma_{\text{eff}} \neq 1$. The same analysis also extends to ridge-regularised downstream regression; the formulation and results are given in Appendix D.

The pretraining and downstream population covariances follow rank-one spiked models

$$\mathbf{\Sigma}_u = \mathbf{I}_p + (\lambda_u - 1)\mathbf{v}\mathbf{v}^\top, \qquad \mathbf{\Sigma}_l = \mathbf{I}_p + (\lambda_l - 1)\mathbf{q}\mathbf{q}^\top, \tag{2}$$

where $\lambda_u, \lambda_l > 1$ and $\mathbf{v}, \mathbf{q} \in \mathbb{R}^p$ are deterministic unit vectors. Finally, we define the following overlaps:

$$r := \mathbf{v}^\top\mathbf{q} \qquad r_v := \mathbf{w}^{\star\top}\mathbf{v}/\|\mathbf{w}^\star\| \qquad r_q := \mathbf{w}^{\star\top}\mathbf{q}/\|\mathbf{w}^\star\|$$

4 Theoretical analysis of two-stage learning

In this section, we derive asymptotic expressions for the estimation, generalisation, and training errors of the estimator in (1) and find the representation dimensionality α that minimizes the errors. We then characterise when the optimal α shifts from retaining all PCs to retaining only a few PCs, as a function of n_u and n_l. We further study how n_u, n_l, λ and the signal-to-noise ratio $\text{SNR} = \frac{\|\mathbf{w}^*\|^2}{\sigma^2}$ affect both the errors and the optimal α, as well as the limiting behaviour with infinite pretraining data ($\gamma_u = 0$). Proofs of all results are deferred to Appendix A.

4.1 Asymptotic expressions

The following deterministic limits describe how the pretraining projections $\mathcal{P}_m$ and $\mathcal{P}_\perp$ act on fixed directions $\mathbf{a}, \mathbf{b}$. For the retained and discarded pretraining subspaces,

$$\frac{\mathbf{a}^\top\mathcal{P}_\perp\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \xrightarrow{a.s.} \bar{p}_{\perp,\mathbf{a},\mathbf{b}}, \qquad \frac{\mathbf{a}^\top\mathcal{P}_m\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \xrightarrow{a.s.} \bar{p}_{\mathbf{a},\mathbf{b}}. \tag{3}$$

These limits are obtained by expanding the projections over the sample PCs, e.g., $\mathbf{a}^\top\mathcal{P}_\perp\mathbf{b} = \sum_{i=m+1}^{p}(\mathbf{a}^\top\mathbf{u}_i)(\mathbf{b}^\top\mathbf{u}_i)$, and relating sample-PC overlaps to their population counterparts using random-matrix theory (Appendix A.6).

To characterise the additional projection induced by finite downstream data, define

$$\Pi := (\mathbf{X}_l\mathcal{P}_m)^\dagger(\mathbf{X}_l\mathcal{P}_m),$$

the projector onto the row space of the projected downstream dataset. When $\gamma_{\text{eff}} < 1$, $\mathbf{X}_l\mathbf{U}_m$ has full column rank and $\Pi = \mathcal{P}_m$, so all retained directions are identifiable. When $\gamma_{\text{eff}} > 1$, its nullspace removes part of the retained subspace, with the reduction depending on γ_{eff} and the effective downstream spike $\bar{\lambda}_l := 1 + (\lambda_l - 1)\bar{p}_{\mathbf{q},\mathbf{q}}$:

$$\frac{\mathbf{a}^\top\Pi\mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \xrightarrow{a.s.} \bar{\pi}_{\mathbf{a},\mathbf{b}} = \begin{cases} \bar{p}_{\mathbf{a},\mathbf{b}}, & \gamma_{\text{eff}} < 1, \\ \dfrac{\bar{p}_{\mathbf{a},\mathbf{b}}}{\gamma_{\text{eff}}} + \dfrac{\bar{p}_{\mathbf{a},\mathbf{q}}\bar{p}_{\mathbf{b},\mathbf{q}}}{\bar{p}_{\mathbf{q},\mathbf{q}}}\left(\dfrac{\bar{\lambda}_l}{\gamma_{\text{eff}} - 1 + \bar{\lambda}_l} - \dfrac{1}{\gamma_{\text{eff}}}\right), & \gamma_{\text{eff}} > 1. \end{cases} \tag{4}$$

These quantities determine the asymptotic errors below.

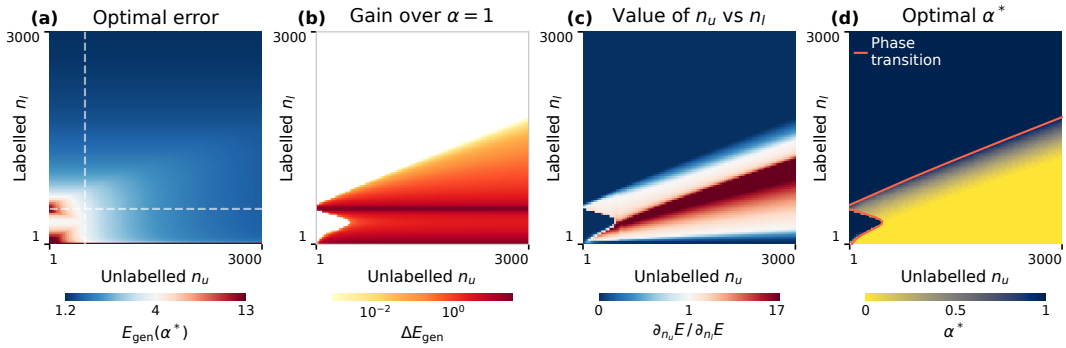


Figure 2: **Benefits of optimizing representation size**. **(a)** Minimum achievable generalisation error with optimal $\alpha^\star$. **(b)** Gain in generalisation error ($\Delta E_{\text{gen}} = E_{\text{gen}}(1) - E_{\text{gen}}(\alpha^*)$) from using $\alpha^\star$ instead of all PCs ($\alpha = 1$). **(c)** Marginal rate of substitution between unlabelled and labelled data, $(\partial E_\infty^{\text{gen}}/\partial n_u) \,/\, (\partial E_\infty^{\text{gen}}/\partial n_l)$; values above 1 (red) indicate that unlabelled data reduces the error more than labelled data, and vice-versa below 1 (blue). **(d)** Generalisation-optimal representation size $\alpha^\star$.[1]Simulation details are given in Appendix H.

Theorem 1 (Asymptotic estimation error)**.** *As $n_l, n_u, p, m \to \infty$ with $p/n_l \to \gamma_l$, $p/n_u \to \gamma_u$, $m/p \to \alpha$ and $\gamma_{\text{eff}} \neq 1$, the estimation error converges a.s. to*

$$E_\infty^{\text{est}} = \bar{w}^* \Bigg(\underbrace{1 - \bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}}_{\textit{missing signal}} + \underbrace{\frac{(\lambda_l - 1)^2}{\bar{\lambda}_l^2} \bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}\, \bar{\pi}_{\mathbf{q},\mathbf{q}}}_{\textit{leaked signal}} \Bigg) + \underbrace{\frac{\min\{\gamma_{\text{eff}}, 1\}}{|\gamma_{\text{eff}} - 1|} \left(\bar{w}^* \bar{\sigma}^2_{\text{eff}} + \sigma^2 \right)}_{\textit{variance}}, \tag{5}$$

where $\bar{w}^*$ is the deterministic limit of $\|\mathbf{w}^\star\|^2$, $\bar{\lambda}_l = 1 + (\lambda_l - 1)\bar{p}_{\mathbf{q},\mathbf{q}}$ is the effective spike eigenvalue, and the variance from the unrecoverable part of $\mathbf{w}^\star$ is $\bar{\sigma}^2_{\text{eff}} = \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star} + \frac{\lambda_l - 1}{\bar{\lambda}_l} \bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}$.

The expression in (5) admits a simple geometric interpretation: the estimator observes $\mathbf{w}^\star$ through the subspace induced by $\mathbf{X}_l \mathcal{P}_m$, and the error reflects how $\mathbf{w}^\star$ decomposes relative to this subspace. By decomposing into two components, aligned or orthogonal to $\mathbf{U}_m$, $\mathbf{w}^\star = \mathbf{w}^\star{}_\parallel + \mathbf{w}^\star{}_\perp$, the error splits into three contributions: *(a)* the *missing signal*, *(b)* the *leaked signal*, and *(c)* the *variance* term.

(a) The term $1 - \bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}$ captures the *missing signal*, i.e. directions that lie in the retained subspace but cannot be reliably estimated from finite samples.

(b) A distinct contribution comes from the *leak*. In an isotropic model, the components $\mathbf{X}_l \mathcal{P}_m$ and $\mathbf{X}_l \mathcal{P}_\perp$ are independent, so $\mathbf{w}^\star{}_\perp$ is completely invisible to the estimator. Here, the spike in the covariance couples these two parts of the data, so directions outside the retained subspace become correlated with it. As a result, part of $\mathbf{w}^\star{}_\perp$ leaks into Π, where it appears as a spurious signal aligned with $\Pi\mathbf{q}$, producing the bias term $\frac{(\lambda_l - 1)^2}{\bar{\lambda}_l^2} \bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}\, \bar{\pi}_{\mathbf{q},\mathbf{q}}$. For a detailed discussion of parameter effects on the leak and a breakdown of different terms in the estimation error, see Figure 6 in Appendix E.

(c) The remaining part of $\mathbf{w}^\star{}_\perp$ does not couple back into the learned subspace. Together with label noise, this contributes to the *variance*, which exhibits a double descent [34] controlled by $\gamma_{\text{eff}} = \alpha\gamma_l$. Thus, even without explicit regularisation in the regression step, the model mitigates double descent by controlling the representation size α, see Figure 6 in Appendix E.

Theorem 2 (Asymptotic generalisation error)**.** *Under the same regime, the generalisation error converges a.s. to*

$$E_\infty^{\text{gen}} = E_\infty^{\text{est}} + \bar{w}^*(\lambda_l - 1) \Bigg(\underbrace{r_q - \bar{\pi}_{\mathbf{w}^\star,\mathbf{q}}}_{\textit{missing signal}} - \underbrace{\frac{\lambda_l - 1}{\bar{\lambda}_l} \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}\, \bar{\pi}_{\mathbf{q},\mathbf{q}}}_{\textit{leaked signal}} \Bigg)^2 + \sigma^2.$$

The calculation of the generalisation error builds on that of the estimation error, with an additional contribution measuring how the prediction error $\mathbf{w}^\star - \hat{\mathbf{w}}$ projects onto the spike $\mathbf{q}$ in the test population covariance. The additional error depends on the mismatch between the true alignment r_q

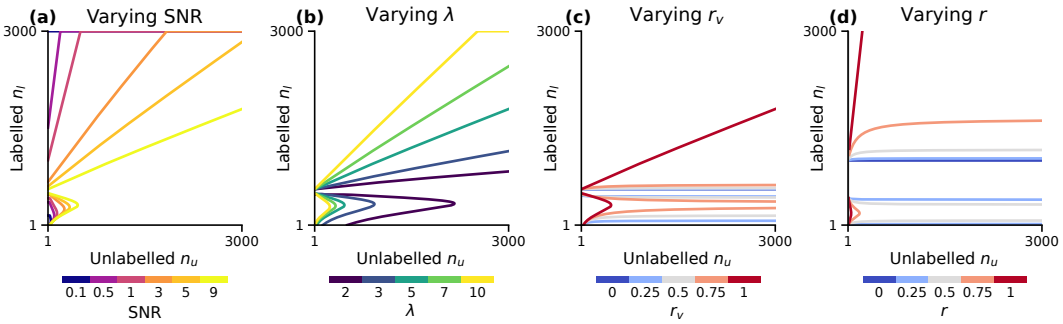


Figure 3: **When is compression useful**. **Phase boundaries in the** (n_u, n_l) **plane for (a)** varying SNR **(b)** varying λ **(c)** varying r_v **(d)** varying r. Non-varying parameters are kept fixed, see simulation details are given in Appendix H.

and the alignment recovered by the estimator. As in the estimation error, this mismatch decomposes into a *missing signal* term and a *leaked signal* term, arising from finite-sample limitations and the coupling induced by the spike. The final σ^2 term reflects the irreducible error due to label noise. For additional discussions and a breakdown of different terms in the generalisation error, see Figure 7 in Appendix E.

Theorem 3 (Asymptotic training error)**.** *Under the same regime, the training error converges a.s. to*

$$E_\infty^{\text{train}} = \left(\bar{w}^* \bar{\sigma}_{\text{eff}}^2 + \sigma^2\right) \max(0,\, 1 - \gamma_{\text{eff}})\,.$$

This expression has the standard interpolation form. When $\gamma_{\text{eff}} > 1$, the model is overparameterised and perfectly fits the training data, yielding zero error. When $\gamma_{\text{eff}} < 1$, the model cannot interpolate, and the error decreases linearly with γ_{eff}, with slope set by the total residual variance $\bar{w}^* \bar{\sigma}_{\text{eff}}^2 + \sigma^2$.

Altogether, these results provide exact expressions for generalisation, estimation, and training errors, predicting several key trends. In Figure 1**(b-c)**, we illustrate these errors as functions of the compression ratio α under varying data structures λ. The plots demonstrate an excellent agreement between theoretical predictions and experimental results for moderate values of n_u, n_l, p. Increasing λ reduces the bias from the missing signal, the leak and also the variance. However, it does not reduce the errors across α uniformly, as errors with low α benefit more from stronger structure than errors with high α. We validate our theory across a range of parameter settings, with experiments exploring variations in the signal-to-noise ratio, the alignments r, r_v, r_q and the number of unlabelled and labelled samples n_u, n_l (see Figure 5 in Appendix E).

Importantly, optimising the representation dimensionality reduces generalisation error across a broad region of parameter space relative to full-rank representations ($\alpha = 1$, Figure 2**(b)**), with the optimum shown in Figure 2**(a)**. These gains are particularly pronounced in regimes with substantial pretraining data and limited downstream samples, where standard PCR or regular regression fail to achieve comparable performance even in misaligned task settings (see Figure 22 in Appendix G.1 for a comparison).

Next, we study the trade-off between pretraining and supervision at the optimal $\alpha^\star$ by quantifying how much error reduction is obtained from additional unlabelled versus labelled samples (Figure 2**(c)**). The resulting phase map reveals a nontrivial dependence on both data sources. In particular, there exists a broad interior region of the phase diagram—spanning moderate to large n_u and increasing n_l—in which additional pretraining data yields larger reductions in generalisation error than labelled samples. Interestingly, this regime is closely linked to the optimal representation size $\alpha^\star$ that minimises generalisation error (see Figure 2**(d)**), which is discussed below.

4.2 Optimal representation size

Generalisation in two-stage learning can be hindered by many factors, including misalignment between the pretraining and downstream datasets, weak signal, task difficulty, and limited sample sizes (Theorem 2). Yet representation dimensionality provides a controllable means of improving

[1] If m does not grow with p (e.g., $m = 1$), then $m/p \to \alpha = 0$. Thus, in simulations, $\alpha \approx 0$ refers to taking m not growing with p (e.g., retaining only the top principal component $m = 1$).

performance (Figures 1 and 5): choosing the representation size appropriately can substantially mitigate these sources of degradation. We therefore characterise the optimal retained fraction $\alpha = m/p$ as a function of n_u, n_l, the SNR, the spike strength λ, and the alignment of the task $\mathbf{w}^\star$ with the salient data directions $\mathbf{v}$ and $\mathbf{q}$.

Generalisation error. We find that the optimal level of compression minimising generalisation error adapts to the availability of unlabelled and labelled data. For fixed SNR, λ and an aligned task $\mathbf{w}^\star \parallel \mathbf{v}$, Figure 2**(d)** reveals the following three regimes.

- *(i)* Scarce n_l and n_u: not enough pretraining samples to reliably estimate PCs, thus downstream tasks should not be compressed ($\alpha^\star \approx 1$).
- *(ii)* Large n_u, low n_l: maintain a compact representation subspace ($\alpha^\star < 1$).
- *(iii)* High n_l: little compression needed ($\alpha^\star \approx 1$), as the number of task samples is enough to learn.

We argue that regime *(ii)* is the most representative of modern machine-learning pipelines, where pretraining data is abundant but downstream task-specific data are limited. In practice, modern models are typically pretrained on large-scale datasets [2, 79] and then fine-tuned on substantially smaller task-specific datasets [3]. Altogether, we observe two phase transitions: one between regimes *(i)* and *(ii)*, and another between regimes *(ii)* and *(iii)*. We discuss them below in detail.

Phase transitions for generalisation error. We derive approximate analytical phase transition lines as a function of signal-to-noise ratio (SNR), pretraining and downstream spike strengths (λ_u, λ_l), task and spike alignments and sample allocation between pretraining and downstream tasks (n_u and n_l). These approximations characterise the boundaries separating regimes in which different representation dimensions ($\alpha^\star$) are optimal, and they provide a tractable description of the underlying transitions observed in simulations.

First we consider the transition between regions *(i)* and *(ii)* when $\gamma_l > 1$. In this regime, the generalisation error exhibits a double descent, with competing minima at large and small α (see Figure 2**(b)**). Although their exact locations can in principle be obtained from the closed-form error expressions, the resulting formulas are not analytically transparent. Instead, we provide an interpretable approximation by comparing the error at the two extremes $\alpha = 1$ and $\alpha = 0$, corresponding to retaining all components and only a fixed positive number of leading PCs.

Corollary 1 (Phase transition for $\gamma_l > 1$)**.** *Let $\gamma_l > 1$, $\mathrm{SNR} \to S = \bar{w}^*/\sigma^2$. The overlap between the leading PC of the population and sample covariance of the dataset $\mathcal{D}_u$ has a deterministic limit*

$$c_{\gamma_u} = \mathbf{1}_{\gamma_u < (\lambda_u - 1)^2} \frac{1 - \gamma_u/(\lambda_u - 1)^2}{1 + \gamma_u/(\lambda_u - 1)}.$$

Then, the transition between the small-α and large-α minima is approximated by the endpoint condition $E_\infty^{\mathrm{gen}}(0) = E_\infty^{\mathrm{gen}}(1)$, which yields a parametric curve

$$1 - r_v^2\, c_{\gamma_u} + \frac{(\lambda_l - 1)(r_q - r_v r\, c_{\gamma_u})^2}{1 + (\lambda_l - 1) r^2\, c_{\gamma_u}} = (\gamma_l - 1)\left[\frac{1 - r_q^2}{\gamma_l} + \frac{r_q^2 \lambda_l \gamma_l}{(\gamma_l - 1 + \lambda_l)^2}\right] + \frac{1}{S(\gamma_l - 1)} \quad (6)$$

When the competing minima lie close to the two endpoints, crossing the curve (6) approximately marks an abrupt switch in the optimal representation size $\alpha^\star$ between the small- and large-α regimes. The derivation is given in Appendix B.1.

Corollary 2 characterises the loss of local stability of the full representation in the underparametrised regime $\gamma_l < 1$. When the generalisation error exhibits a single minimum, as in the shared-spike examples (see e.g. Figure 5**(e,g,h)** in Appendix E.1), this describes the transition between regions *(ii)* and *(iii)*. The point $\alpha = 1$ is locally optimal as long as decreasing α increases the generalisation error. The boundary is determined by the condition $\left.\frac{\partial E_\infty^{\mathrm{gen}}}{\partial \alpha}\right|_{\alpha = 1^-} = 0$.

Corollary 2 (Phase transition for $\gamma_l < 1$)**.** *Let $\gamma_l < 1$, $\mathrm{SNR} \to S = \bar{w}^*/\sigma^2$, and denote by $D_{\mathbf{v}}(\gamma_u; \lambda_u)$ the α-derivative of $\bar{p}_{\mathbf{v},\mathbf{v}}$ at $\alpha = 1^-$. Then, the solution $\alpha = 1$ is locally stable if and only if $\gamma_l < \gamma_l^{crit}$:*

$$\gamma_l^{\mathrm{crit}} = \frac{S\left[(1 - r_v^2) + r_v^2\, D_{\mathbf{v}}(\gamma_u; \lambda_u)\right]}{1 + S\left[(1 - r_v^2) + r_v^2\, D_{\mathbf{v}}(\gamma_u; \lambda_u)\right]}, \qquad D_{\mathbf{v}}(\gamma_u; \lambda_u) = \begin{cases} \dfrac{\lambda_u \gamma_u}{(\lambda_u - 1 + \sqrt{\gamma_u})^2}, & \gamma_u < 1, \\ \dfrac{1}{\lambda_u}, \dfrac{\gamma_u}{\gamma_u - 1 + \lambda_u}, & \gamma_u > 1. \end{cases}$$

Depending on the regime, the resulting boundary induces either a linear or nonlinear scaling of the critical number of labelled samples n_l with the unlabelled sample size n_u. Below this threshold, the optimal representation shifts to $\alpha < 1$, and compression improves downstream generalisation. The derivation of this result is in Appendix B.2.

Of particular interest is the limit $\gamma_u \to 0$, corresponding to infinite unlabelled data—a regime reminiscent of modern pretraining datasets [1, 2, 80]. Here, errors and phase boundaries are independent of γ_u and have simple dependencies on r_v,, SNR, λ, and r. Precise expressions are in Appendix C (Corollary 3), where we also consider the limit of infinite labelled data, $\gamma_l \to 0$ (Corollary 4). In this last regime, the generalisation error simplifies (see Eq. (45)), the double-descent behaviour disappears, and the effective projector $(\mathbf{X}\mathcal{P}_m)^\dagger \mathbf{X}\mathcal{P}_m$ reduces to the population projector onto $\mathcal{P}_m$.

The full phase transition curve, derived in Corollaries 1 and 2, matches the phase transition observed in the heat map of Figure 2**(d)** (shown in white; see also additional heatmaps in Figures 8 and 10 in Appendix E.2). Increasing the SNR progressively shrinks the region where compression is beneficial (Figure 3**(a)**), making full-rank representations ($\alpha = 1$) optimal over a wider range of sample sizes. The converse is true when increasing the spike strength λ, the task-spike alignment r_v or the spikes-alignment r (Figure 3**(b, c, d)**). These trends are evident from Corollary 2: when $\gamma_l < 1$, the stability boundary depends on $S(1 - r_v^2)$, so increasing SNR or decreasing r_v directly reduces γ_l^{crit}. Additional heatmaps for α in Appendix E.2 (Figures 8, 10, 12) further validate the joint dependence on SNR, λ, r_v and r, as well as the match with the theoretically derived transitions. In particular, increasing SNR, and decreasing λ, r_v and r all act to suppress the benefits of compression: for $\gamma_l < 1$ the phase boundary flattens, while for $\gamma_l > 1$ it shifts along the n_l axis, reducing the region where $\alpha < 1$ is optimal. Intuitively, a weaker spike is harder to estimate, lower alignment reduces the relevance of prior structure, and higher SNR makes the downstream task easy to learn from scratch. These effects diminish the value of compression, even when labelled data is abundant.

Training error. Our analysis further characterises the optimal value of α for minimising the training error. In the oversampled regime ($n_l > p$, corresponding to $\gamma_l < 1$), the optimal representation is full-rank, i.e., $\alpha = 1$. Conversely, in the undersampled regime ($n_l < p$, corresponding to $\gamma_l > 1$), the training error is equal to zero. In this case, the optimal dimensionality matches the sample size, i.e., $\alpha = n_l/p$, since increasing the dimension beyond this point provides no additional information. This is illustrated in Figures 9 and 11 in Appendix E.2.

Altogether, our exact asymptotic analysis provides a theoretical framework to understand when low- or high-dimensional representations are optimal for downstream generalisation. We next validate such framework in practical settings.

5 Validation in autoencoders and pretrained transformers

In this section, we evaluate the extent to which theoretical predictions are valid beyond the idealised setting employed in their derivation. We first consider a controlled synthetic experiment in which the representation learned by an autoencoder approximates PCA. We then relax the assumptions of linearity and spiked isotropic covariance, allowing for a more general covariance structure. Finally, we show that our predictions extend to representations learned by modern neural networks.

5.1 Autoencoders as learned representations

We train the autoencoder on a pretraining task and use the learned representations for regression on downstream tasks. This approach is motivated by the fact that, in several regimes, autoencoders with a bottleneck dimension m recover the subspace spanned by the top-m principal components of the pretraining data [81–84]. We discuss below the regimes under which this correspondence holds.

Spiked isotropic covariance. We first train autoencoders on the spiked identity covariance structures considered in the theory. For linear autoencoders, the optimal bottleneck size, defined as the value minimising the generalisation error on a downstream task, closely follows our theoretical results (compare Figure 4**(a)** with Figure 2**(d)**). We highlight that the boundaries of the phase transitions in the heatmaps precisely match the predictions from Section 4.2. Nonlinear autoencoders exhibit the same qualitative behaviour, except when the number of pretraining samples is small, see Figure 4**(b)**. This discrepancy is explained by a data-dependent transition: both linear and nonlinear autoencoders recover PCA only once sufficient pretraining data is available, see Figure 16 in Appendix F.1. In the

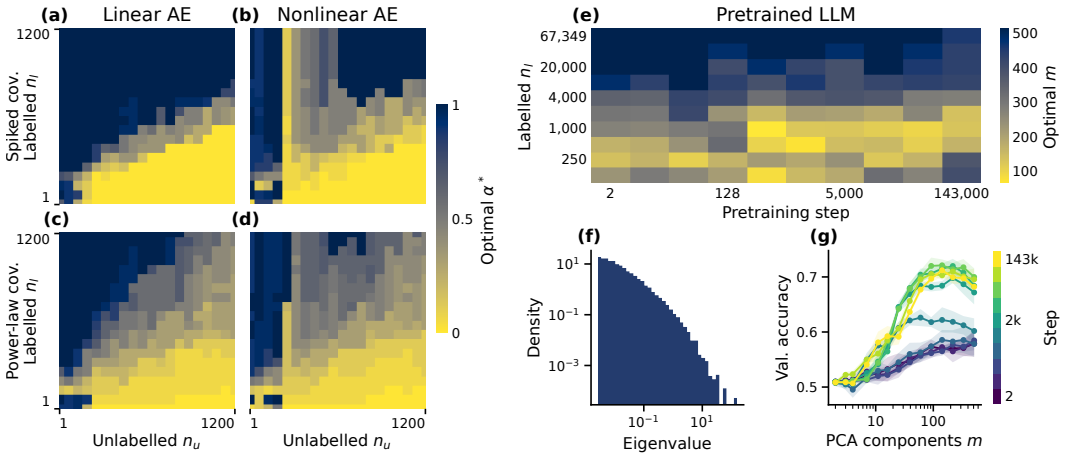


Figure 4: **Optimal representation in autoencoders and pretrained LLMs**. Optimal representation size as a function of unlabelled (n_u) and labelled (n_l) sample sizes. Linear autoencoders are considered in panels **(a,c)** and nonlinear ones in panels **(b,d)**. Autoencoders are trained on Gaussian data with either spiked identity covariance $\mathbf{\Sigma} = \mathbf{I}_p + \lambda \mathbf{v}\mathbf{v}^\top$ (panels **(a,b)**) or a power-law covariance $\mathbf{\Sigma} = \mathrm{diag}(1^{-\beta}, 2^{-\beta}, \ldots, p^{-\beta})$ with $\beta = 0.3$ (panels **(c,d)**). **(e)** In a language transformer (`PYTHIA` models), PCA on representations also improves downstream linear probing on the `sst2` task from the `GLUE` benchmark. **(f)** Eigenvalue spectrum of last-token representations from the final hidden layer. **(g)** Validation accuracy versus representation size m for a small downstream dataset ($n_l = 2000$); colors denote different checkpoints and shaded regions indicate standard deviation across seeds. Simulation details are given in Appendix H.

low-sample regime, autoencoders instead tend to memorise training samples [85–87]. Despite this, when $n_u < p$, higher-dimensional representations still yield lower errors on downstream tasks.

General covariance structure. For a power law covariance, often appearing in representations in modern neural networks [88, 89], the same qualitative behaviour persists. In fact, while phase boundaries shift, the overall structure is preserved: no compression when pretraining data is limited; small bottlenecks for large pretraining and small downstream datasets; and intermediate-sized bottlenecks for large pretraining datasets. This is illustrated in Figure 4**(c,d)**, and it is consistent with PCA-based regression under general covariance, see Figure 14 in Appendix F.1. Furthermore, when optimising for training error, the results closely match the spiked identity case (Figure 14 in Appendix F.1). At lower signal-to-noise ratios, the regime where compression is optimal expands (Figure 15 in Appendix F.1), in agreement with the theoretical analysis. Nonlinear autoencoders again exhibit deviations from PCA at low pretraining sample sizes.

Overall, our empirical results show that the theoretical predictions derived in the idealised setting extend to trained linear and nonlinear autoencoders when enough pretraining data is available, and that the qualitative phase structure persists under general covariance spectra.

5.2 Linear probing transformers

We next examine whether representations learned by modern architectures exhibit similar dimensionality effects under linear probing. We use models from the Pythia family [79], which provide access to intermediate checkpoints throughout pretraining. As these models are trained for $\sim$1.5 epochs on the Pile dataset [80, 90], the training step serves as a proxy for the number of pretraining samples n_u. The learned representations exhibit an approximate power-law structure (Figure 4**(f)**). For each checkpoint, we project representations onto the top-m principal components and train linear probes on a downstream sentiment classification task (SST-2 from GLUE [3]), selecting m to optimise validation or training accuracy. We consider two regimes: one where PCA is applied to pretraining representations, and one where it is applied to representations extracted from the downstream task itself; the latter is explored in Appendix F.2.

The dependence on dataset size is consistent with theoretical predictions, see Figure 4**(e)**. For small downstream datasets, models that are pretrained longer benefit from stronger compression, see Figure 4**(g)**. For larger downstream datasets, performance improves monotonically with m, indicating

that little or no compression is optimal (see the right-most plot of Figure 18 in Appendix F.2). For extremely small downstream datasets ($\sim$100 samples), the optimal dimensionality is high again (see the left-most plot in Figure 18 in Appendix F.2), consistent with the weak alignment between the task signal and dominant pretraining directions. Finally, tuning the representation dimensionality via PCA consistently matches or outperforms using the full feature space, see Figure 19 in Appendix F.2.

We further validate these findings across a broader set of tasks and model scales in Appendix F.3. Extending to nine GLUE datasets and two Pythia model sizes (70M and 410M parameters), we find qualitatively consistent behaviour with no systematic dependence on model scale. We also confirm the theoretical prediction that lower signal-to-noise ratio favours stronger compression: corrupting downstream training labels shifts the validation-optimal dimension downward across datasets and both model sizes (Figures 20 and 21). In summary, these results show that real-world representations exhibit a theoretically predictable qualitative phase structure when optimising dimensionality for downstream generalisation.

6 Discussion

We present a theoretical model quantifying how representation dimensionality affects training and generalisation error on downstream tasks. We demonstrate that *(i)* tuning this dimensionality significantly improves generalisation compared to full representations or naive fixed dimensionality reduction, and that *(ii)* the optimal dimensionality is determined by the amount of pretraining and downstream data, the task signal-to-noise ratio, the strength of structure in the inputs, and the alignment between task-relevant signals and dominant covariance directions. In particular, when pretraining data is abundant and downstream data is limited, the optimal representations are maximally compressed. Furthermore, we unveil a trade-off between pretraining and supervision, identifying regimes where unlabelled data yields larger gains than labelled samples. These results provide a principled explanation for why large models trained on extensive data and exhibiting low intrinsic dimensionality [11], are especially effective for fine-tuning and linear probing. In Appendix G.2, we discuss how these findings could extend to the fine-tuning setting.

The question of optimal representations is central to machine learning as well as neuroscience, where internal representations support categorisation, reasoning, planning, navigation, and decision-making [91–93]. A key property of these representations is their dimensionality. High-dimensional representations may support flexible readout across tasks [94–97], whereas low-dimensional representations are associated with abstraction, efficient coding, or structured prior knowledge [27, 98]. Our results provide a quantitative characterization of how task structure and data availability determine the optimal representation tradeoff in a simplified setting, offering predictions for how biological systems might organize representations for effective learning.

Several limitations motivate future work such as considering richer covariance models. Richer covariance models require matrix-valued rather than rank-one resolvent arguments. Another important direction is to replace the closed-form regression estimator with gradient-based training, enabling the study of optimisation-induced bias and its dependence on initialisation from pretraining. Finally, extending the framework beyond PCA-based projections to learned feature representations would allow for a closer connection to empirical observations in deep networks.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF) (10.55776/COE12), Gatsby Charitable Foundation (GAT3850 and GAT4058), Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z). For the purpose of open access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. CD was supported by the Harvard Center for Brain Science (CBS)–NTT Fellowship Program on the Physics of Intelligence.

Contributions

A.S. and R.S. initiated and conceptualized the project. V.N. and C.D. developed the theoretical analysis and derived the main results, with V.N. leading the derivations. V.N. and C.D. designed and performed the experiments. M.M. co-supervised the project and contributed to the technical development of the work, including proofs, experimental design, and interpretation of results. A.S. co-supervised the project and contributed to the conceptual framing of the project, interpretation of the theoretical results, experimental design and overall scientific direction. V.N. and C.D. wrote the manuscript. M.M. and A.S. reviewed and edited the paper. All authors discussed the results and commented on the manuscript.

References

[1] OpenAI et al. GPT-4 Technical Report. *arXiv:2303.08774*, 2024.

[2] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[7] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[8] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*. PMLR, 2020.

[9] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[10] Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. In *International Conference on Machine Learning*. PMLR, 2025.

[11] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[12] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[13] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

[14] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11, 2010.

[15] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 2010.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[17] Alex Martin. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58, 2007.

[18] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37, 2016.

[19] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 2019.

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114, 2017.

[21] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*. PMLR, 2017.

[22] Lukas Braun, Clémentine Carla Juliette Dominé, James E Fitzgerald, and Andrew M Saxe. Exact learning dynamics of deep linear networks with prior knowledge. In *Advances in Neural Information Processing Systems*, 2022.

[23] David Badre, Apoorva Bhandari, Haley Keglovits, and Atsushi Kikumoto. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38, 2021.

[24] Lara M. Boyle, Lorenzo Posani, Sarah Irfan, Steven A. Siegelbaum, and Stefano Fusi. Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron*, 112, 2024.

[25] Hristos S. Courellis, Juri Minxha, Araceli R. Cardenas, Daniel L. Kimmel, Chrystal M. Reed, Taufik A. Valiante, C. Daniel Salzman, Adam N. Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632, 2024.

[26] Karyna Mishchanchuk, Gabrielle Gregoriou, Albert Qü, Alizée Kastler, Quentin J. M. Huys, Linda Wilbrecht, and Andrew F. MacAskill. Hidden state inference requires abstract contextual representations in the ventral hippocampus. *Science*, 386, 2024.

[27] Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26, 2023.

[28] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[29] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 2018.

[30] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50, 2022.

[31] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[32] Denny Wu and Ji Xu. On the optimal weighted \ell_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[33] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117, 2020.

[34] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116, 2019.

[35] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132, 2020.

[36] Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *Journal of Machine Learning Research*, 27, 2026.

[37] Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In *International Conference on Machine Learning*, 2025.

[38] Edwige Cyffers, Alireza Mirrokni, and Marco Mondelli. Optimal regularization for performative learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[39] Parham Rezaei, Filip Kovacevic, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2026.

[40] M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations*, 2025.

[41] Germain Kolossov, Andrea Montanari, and Pulkit Tandon. Towards a statistical theory of data selection under weak supervision. In *International Conference on Learning Representations*, 2024.

[42] Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *Advances in Neural Information Processing Systems*, 37, 2024.

[43] Ian T. Jolliffe. A Note on the Use of Principal Components in Regression. *Applied Statistics*, 31, 1982.

[44] Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[45] Alden Green and Elad Romanov. The high-dimensional asymptotics of principal component regression. *The Annals of Statistics*, 53, 2025.

[46] William F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60, 1965.

[47] Paramveer S. Dhillon, Dean P. Foster, Sham M. Kakade, and Lyle H. Ungar. A risk comparison of ordinary least squares vs ridge regression. *Journal of Machine Learning Research*, 14, 2013.

[48] Jianfeng Yao, Zhidong Bai, and Shui-Rong Zheng. *Large sample covariance matrices and high-dimensional data analysis*. Cambridge series in statistical and probabilistic mathematics. Cambridge university press, 2015.

[49] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22, 2010.

[50] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, PP, 2020.

[51] Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 2022.

[52] Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2019.

[53] Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: a theory of transfer learning in high-dimensional regression. In *Forty-second International Conference on Machine Learning*, 2025.

[54] Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Re, and Weijie J Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26, 2025.

[55] Yanke Song, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv:2406.13944*, 2024.

[56] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[57] Eyar Azar and Boaz Nadler. Semi-supervised sparse gaussian classification: Provable benefits of unlabeled data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[58] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv:1902.09229*, 2019.

[59] Tianyi Zhang and Tatsunori B. Hashimoto. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[60] Taj Jones-Mccormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267. PMLR, 2025.

[61] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*, 2018.

[62] Bin Wang, W. Jeffrey Johnston, and Stefano Fusi. A mathematical theory for understanding when abstract representations emerge in neural networks. *arXiv:2510.09816*, 2026.

[63] Nicolas Anguita, Francesco Locatello, Andrew M. Saxe, Marco Mondelli, Flavia Mancini, Samuel Lippl, and Clementine Domine. A theory of how pretraining shapes inductive bias in fine-tuning. *arXiv:2602.20062*, 2026.

[64] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29, 2001.

[65] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33, 2005.

[66] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1, 1967.

[67] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York, 2010.

[68] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97, 2006.

[69] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 2007.

[70] Zhidong Bai and Jianfeng Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106, 2012.

[71] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[72] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[73] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[74] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[75] Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue M Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *International Conference on Machine Learning*, 2024.

[76] Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

[77] Rishi Sonthalia, Michael Murray, and Guido F. Montúfar. Low rank gradients and where to find them. In *Advances in Neural Information Processing Systems*, 2025.

[78] Daniel Gedon, Antônio H. Ribeiro, and Thomas B. Schön. No double descent in principal component regression: A high-dimensional analysis. In *International Conference on Machine Learning*, 2024.

[79] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

[80] Stella Biderman, Kieran Bicheno, and Leo Gao. Datasheet for the pile. *arXiv:2201.07311*, 2022.

[81] Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. *Journal of Statistical Mechanics: Theory and Experiment*, 2023, 2023.

[82] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2, 1989.

[83] Elad Plaut. From Principal Subspaces to Principal Components with Linear Autoencoders. *arXiv:1804.10253*, 2018.

[84] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59, 1988.

[85] Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[86] Yibo Jiang and Cengiz Pehlevan. Associative memory in iterated overparameterized sigmoid autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[87] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized Neural Networks Implement Associative Memory. *Proceedings of the National Academy of Sciences*, 117, 2020.

[88] Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Aaron Richards. α-ReQ : Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems*, 2022.

[89] Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake Aaron Richards. Tracing the representation geometry of language models from pretraining to post-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[90] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv:2101.00027*, 2020.

[91] Mark G Stokes, Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan. Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78, 2013.

[92] Aldo Genovesio, Peter J Brasted, Andrew R Mitz, and Steven P Wise. Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47, 2005.

[93] Keisetsu Shima and Jun Tanji. Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of neurophysiology*, 84, 2000.

[94] David Raposo, Matthew T Kaufman, and Anne K Churchland. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17, 2014.

[95] Evelyn Tang, Marcelo G Mattar, Chad Giusti, David M Lydon-Staley, Sharon L Thompson-Schill, and Danielle S Bassett. Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature neuroscience*, 22, 2019.

[96] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497, 2013.

[97] Silvia Bernardi, Marcus K Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183, 2020.

[98] Kay Tye, Earl Miller, Felix Taschbach, Marcus Benna, Mattia Rigotti, and Stefano Fusi. Mixed selectivity: Cellular computations for complexity. *Neuron*, 112, 2024.

[99] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2026.

[100] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81, 2011.

[101] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 2019.

[102] Clémentine CJ Dominé, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro AM Mediano, and Andrew M Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv:2409.14623*, 2024.

[103] Yoonsoo Nam, Seok Hyeong Lee, Clémentine Carla Juliette Dominé, Yeachan Park, Charles London, Wonyl Choi, Niclas Alexander Göring, and Seungjai Lee. Position: Solve layerwise linear models first to understand neural dynamical phenomena (Neural collapse, emergence, Lazy/Rich regime, and grokking). In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[104] Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *Advances in Neural Information Processing Systems*, 37, 2024.

Appendix

Appendix contents

A Assumptions and proofs 19
A.1 Assumptions 19
A.2 Main ideas of the proofs 20
A.3 Proof for the estimation error 21
A.4 Proof for the generalisation error 25
A.5 Proof for the training error 27
A.6 Auxiliary lemmas 28
B Phase transitions 34
B.1 Phase transition for $\gamma_l > 1$ 34
B.2 Phase transition for $\gamma_l < 1$ 37
C Infinite data limits 38
D Regularized regression 40
D.1 Decomposition of the ridge estimator 42
D.2 Asymptotic estimation error for ridge regression 43
D.3 Asymptotic generalisation error for ridge regression 44
D.4 Asymptotic training error for ridge regression 45
E Theory-experiment validation 45
E.1 Error curves 46
E.2 Optimal dimensionality 48
F Additional experiments 49
F.1 Additional autoencoder experiments 49
F.2 Additional transformer experiments 52
F.3 More datasets 53
G Further discussion 56
G.1 Relation to classical principal component regression 56
G.2 Fine-tuning 60
H Implementation details 61
H.1 Figure 1 61
H.2 Figure 2 61
H.3 Figure 3 62
H.4 Figure 4 62
H.5 Compute 64
H.6 LLM Usage 64

A Assumptions and proofs

A.1 Assumptions

Assumption 1 (High-dimensional asymptotic regime)**.** *Let the number of features be p, and n_l, n_u, m be such that*

$$\frac{p}{n_l} \to \gamma_l \in (0, \infty), \qquad \frac{p}{n_u} \to \gamma_u \in (0, \infty), \qquad \frac{m}{p} \to \alpha \in [0, 1].$$

Furthermore, we define $\gamma_{\text{eff}} := \frac{m}{n_l} \to \alpha\gamma_l \in [0, \infty)$. We assume $1 \leq m \leq p$. In particular, $\alpha = 0$ means $m/p \to 0$ while retaining at least the leading empirical principal component; it does not mean $m = 0$. This includes both fixed m and sequences $m \to \infty$ with $m = o(p)$. For $\alpha = 0$, $\gamma_{\text{eff}} = 0$.

In words, α is the fraction of dimensions to keep, γ_u and γ_l are the aspect ratios the prior and downstream datasets, and γ_{eff} is the effective aspect ratio of the compressed downstream dataset.

Assumption 2 (The spiked covariance models)**.** *For each p, the pretraining and downstream population covariance matrices are*

$$\mathbf{\Sigma}_u = \mathbf{I}_p + (\lambda_u - 1)\mathbf{v}\mathbf{v}^\top, \qquad \mathbf{\Sigma}_l = \mathbf{I}_p + (\lambda_l - 1)\mathbf{q}\mathbf{q}^\top,$$

where $\lambda_u, \lambda_l > 1$, $\mathbf{v}, \mathbf{q} \in \mathbb{R}^p$ are deterministic unit vectors.

Note that the eigendecomposition of $\mathbf{\Sigma}_u$ is

$$\mathbf{\Sigma}_u = \mathbf{V}\text{diag}(\lambda_u, 1, \ldots, 1)\mathbf{V}^\top,$$

with $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p] \in \mathbb{R}^{p\times p}$. As we often refer directly to the spike eigenvector, we keep the simpler notation $\mathbf{v}_1 = \mathbf{v}$. Similarly for $\mathbf{\Sigma}_l$ where we use $\mathbf{q}_1 = \mathbf{q}$.

Assumption 3 (Data generation)**.** *The matrices $\mathbf{X}_l \in \mathbb{R}^{n_l\times p}$ and $\mathbf{X}_u \in \mathbb{R}^{n_u\times p}$ are independent of each other. Each matrix has independent, identically distributed rows, with*

$$\mathbf{x}_l \sim \mathcal{N}(0, \mathbf{\Sigma}_l), \qquad \mathbf{x}_u \sim \mathcal{N}(0, \mathbf{\Sigma}_u).$$

Equivalently, let $\mathbf{Z}_u \in \mathbb{R}^{n_u\times p}$ and $\mathbf{Z}_l \in \mathbb{R}^{n_l\times p}$ be mutually independent matrices with independent $\mathcal{N}(0, 1)$ entries. Then $\mathbf{X}_u = \mathbf{Z}_u\mathbf{\Sigma}_u^{1/2}$ and $\mathbf{X}_l = \mathbf{Z}_l\mathbf{\Sigma}_l^{1/2}$. The label-noise vector $\xi \sim \mathcal{N}(0, \sigma^2\mathbf{I}_{n_l})$ is independent of both datasets.

Assumption 4 (Task weight vector)**.** *The deterministic task vector satisfies*

$$\|\mathbf{w}^\star\|^2 \longrightarrow \bar{w}^* \in (0, \infty).$$

The deterministic unit vectors $\mathbf{v}, \mathbf{q}$ and the task vector have convergent signed normalized overlaps:

$$\frac{\mathbf{w}^{\star\top}\mathbf{v}}{\|\mathbf{w}^\star\|} \to r_v, \qquad \frac{\mathbf{w}^{\star\top}\mathbf{q}}{\|\mathbf{w}^\star\|} \to r_q, \qquad \mathbf{v}^\top\mathbf{q} \to r.$$

Furthermore, let the alignment between the true signal $\mathbf{w}^\star$ and the pretraining covariance structure $\mathbf{\Sigma}_u$ be given by the spectral measure

$$\hat{G}_{\mathbf{w}^\star}(\tau) = \frac{1}{\|\mathbf{w}^\star\|^2}\sum_{i=1}^{p}(\mathbf{w}^{\star\top}\mathbf{v}_i)^2\mathbf{1}\{\lambda_i(\mathbf{\Sigma}_u) \leq \tau\}.$$

Then, there exists a limiting spectrum $G_{\mathbf{w}^\star}$ such that $\hat{G}_{\mathbf{w}^\star} \Rightarrow G_{\mathbf{w}^\star}$ almost surely. Similarly, we define the alignment between the downstream spike $\mathbf{q}$ and the pretraining covariance $\mathbf{\Sigma}_u$

$$\hat{G}_{\mathbf{q}}(\tau) = \sum_{i=1}^{p}(\mathbf{q}^\top\mathbf{v}_i)^2\mathbf{1}\{\lambda_i(\mathbf{\Sigma}_u) \leq \tau\}.$$

with a limiting spectrum $\hat{G}_{\mathbf{q}} \Rightarrow G_{\mathbf{q}}$.

Sample covariance eigenspectrum. We further note that sample covariance matrices $\mathbf{S}_u = \frac{1}{n_u}\mathbf{X}_u^\top \mathbf{X}_u$ and $\mathbf{S}_l = \frac{1}{n_l}\mathbf{X}_l^\top \mathbf{X}_l$ have empirical spectral measures

$$\hat{F}^{\mathbf{S}_u}(\theta) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}\{\lambda_i(\mathbf{S}_u) \le \theta\}, \qquad \hat{F}^{\mathbf{S}_l}(\theta) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}\{\lambda_i(\mathbf{S}_l) \le \theta\},$$

which converge weakly almost surely to Marcenko-Pastur laws [66, 67] with aspect ratios γ_u and γ_l respectively:

$$\hat{F}^{\mathbf{S}_u} \Rightarrow F_{\gamma_u}, \qquad \hat{F}^{\mathbf{S}_l} \Rightarrow F_{\gamma_l}.$$

For completeness, we state the Marcenko-Pastur density with an aspect ratio γ:

$$\begin{aligned} f_\gamma(\lambda) &= \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda\gamma} + \mathbf{1}_{\gamma>1}\left(1 - \frac{1}{\gamma}\right)\delta(\lambda), \\ \lambda_+ &= (1+\sqrt{\gamma})^2, \\ \lambda_- &= (1-\sqrt{\gamma})^2, \end{aligned} \tag{7}$$

with λ_+, λ_- denoting the edges of the support of the distribution.

A.2 Main ideas of the proofs

The key idea of each proof is to decompose the model's estimator $\hat{\mathbf{w}}$ into a signal term, a bias term arising from the component of $\mathbf{w}^\star$ outside the relevant subspace, and two variance terms arising from noise and model mismatch. The Gaussian structure of the data allows us to express the component of $\mathbf{X}_l$ orthogonal to the prior subspace $\mathbf{U}_m$ as a linear function of $\mathbf{X}_l\mathbf{U}_m$ plus independent noise $\mathbf{E}$. This reduces the analysis to understanding various projections onto the row space of spiked Gaussian matrices.

Selection of the retained PCA subspace. The columns of $\mathbf{U}_m$ are eigenvectors associated with the m largest eigenvalues of $\mathbf{S}_u$. If $m > \operatorname{rank}(\mathbf{S}_u)$, we retain all positive-eigenvalue directions and complete the subspace with a Haar-uniform $(m - \operatorname{rank}(\mathbf{S}_u))$-dimensional subspace of $\ker(\mathbf{S}_u)$, conditional on $\mathbf{X}_u$. The auxiliary randomness used for this completion is independent of the downstream data and label noise.

Conditional Gaussian decomposition under prior subspace projections. Let $\mathbf{X}_l \in \mathbb{R}^{n_l \times p}$ be a random matrix whose rows $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_l)$ are i.i.d., and let $\mathbf{U}_m \in \mathbb{R}^{p\times m}$ be a random matrix with orthonormal columns. Let $\mathbf{U}_\perp \in \mathbb{R}^{p\times(p-m)}$ be such that $[\mathbf{U}_m\ \mathbf{U}_\perp]$ is an orthogonal matrix. Define

$$\mathbf{X}_m = \mathbf{X}_l\mathbf{U}_m \in \mathbb{R}^{n_l\times m}, \qquad \mathbf{X}_\perp = \mathbf{X}_l\mathbf{U}_\perp \in \mathbb{R}^{n_l\times(p-m)}.$$

Then, conditioning on $\mathbf{U}_m$, the rows of $(\mathbf{X}_m, \mathbf{X}_\perp)$ are jointly Gaussian with covariance

$$\begin{bmatrix} \mathbf{\Sigma}_{mm} & \mathbf{\Sigma}_{m\perp} \\ \mathbf{\Sigma}_{\perp m} & \mathbf{\Sigma}_{\perp\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_m^\top\mathbf{\Sigma}_l\mathbf{U}_m & \mathbf{U}_m^\top\mathbf{\Sigma}_l\mathbf{U}_\perp \\ \mathbf{U}_\perp^\top\mathbf{\Sigma}_l\mathbf{U}_m & \mathbf{U}_\perp^\top\mathbf{\Sigma}_l\mathbf{U}_\perp \end{bmatrix}.$$

Moreover, due to the conditional distribution, $\mathbf{X}_\perp$ given $\mathbf{X}_m$ satisfies

$$\mathbf{X}_\perp = \mathbf{X}_m\mathbf{\Sigma}_{mm}^{-1}\mathbf{\Sigma}_{m\perp} + \mathbf{E}, \tag{8}$$

where $\mathbf{E}$ is independent of $\mathbf{X}_m$ and its rows are i.i.d. with distribution $\mathbf{e}_i \sim \mathcal{N}\big(0,\ \bar{\mathbf{\Sigma}}\big)$, with $\bar{\mathbf{\Sigma}} := \mathbf{\Sigma}_{\perp\perp} - \mathbf{\Sigma}_{\perp m}\mathbf{\Sigma}_{mm}^{-1}\mathbf{\Sigma}_{m\perp}$. Under the spiked covariance model of Assumption 2, the population covariance $\mathbf{\Sigma}_{mm}$ of $\mathbf{X}_m$ equals

$$\mathbf{\Sigma}_{mm} = \mathbf{I}_m + (\tilde{\lambda}_l - 1)\tilde{\mathbf{q}}\tilde{\mathbf{q}}^\top, \tag{9}$$

where $\tilde{\lambda}_l = 1 + (\lambda_l - 1)\|\mathcal{P}_m \mathbf{q}\|^2$ and $\tilde{\mathbf{q}} = \frac{\mathbf{U}_m^\top \mathbf{q}}{\|\mathcal{P}_m \mathbf{q}\|}$. Furthermore, after some algebraic manipulations using the Sherman-Morrison formula, it follows that

$$\boldsymbol{\Sigma}_{mm}^{-1} = \mathbf{I}_m - \frac{\tilde{\lambda}_l - 1}{\tilde{\lambda}_l} \tilde{\mathbf{q}}\tilde{\mathbf{q}}^\top, \tag{10}$$

$$\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp} = \frac{(\lambda_l - 1)\|\mathcal{P}_m \mathbf{q}\|}{\tilde{\lambda}_l} \tilde{\mathbf{q}}\mathbf{q}^\top \mathbf{U}_\perp, \tag{11}$$

$$\bar{\boldsymbol{\Sigma}} = \mathbf{I}_{p-m} + \frac{\lambda_l - 1}{\tilde{\lambda}_l} \mathbf{U}_\perp^\top \mathbf{q}\mathbf{q}^\top \mathbf{U}_\perp. \tag{12}$$

Decomposing the model's weight estimate. From the definition of $\hat{\mathbf{w}}$ in Equation (1) and the decomposition in Equation (8), we can write:

$$\begin{aligned}\hat{\mathbf{w}} &= (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathcal{P}_m \mathbf{w}^\star + (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathcal{P}_\perp \mathbf{w}^\star + (\mathbf{X}_l \mathcal{P}_m)^\dagger \xi \\ &= (\mathbf{X}_m \mathbf{U}_m^\top)^\dagger \mathbf{X}_m \mathbf{U}_m^\top \mathbf{w}^\star + (\mathbf{X}_m \mathbf{U}_m^\top)^\dagger \mathbf{X}_m \boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp} \mathbf{U}_\perp^\top \mathbf{w}^\star + (\mathbf{X}_m \mathbf{U}_m^\top)^\dagger \mathbf{E}\mathbf{U}_\perp^\top \mathbf{w}^\star + (\mathbf{X}_m \mathbf{U}_m^\top)^\dagger \xi.\end{aligned}$$

Then, using Equation (11) leads to

$$\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp}\mathbf{U}_\perp^\top \mathbf{w}^\star = \underbrace{\frac{(\lambda_l - 1)(\mathbf{w}^{\star\top} \mathcal{P}_\perp \mathbf{q})}{\tilde{\lambda}_l}}_{A} \mathbf{U}_m^\top \mathbf{q} \tag{13}$$

and by defining

$$\Pi_m := \mathbf{X}_m^\dagger \mathbf{X}_m \in \mathbb{R}^{m \times m} \qquad \Pi := (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathcal{P}_m = \mathbf{U}_m \Pi_m \mathbf{U}_m^\top \in \mathbb{R}^{p \times p},$$

we obtain

$$\begin{aligned}\hat{\mathbf{w}} = &\underbrace{\Pi \mathbf{w}^\star}_{\text{signal from } \mathcal{P}_m \text{ subspace}} + \underbrace{A \cdot \Pi \mathbf{q}}_{\text{signal from } \mathcal{P}_\perp \text{ subspace that leaked into } \mathcal{P}_m} \\ &+ \underbrace{(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{E} \mathbf{w}^\star{}_\perp}_{\text{variance from model mismatch}} + \underbrace{(\mathbf{X}_l \mathcal{P}_m)^\dagger \xi}_{\text{variance from label noise}},\end{aligned} \tag{14}$$

where $\mathbf{w}^\star{}_\perp := \mathbf{U}_\perp^\top \mathbf{w}^\star$. Note that the component of $\mathbf{w}^\star$ not involving label noise is

$$(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star = \Pi \mathbf{w}^\star + A\Pi \mathbf{q} + (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{E} \mathbf{w}^\star{}_\perp. \tag{15}$$

The model's weight estimate decomposes into the four parts discussed below.

- $\Pi \mathbf{w}^\star$: part of $\mathbf{w}^\star$ that lies in the subspace formed by $\mathcal{P}_m$ and is identifiable from the available data. It is equal to $\mathcal{P}_m \mathbf{w}^\star$ if $m < n_l$ (i.e., $\gamma_{\text{eff}} < 1$), since in that case $\Pi_m = \mathbf{I}_m$ (see Lemma 6); if $m > n_l$ (i.e., $\gamma_{\text{eff}} > 1$), this term causes bias due to the finite sample size.
- $A\Pi \mathbf{q}$: a part of $\mathbf{w}^\star$ that lies outside the subspace $\mathcal{P}_m$ can leak into $\mathcal{P}_m$, causing the model to falsely reconstruct it. The reason is the existence of the population covariance spike which makes the rows of $\mathbf{X}_l \mathcal{P}_\perp$ and $\mathbf{X}_l \mathcal{P}_m$ correlated, so the part outside the subspace leaks back into the subspace through this covariance structure.
- $(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{E} \mathbf{w}^\star{}_\perp$: the part of the $\mathbf{w}^\star$ that lies outside the subspace $\mathcal{P}_m$ and does not leak into $\mathcal{P}_m$ behaves like noise, eventually causing a double-descent-like curve that scales with $\|\mathbf{w}^\star\|^2$.
- $(\mathbf{X}_l \mathcal{P}_m)^\dagger \xi$: label noise variance, causing another double-descent-like curve that scales with σ^2.

A.3 Proof for the estimation error

Proof of Theorem 1. The proof proceeds in three stages, exploiting the conditional independence structure of the data. First, we use the joint Gaussianity of $\mathbf{X}_m$ and $\mathbf{X}_\perp$ to rewrite $\mathbf{X}_\perp$ in terms of $\mathbf{X}_m$ and an independent noise matrix $\mathbf{E} \in \mathbb{R}^{n_l \times (p-m)}$ (see Section A.2). We then average over $\mathbf{E}$ conditioned on $\mathbf{X}_m$ and $\mathbf{U}_m$, then average over $\mathbf{X}_m$ conditioned on $\mathbf{U}_m$, and finally over $\mathbf{U}_m$.

Step 1: Decomposition of the error. The estimation error splits into bias and variance:

$$\begin{aligned}\mathbb{E}_\xi\left[\|\mathbf{w}^\star-\hat{\mathbf{w}}\|^2\right] &= \left\|\left(\mathbf{I}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{X}_l\right)\mathbf{w}^\star\right\|^2+\mathbb{E}_\xi\left[\xi^\top(\mathbf{X}_l\mathcal{P}_m)^{\dagger\top}(\mathbf{X}_l\mathcal{P}_m)^\dagger\xi\right]\\ &\quad-2\mathbb{E}_\xi\left[\xi^\top(\mathbf{X}_l\mathcal{P}_m)^{\dagger\top}\left(\mathbf{I}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{X}_l\right)\mathbf{w}^\star\right]\\ &= \left\|\left(\mathbf{I}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{X}_l\right)\mathbf{w}^\star\right\|^2+\mathbb{E}_\xi\left[\mathrm{Tr}\left[\xi\xi^\top(\mathbf{X}_l\mathcal{P}_m)^{\dagger\top}(\mathbf{X}_l\mathcal{P}_m)^\dagger\right]\right]\\ &= \left\|\left(\mathbf{I}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{X}_l\right)\mathbf{w}^\star\right\|^2+\sigma^2\mathrm{Tr}\left[(\mathcal{P}_m\mathbf{X}_l^\top\mathbf{X}_l\mathcal{P}_m)^\dagger\right].\end{aligned}$$

Substituting the result in (15) further decomposes the bias:

$$\begin{aligned}\frac{1}{\|\mathbf{w}^\star\|^2}\left\|\left(\mathbf{I}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{X}_l\right)\mathbf{w}^\star\right\|^2 &= \frac{1}{\|\mathbf{w}^\star\|^2}\|\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q}-(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star\|^2\\ &= \frac{1}{\|\mathbf{w}^\star\|^2}\|\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q}\|^2+\frac{1}{\bar{w}^*}\left\|(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star\right\|^2\\ &\quad-2\underbrace{\frac{1}{\|\mathbf{w}^\star\|^2}(\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q})^\top(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star}_{\text{cross term } T}.\end{aligned}$$

Step 2: E-independent term. The first term $\|\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q}\|^2$ does not depend on $\mathbf{E}$, so its conditional expectation on $\mathbf{E}$ is itself. We expand using the property of the projection matrix Π that $\Pi^2=\Pi$:

$$\frac{1}{\|\mathbf{w}^\star\|^2}\|\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q}\|^2=1-\frac{\|\Pi\mathbf{w}^\star\|^2}{\|\mathbf{w}^\star\|^2}+\frac{A^2}{\|\mathbf{w}^\star\|^2}\|\Pi\mathbf{q}\|^2, \tag{16}$$

where the cross term $-2A\mathbf{w}^{\star\top}\Pi\mathbf{q}+2A\mathbf{w}^{\star\top}\Pi\Pi\mathbf{q}=0$. Substituting $A=\frac{\lambda_l-1}{\tilde{\lambda}_l}(\mathbf{q}^\top\mathcal{P}_\perp\mathbf{w}^\star)$ gives

$$1-\frac{\|\Pi\mathbf{w}^\star\|^2}{\|\mathbf{w}^\star\|^2}+\frac{(\lambda_l-1)^2}{\tilde{\lambda}_l^2}\frac{(\mathbf{q}^\top\mathcal{P}_\perp\mathbf{w}^\star)^2}{\|\mathbf{w}^\star\|^2}\|\Pi\mathbf{q}\|^2\Rightarrow 1-\bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}+\frac{(\lambda_l-1)^2}{\bar{\lambda}_l^2}\bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}\bar{\pi}_{\mathbf{q},\mathbf{q}} \text{ almost surely,} \tag{17}$$

which matches the first two terms of the bias. The projections $\frac{\|\Pi\mathbf{w}^\star\|^2}{\|\mathbf{w}^\star\|^2}$ and $\|\Pi\mathbf{q}\|^2$ converge almost surely to $\bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star},\bar{\pi}_{\mathbf{q},\mathbf{q}}$ in Equation (41) according to Lemma 6. The quantity $\frac{\mathbf{q}^\top\mathcal{P}_\perp\mathbf{w}^\star}{\|\mathbf{w}^\star\|}$ converges almost surely to $\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}$ in Equation (36) according to Lemma 5, and $\tilde{\lambda}_l=1+(\lambda_l-1)\|\mathcal{P}_m\mathbf{q}\|^2$ converges to $\bar{\lambda}_l$ almost surely by the continuous mapping theorem and the almost sure convergence of $\|\mathcal{P}_m\mathbf{q}\|^2\to\bar{p}_{\mathbf{q},\mathbf{q}}$ from Equation (32) (see Lemma 4). By the continuous mapping theorem, the combined expression in (17) converges almost surely.

Step 3: Cross term linear in E. We write the cross term as

$$T=\mathbf{c}^\top\mathbf{E}\mathbf{d},$$

with

$$\begin{aligned}\mathbf{c}&:=\frac{1}{\|\mathbf{w}^\star\|}(\mathbf{X}_l\mathcal{P}_m)^{\dagger\top}(\mathbf{w}^\star-\Pi\mathbf{w}^\star-A\Pi\mathbf{q})\in\mathbb{R}^{n_l},\\ \mathbf{d}&:=\frac{1}{\|\mathbf{w}^\star\|}\mathbf{U}_\perp^\top\mathbf{w}^\star\in\mathbb{R}^{p-m}.\end{aligned}$$

Conditioned on $(\mathbf{X}_m,\mathbf{U}_m)$, the vectors $\mathbf{c}$ and $\mathbf{d}$ are deterministic and independent of $\mathbf{E}$. Since the rows of $\mathbf{E}$ are i.i.d. Gaussian with covariance $\bar{\mathbf{\Sigma}}$, it follows that

$$T\mid(\mathbf{X}_m,\mathbf{U}_m)\sim\mathcal{N}\big(0,\,(\mathbf{d}^\top\bar{\mathbf{\Sigma}}\mathbf{d})\,\|\mathbf{c}\|^2\big).$$

We now bound the variance. First,

$$\mathbf{d}^\top\bar{\mathbf{\Sigma}}\mathbf{d}=\frac{\|\mathcal{P}_\perp\mathbf{w}^\star\|^2}{\|\mathbf{w}^\star\|^2}+\frac{\lambda_l-1}{\tilde{\lambda}_l}\frac{(\mathbf{w}^{\star\top}\mathcal{P}_\perp\mathbf{q})^2}{\|\mathbf{w}^\star\|^2},$$

which by the almost sure limits of $\|\mathcal{P}_\perp\mathbf{w}^\star\|^2$, $\tilde{\lambda}_l$, and $\mathbf{w}^{\star\top}\mathcal{P}_\perp\mathbf{q}$, converges almost surely to

$$\bar{\sigma}^2_{\text{eff}}=\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star}+\frac{\lambda_l-1}{\bar{\lambda}_l}\bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}.$$

In particular, there exists a deterministic constant C_1 such that for all sufficiently large p,

$$\mathbf{d}^\top \bar{\mathbf{\Sigma}} \mathbf{d} \le C_1 \quad \text{almost surely.}$$

It therefore remains to control $\|\mathbf{c}\|^2$. We have

$$\|\mathbf{c}\|^2 \le \frac{1}{\|\mathbf{w}^\star\|^2} \frac{\|\mathbf{w}^\star - \Pi \mathbf{w}^\star - A\Pi \mathbf{q}\|^2}{\sigma_{\min}(\mathbf{X}_m)^2}.$$

By Step 2, the numerator is bounded almost surely for all sufficiently large p. For the denominator, writing $\mathbf{X}_m = \mathbf{G}\mathbf{\Sigma}_{mm}^{1/2}$ (see Lemma 1) gives

$$\sigma_{\min}(\mathbf{X}_m) \ge \sigma_{\min}(\mathbf{G}) \sigma_{\min}(\mathbf{\Sigma}_{mm}^{1/2}).$$

Note that $\sigma_{\min}(\mathbf{\Sigma}_{mm}^{1/2}) = \min\{1, \sqrt{\tilde{\lambda}_l}\}$ for any p. By the singular value inequality for Gaussian random matrices (see e.g. Exercise 7.13 in [99]), for all $t > 0$,

$$\mathbb{P}\big(\sigma_{\min}(\mathbf{G}) \le |\sqrt{n_l} - \sqrt{m}| - t\big) \le 2e^{-t^2/2}.$$

Since $\gamma_{\text{eff}} = m/n_l \neq 1$, we have

$$|\sqrt{n_l} - \sqrt{m}| = \sqrt{p} \frac{|1 - \sqrt{\gamma_{\text{eff}}}|}{\sqrt{\gamma_l}} = \sqrt{p}\, c_\gamma.$$

Choosing $t = \frac{1}{2} c_\gamma \sqrt{p}$ and setting $c = c_\gamma^2/8$, we obtain

$$\mathbb{P}\left(\sigma_{\min}(\mathbf{G}) \le \frac{1}{2} c_\gamma \sqrt{p}\right) \le 2 \exp\left(-\frac{c_\gamma^2}{8} p\right).$$

Noting that $\tilde{\lambda}_l \le \lambda_l$, and setting $a_\gamma := \frac{1}{4} c_\gamma^2 \min\{1, \lambda_l\}$, $b_\gamma := \frac{1}{8} c_\gamma^2$, we have

$$\mathbb{P}\big(\sigma_{\min}(\mathbf{X}_m)^2 \le a_\gamma p\big) \le 2e^{-b_\gamma p}.$$

Therefore, there exists $C_2 > 0$ such that

$$\mathbb{P}\left(\|\mathbf{c}\|^2 \le \frac{C_2}{p}\right) \ge 1 - 2e^{-b_\gamma p}.$$

Define the event

$$\mathcal{E}_p := \left\{ \|\mathbf{c}\|^2 \le \frac{C_2}{p} \right\}.$$

Then

$$\mathbb{P}(\mathcal{E}_p^c) \le 2e^{-b_\gamma p}.$$

For any $\delta > 0$,

$$\mathbb{P}(|T| \ge \delta) = \mathbb{P}(|T| \ge \delta, \mathcal{E}_p^c) + \mathbb{P}(|T| \ge \delta, \mathcal{E}_p) \le \mathbb{P}(\mathcal{E}_p^c) + \mathbb{P}(|T| \ge \delta, \mathcal{E}_p).$$

Noting that there exists C_1 s.t. $\mathbf{d}^\top \bar{\mathbf{\Sigma}} \mathbf{d} \le C_1$ for any large p, on $\mathcal{E}_p$,

$$\mathrm{Var}(T \mid \mathbf{X}_m, \mathbf{U}_m) \le \frac{C_1 C_2}{p}.$$

Thus, on $\mathcal{E}_p$, we have

$$\mathbb{P}(|T| \ge \delta \mid \mathbf{X}_m, \mathbf{U}_m) \le 2 \exp(-c_\delta p),$$

and since this holds for any $\mathbf{X}_m, \mathbf{U}_m$, by the law of total probability, we have

$$\mathbb{P}(|T| \ge \delta, \mathcal{E}_p) \le 2e^{-c_\delta p}.$$

Therefore,

$$\mathbb{P}(|T| \ge \delta) \le 2e^{-b_\gamma p} + 2e^{-c_\delta p} = Ce^{-cp}.$$

Since $\sum_p Ce^{-cp} < \infty$, the Borel–Cantelli lemma implies

$$T \xrightarrow{a.s.} 0.$$

Step 4: Average of the trace terms. The variance term contains

$$\mathrm{Tr}[(\mathcal{P}_m \mathbf{X}_l^\top \mathbf{X}_l \mathcal{P}_m)^\dagger] = \mathrm{Tr}[(\mathbf{X}_m^\top \mathbf{X}_m)^\dagger].$$

We note that the term $\mathrm{Tr}[(\mathbf{X}_m \mathbf{X}_m^\top)^\dagger]$ is equivalent to the above, since the non-zero eigenvalues of $\mathbf{X}_m^\top \mathbf{X}_m$ and $\mathbf{X}_m \mathbf{X}_m^\top$ are the same. By Lemma 2, the trace term has a limiting value $\frac{\min\{1,\gamma_{\text{eff}}\}}{|\gamma_{\text{eff}}-1|}$ almost surely.

Step 5: Quadratic term in E. Define

$$Q := \mathbf{w}^{\star\top} \mathbf{U}_\perp \mathbf{E}^\top (\mathbf{X}_m \mathbf{X}_m^\top)^\dagger \mathbf{E} \mathbf{U}_\perp^\top \mathbf{w}^\star.$$

Let

$$\mathbf{a} := \mathbf{U}_\perp^\top \mathbf{w}^\star, \qquad \mathbf{A} := (\mathbf{X}_m \mathbf{X}_m^\top)^\dagger.$$

Conditioned on $(\mathbf{X}_m, \mathbf{U}_m)$, the vector $\mathbf{a}$ and the matrix $\mathbf{A}$ are fixed and independent of $\mathbf{E}$. Since the rows of $\mathbf{E}$ are i.i.d. Gaussian with covariance $\bar{\mathbf{\Sigma}}$, the vector

$$\mathbf{c} := \mathbf{E}\mathbf{a}$$

is Gaussian with

$$\mathbf{c} \mid (\mathbf{X}_m, \mathbf{U}_m) \sim \mathcal{N}\big(0,\, (\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}) \mathbf{I}_{n_l}\big).$$

Therefore

$$Q = \mathbf{c}^\top \mathbf{A} \mathbf{c}.$$

First, the conditional expectation is

$$\mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] = (\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}) \mathrm{Tr}(\mathbf{A}).$$

Here

$$\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a} = \mathbf{w}^{\star\top} \mathbf{U}_\perp \bar{\mathbf{\Sigma}} \mathbf{U}_\perp^\top \mathbf{w}^\star = \|\mathbf{w}^\star\|^2 \sigma_{\text{eff}}^2.$$

By the deterministic equivalents used above,

$$\|\mathbf{w}^\star\|^2 \sigma_{\text{eff}}^2 \xrightarrow{a.s.} \bar{w}^* \bar{\sigma}_{\text{eff}}^2.$$

Moreover, by the pseudo-inverse trace computation in Step 4,

$$\mathrm{Tr}\big[(\mathbf{X}_m \mathbf{X}_m^\top)^\dagger\big] \xrightarrow{a.s.} \frac{\min\{1, \gamma_{\text{eff}}\}}{|\gamma_{\text{eff}} - 1|}.$$

Thus

$$\mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] \xrightarrow{a.s.} \bar{w}^* \bar{\sigma}_{\text{eff}}^2 \frac{\min\{1, \gamma_{\text{eff}}\}}{|\gamma_{\text{eff}} - 1|}.$$

It remains to show concentration around this conditional expectation. Write

$$\mathbf{c} = \sqrt{\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}}\, \mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_{n_l}),$$

conditionally on $(\mathbf{X}_m, \mathbf{U}_m)$. Then

$$Q - \mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] = (\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}) \left(\mathbf{z}^\top \mathbf{A} \mathbf{z} - \mathrm{Tr}(\mathbf{A})\right).$$

By the Hanson–Wright inequality, for every fixed $\delta > 0$,

$$\mathbb{P}_{\mathbf{E}}(|Q - \mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m]| \geq \delta \mid \mathbf{X}_m, \mathbf{U}_m) \leq 2 \exp\left[-c \min\left(\frac{\delta^2}{(\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a})^2 \|\mathbf{A}\|_F^2}, \frac{\delta}{(\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}) \|\mathbf{A}\|}\right)\right].$$

We now bound the quantities appearing in this inequality. As above,

$$\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a} = \|\mathbf{w}^\star\|^2 \sigma_{\text{eff}}^2$$

is bounded for all sufficiently large p almost surely. Furthermore, since $\gamma_{\text{eff}} \neq 1$, the nonzero spectrum of $\frac{1}{n_l} \mathbf{X}_m^\top \mathbf{X}_m$ is bounded away from zero almost surely. Equivalently, the nonzero singular values of $\mathbf{X}_m$ are of order $\sqrt{p}$ (see argument in Step 3). Hence

$$\|\mathbf{A}\| = \|(\mathbf{X}_m \mathbf{X}_m^\top)^\dagger\| = O\left(\frac{1}{p}\right),$$

and

$$\|\mathbf{A}\|_F^2 = \sum_{i:\sigma_i(\mathbf{X}_m)>0} \frac{1}{\sigma_i(\mathbf{X}_m)^4} = O\left(\frac{1}{p}\right).$$

Consequently, for all sufficiently large p,

$$\mathbb{P}_{\mathbf{E}}(|Q - \mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m]| \geq \delta \mid \mathbf{X}_m, \mathbf{U}_m) \leq 2e^{-c_\delta p}.$$

The right-hand side is summable in p, so Borel–Cantelli gives

$$Q - \mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] \xrightarrow{a.s.} 0.$$

Combining this concentration statement with the limit of the conditional expectation yields

$$Q \xrightarrow{a.s.} \bar{w}^* \bar{\sigma}_{\text{eff}}^2 \frac{\min\{1, \gamma_{\text{eff}}\}}{|\gamma_{\text{eff}} - 1|}.$$

Combining. Summing the contributions from Steps 1–5 and applying the continuity mapping theorem concludes the proof.

□

A.4 Proof for the generalisation error

Proof of Theorem 2. We decompose the error and take averages over $\mathbf{E}$, $\mathbf{X}_m$, $\mathbf{U}_m$.

Step 1: Decomposition of the generalisation error. We start from the definition of the generalisation error

$$E^{\text{gen}} = \mathbb{E}_{\mathbf{x},y,\xi}\left[\|y - \mathbf{x}^\top \hat{\mathbf{w}}\|^2 \mid \mathbf{w}^\star, \mathbf{X}_l, \mathbf{X}_u\right]. \tag{18}$$

Using the data model $y = \mathbf{x}^\top \mathbf{w}^\star + \xi$ and independence of the test noise, this becomes

$$\begin{aligned} E^{\text{gen}} &= \mathbb{E}_{\mathbf{x},\xi}\left[(\mathbf{w}^\star - \hat{\mathbf{w}})^\top \mathbf{x}\mathbf{x}^\top (\mathbf{w}^\star - \hat{\mathbf{w}}) \mid \mathbf{w}^\star, \mathbf{X}_l, \mathbf{X}_u\right] + \sigma^2 \\ &= \mathbb{E}_{\xi}\left[(\mathbf{w}^\star - \hat{\mathbf{w}})^\top \mathbf{\Sigma}_l (\mathbf{w}^\star - \hat{\mathbf{w}})\right] + \sigma^2. \end{aligned} \tag{19}$$

Using the spiked covariance model $\mathbf{\Sigma}_l = \mathbf{I}_p + (\lambda_l - 1)\mathbf{q}\mathbf{q}^\top$, we obtain

$$E^{\text{gen}} = E^{\text{est}} + (\lambda_l - 1)\,\mathbb{E}_{\xi}\left[\left((\mathbf{w}^\star - \hat{\mathbf{w}})^\top \mathbf{q}\right)^2\right] + \sigma^2. \tag{20}$$

Substituting $\hat{\mathbf{w}} = (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star + (\mathbf{X}_l\mathcal{P}_m)^\dagger \xi$ and expanding the quadratic form, the cross term vanishes in ξ expectation and we obtain a bias contribution

$$\left((\mathbf{w}^\star - (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star)^\top \mathbf{q}\right)^2 \tag{21}$$

and a variance contribution proportional to

$$\mathbf{q}^\top (\mathcal{P}_m \mathbf{X}_l^\top \mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{q}. \tag{22}$$

Using the same decomposition and conditional Gaussian argument as in Equation (15), $(\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star = \Pi \mathbf{w}^\star + A\Pi\mathbf{q} + (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{E}\mathbf{w}^\star{}_\perp$, we further decompose the bias

$$(\mathbf{w}^\star - (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star)^\top \mathbf{q} = \mathbf{w}^{\star\top}\mathbf{q} - \mathbf{q}^\top \Pi \mathbf{w}^\star - A\mathbf{q}^\top \Pi \mathbf{q} - \mathbf{q}^\top (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{E}\mathbf{w}^\star{}_\perp, \tag{23}$$

with the squared term being:

$$\underbrace{(\mathbf{w}^{\star\top}\mathbf{q} - \mathbf{q}^\top \Pi \mathbf{w}^\star - \frac{\lambda_l - 1}{\tilde{\lambda}_l}(\mathbf{w}^{\star\top}\mathcal{P}_\perp \mathbf{q})(\mathbf{q}^\top \Pi \mathbf{q}))^2}_{\mathbf{E}\text{–independent}} + \underbrace{(\mathbf{q}^\top (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{E}\mathbf{w}^\star{}_\perp)^2}_{\text{quadratic in } \mathbf{E}}$$

$$\underbrace{- 2\,(\mathbf{w}^{\star\top}\mathbf{q} - \mathbf{q}^\top \Pi \mathbf{w}^\star - \frac{\lambda_l - 1}{\tilde{\lambda}_l}(\mathbf{w}^{\star\top}\mathcal{P}_\perp \mathbf{q})(\mathbf{q}^\top \Pi \mathbf{q}))(\mathbf{q}^\top (\mathbf{X}_l\mathcal{P}_m)^\dagger \mathbf{E}\mathbf{w}^\star{}_\perp)}_{\text{cross term}}.$$

Step 2: E-independent term. This component is made up of terms that have almost sure limits. In particular:

$$\begin{aligned}
\mathbf{q}^\top \Pi \mathbf{w}^\star &\xrightarrow{a.s.} \sqrt{\bar{w}^*}\bar{\pi}_{\mathbf{w}^\star,\mathbf{q}} && \text{by Lemma 7}\\
\mathbf{q}^\top \Pi \mathbf{q} &\xrightarrow{a.s.} \bar{\pi}_{\mathbf{q},\mathbf{q}} && \text{by Lemma 7}\\
\mathbf{w}^{\star\top} \mathcal{P}_\perp \mathbf{q} &\xrightarrow{a.s.} \sqrt{\bar{w}^*}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}} && \text{by Lemma 5}\\
\tilde{\lambda}_l &\xrightarrow{a.s.} \bar{\lambda}_l && \text{by Lemma 4}\\
\mathbf{w}^{\star\top} \mathbf{q} &\xrightarrow{a.s.} \sqrt{\bar{w}^*} r_q && \text{by model assumptions}
\end{aligned}$$

Thus, by the continuous mapping theorem, we have the almost sure convergence of the entire $\mathbf{E}$-independent term to

$$\bar{w}^* \left(r_q - \bar{\pi}_{\mathbf{w}^\star,\mathbf{q}} - \frac{\lambda_l - 1}{\bar{\lambda}_l} \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}} \bar{\pi}_{\mathbf{q},\mathbf{q}} \right)^2.$$

Step 3: Cross term linear in E. The linear $\mathbf{E}$ term has the form

$$T = \mathbf{c}^\top \mathbf{E} \mathbf{d},$$

where

$$\mathbf{c} = C \left(\mathbf{X}_l \mathcal{P}_m\right)^{\dagger\top} \mathbf{q}, \qquad \mathbf{d} = \mathbf{w}^\star{}_\perp,$$

and C is the deterministic prefactor defined in Step 2, which converges almost surely to an $O(1)$ limit.

This term has the same structure as the term linear in $\mathbf{E}$ in Theorem 1, Step 3, with the vector

$$\mathbf{w}^\star - \Pi \mathbf{w}^\star - A \Pi \mathbf{q}$$

replaced by the unit vector $\mathbf{q}$ and multiplied by the scalar $C = O(1)$. The same arguments apply and we conclude that

$$T \xrightarrow{a.s.} 0.$$

Step 4: Quadratic term in E. Define

$$Q := \mathbf{w}^\star{}_\perp^\top \mathbf{E}^\top (\mathbf{X}_l \mathcal{P}_m)^{\dagger\top} \mathbf{q}\mathbf{q}^\top (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{E} \mathbf{w}^\star{}_\perp.$$

Let

$$\mathbf{a} := \mathbf{w}^\star{}_\perp, \qquad \mathbf{A} := (\mathbf{X}_l \mathcal{P}_m)^{\dagger\top} \mathbf{q}\mathbf{q}^\top (\mathbf{X}_l \mathcal{P}_m)^\dagger.$$

Conditioned on $(\mathbf{X}_m, \mathbf{U}_m)$, the vector $\mathbf{a}$ and matrix $\mathbf{A}$ are independent of $\mathbf{E}$, so $Q = \mathbf{a}^\top \mathbf{E}^\top \mathbf{A} \mathbf{E} \mathbf{a}$ is a quadratic Gaussian form.

This is exactly the same object as in Theorem 1, Step 5, with a different choice of $\mathbf{A}$. The concentration argument therefore carries over verbatim once we verify that the same norm bounds hold.

First, as in Theorem 1, Step 3,

$$\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a} = O(1) \qquad \text{almost surely.}$$

Next, since $\mathbf{A}$ is rank one,

$$\|\mathbf{A}\| = \|(\mathbf{X}_l \mathcal{P}_m)^{\top\dagger} \mathbf{q}\|^2, \qquad \|\mathbf{A}\|_F^2 = \|\mathbf{A}\|^2.$$

Using the bounds on $\|(\mathbf{X}_l \mathcal{P}_m)^\dagger\|_{\mathrm{op}}$ established in Theorem 1, Step 3, together with $\|\mathbf{q}\| = 1$, we obtain

$$\|\mathbf{A}\| = O\left(\frac{1}{p}\right), \qquad \|\mathbf{A}\|_F^2 = O\left(\frac{1}{p^2}\right), \qquad \text{almost surely.}$$

The conditions required for the Hanson–Wright concentration used in Theorem 1, Step 5 are satisfied. Therefore, by the same argument as in Theorem 1, Step 5,

$$Q - \mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] \xrightarrow{a.s.} 0.$$

Finally, the conditional expectation is

$$\mathbb{E}_{\mathbf{E}}[Q \mid \mathbf{X}_m, \mathbf{U}_m] = (\mathbf{a}^\top \bar{\mathbf{\Sigma}} \mathbf{a}) \mathrm{Tr}(\mathbf{A}) = O\left(\frac{1}{p}\right) \xrightarrow{a.s.} 0$$

Step 5: Variance contribution. Finally, we consider the term from Equation (22) and note it is the same as the norm of the matrix $\mathbf{A}$ from Step 4:

$$\mathbf{q}^\top(\mathcal{P}_m\mathbf{X}_l^\top\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{q} = \|(\mathbf{X}_l\mathcal{P}_m)^{\dagger\top}\mathbf{q}\|^2 = \|\mathbf{A}\| = O\left(\frac{1}{p}\right).$$

Thus, this term converges to 0:

$$\mathbf{q}^\top(\mathcal{P}_m\mathbf{X}_l^\top\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{q} \xrightarrow{a.s.} 0.$$

Combining. Thus, in the high-dimensional limit the only additional contribution relative to the estimation error is the spike bias term, and the generalisation error converges to

$$E_\infty^{\text{gen}} = E_\infty^{\text{est}} + \bar{w}^*(\lambda_l - 1)\left(r_q - \bar{\pi}_{\mathbf{w}^\star,\mathbf{q}} - \frac{\lambda_l - 1}{\bar{\lambda}_l}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}\bar{\pi}_{\mathbf{q},\mathbf{q}}\right)^2 + \sigma^2.$$

□

A.5 Proof for the training error

Proof of Theorem 3. Let

$$\mathbf{H} = (\mathbf{X}_l\mathcal{P}_m)(\mathbf{X}_l\mathcal{P}_m)^\dagger$$

and recall that the target vector is $\mathbf{y}_d = \mathbf{X}_l\mathbf{w}^\star + \xi$. Then, the residual vector is:

$$\begin{aligned}\mathbf{y}_d - \mathbf{X}_l\hat{\mathbf{w}} &= \mathbf{y}_d - (\mathbf{X}_l\mathcal{P}_m)(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{y}_d - (\mathbf{X}_l\mathcal{P}_\perp)(\mathbf{X}_l\mathcal{P}_m)^\dagger\mathbf{y}_d \\ &= (\mathbf{I}_{n_l} - \mathbf{H})\mathbf{y}_d \\ &= (\mathbf{I}_{n_l} - \mathbf{H})(\mathbf{X}_l\mathbf{w}^\star + \xi),\end{aligned} \tag{24}$$

where we note that the term multiplying two projections is 0:

$$(\mathbf{X}_l\mathcal{P}_\perp)(\mathbf{X}_l\mathcal{P}_m)^\dagger = \mathbf{X}_l\mathbf{U}_\perp\underbrace{\mathbf{U}_\perp^\top\mathbf{U}_m}_{0}(\mathbf{X}_l\mathbf{U}_m)^\dagger = 0.$$

The training error splits into two terms since the noise ξ is independent and zero-mean:

$$\begin{aligned}E^{\text{train}} &= \frac{1}{n_l}\mathbb{E}_\xi\left[\|(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_l\mathbf{w}^\star + (\mathbf{I}_{n_l} - \mathbf{H})\xi\|^2\right] \\ &= \underbrace{\frac{1}{n_l}\|(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_l\mathbf{w}^\star\|^2}_{\text{Bias}^2} + \underbrace{\frac{1}{n_l}\mathbb{E}_\xi\left[\|(\mathbf{I}_{n_l} - \mathbf{H})\xi\|^2\right]}_{\text{Variance}}.\end{aligned} \tag{25}$$

We start by computing the bias. Using the decomposition $\mathbf{X}_l = \mathbf{X}_l\mathcal{P}_m + \mathbf{X}_l\mathcal{P}_\perp$ and the fact that $(\mathbf{I} - \mathbf{H})\mathbf{X}_l\mathcal{P}_m = 0$, the bias simplifies to:

$$\text{Bias}^2 = \frac{1}{n_l}\|(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_l\mathcal{P}_\perp\mathbf{w}^\star\|^2. \tag{26}$$

Substituting the conditional Gaussian decomposition, we have

$$\mathbf{X}_l\mathcal{P}_\perp = \mathbf{X}_m\mathbf{\Sigma}_{mm}^{-1}\mathbf{\Sigma}_{m\perp}\mathbf{U}_\perp^\top + \mathbf{E}\mathbf{U}_\perp^\top,$$

where $\mathbf{E}$ is independent of $\mathbf{X}_m$. Thus, we obtain

$$\begin{aligned}(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_l\mathcal{P}_\perp\mathbf{w}^\star &= (\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_m\mathbf{\Sigma}_{mm}^{-1}\mathbf{\Sigma}_{m\perp}\mathbf{U}_\perp^\top\mathbf{w}^\star + (\mathbf{I}_{n_l} - \mathbf{H})\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star \\ &= (\mathbf{I}_{n_l} - \mathbf{H})\mathbf{E}\mathbf{w}^\star{}_\perp,\end{aligned} \tag{27}$$

where we used $(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{X}_m = 0$ and defined $\mathbf{w}^\star{}_\perp = \mathbf{U}_\perp^\top\mathbf{w}^\star$. Note that the entries of $\mathbf{E}\mathbf{w}^\star{}_\perp$ are i.i.d. $\mathcal{N}(0, \|\mathbf{w}^\star\|^2\sigma_{\text{eff}}^2)$, with

$$\sigma_{\text{eff}}^2 = \frac{\|\mathcal{P}_\perp\mathbf{w}^\star\|^2}{\|\mathbf{w}^\star\|^2} + \frac{\lambda_l - 1}{\tilde{\lambda}_l}\frac{(\mathbf{q}^\top\mathcal{P}_\perp\mathbf{w}^\star)^2}{\|\mathbf{w}^\star\|^2}. \tag{28}$$

Conditioned on $\mathbf{X}_m$, $\mathbf{H}$ is deterministic, and the term $\|(\mathbf{I}_{n_l} - \mathbf{H})\mathbf{E}\mathbf{w}^\star{}_\perp\|^2$ is a quadratic form $z^\top(\mathbf{I}_{n_l} - \mathbf{H})z$ where $z \sim \mathcal{N}(0, \|\mathbf{w}^\star\|^2\sigma_{\text{eff}}^2\mathbf{I}_{n_l})$. Hence, we have

$$\mathbb{E}_{\mathbf{E}}\left[\text{Bias}^2 \mid \mathbf{X}_m\right] = \|\mathbf{w}^\star\|^2\frac{\sigma_{\text{eff}}^2}{n_l}\operatorname{Tr}(\mathbf{I}_{n_l} - \mathbf{H}) = \|\mathbf{w}^\star\|^2\sigma_{\text{eff}}^2\left(1 - \frac{\min(m, n_l)}{n_l}\right), \tag{29}$$

where the last equality holds almost surely (w.r.t. the randomness in $\mathbf{X}_m$).

Next, to compute the variance, recall that the label noise ξ is isotropic $\mathcal{N}(0, \sigma^2\mathbf{I}_{n_l})$. Thus, we have

$$\frac{1}{n_l}\mathbb{E}_\xi\left[\xi^\top(\mathbf{I}_{n_l} - \mathbf{H})\xi\right] = \frac{\sigma^2}{n_l}\operatorname{Tr}(\mathbf{I}_{n_l} - \mathbf{H}) = \sigma^2\left(1 - \frac{\min(m, n_l)}{n_l}\right), \tag{30}$$

where the last equality again holds almost surely (w.r.t. the randomness in $\mathbf{X}_m$). Note that

$$1 - \frac{\min(m, n_l)}{n_l} = \max(0, 1 - \gamma_{\text{eff}}).$$

To conclude, it suffices to recall from the proof of Theorem 1 that $\sigma_{\text{eff}}^2 \xrightarrow{\text{a.s.}} \bar{\sigma}_{\text{eff}}^2$, and the argument is complete. □

A.6 Auxiliary lemmas

Lemma 1 (Limiting spectral distribution of $\mathbf{X}_m$)**.** *Consider the setup of Section 3, and define*

$$\mathbf{X}_m := \mathbf{X}_l\mathbf{U}_m.$$

Then the empirical spectral measure of $\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{X}_m$ converges weakly almost surely to the Marchenko–Pastur law with aspect ratio γ_{eff}, i.e.,

$$\hat{F}^{\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{X}_m} \Rightarrow F_{\gamma_{\text{eff}}} \qquad \textit{almost surely.}$$

The bounded-m case, including $m = 1$, is discussed in the remark following the proof.

Proof. Fix a realization of the full sequence $(\mathbf{U}_m)_{p\geq 1}$. Conditional on this realization, the rows of $\mathbf{X}_m = \mathbf{X}_l\mathbf{U}_m$ are i.i.d. Gaussian in $\mathbb{R}^m$ with covariance

$$\mathbf{\Sigma}_{mm} := \mathbf{U}_m^\top\mathbf{\Sigma}_l\mathbf{U}_m.$$

Using the form of $\mathbf{\Sigma}_l$ from Assumption 2, we obtain

$$\mathbf{\Sigma}_{mm} = \mathbf{U}_m^\top\Big(\mathbf{I}_p + (\lambda_l - 1)\mathbf{q}\mathbf{q}^\top\Big)\mathbf{U}_m = \mathbf{I}_m + (\lambda_l - 1)(\mathbf{U}_m^\top\mathbf{q})(\mathbf{U}_m^\top\mathbf{q})^\top.$$

Hence, $\mathbf{\Sigma}_{mm} - \mathbf{I}_m$ has rank at most one. Since $\mathbf{\Sigma}_{mm} \succ 0$, define $\mathbf{G} \in \mathbb{R}^{n_l\times m}$ and $\mathbf{W} \in \mathbb{R}^{m\times m}$ by

$$\mathbf{G} := \mathbf{X}_m\mathbf{\Sigma}_{mm}^{-1/2} \qquad \mathbf{W} := \frac{1}{n_l}\mathbf{G}^\top\mathbf{G}.$$

Conditional on $\mathbf{U}_m$, $\mathbf{G}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. The sample covariance can be rewritten:

$$\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{X}_m = \mathbf{\Sigma}_{mm}^{1/2}\left(\frac{1}{n_l}\mathbf{G}^\top\mathbf{G}\right)\mathbf{\Sigma}_{mm}^{1/2} = \mathbf{\Sigma}_{mm}^{1/2}\mathbf{W}\mathbf{\Sigma}_{mm}^{1/2}.$$

Since $\mathbf{\Sigma}_{mm} = \mathbf{I}_m + \mathbf{R}$ with $\operatorname{rank}(\mathbf{R}) \leq 1$, the spectral decomposition shows that $\mathbf{\Sigma}_{mm}^{1/2} - \mathbf{I}_m$ also has rank at most one. Consequently,

$$\mathbf{\Sigma}_{mm}^{1/2}\mathbf{W}\mathbf{\Sigma}_{mm}^{1/2} - \mathbf{W} = (\mathbf{\Sigma}_{mm}^{1/2} - \mathbf{I}_m)\mathbf{W}\mathbf{\Sigma}_{mm}^{1/2} + \mathbf{W}(\mathbf{\Sigma}_{mm}^{1/2} - \mathbf{I}_m),$$

and thus

$$\operatorname{rank}\Big(\mathbf{\Sigma}_{mm}^{1/2}\mathbf{W}\mathbf{\Sigma}_{mm}^{1/2} - \mathbf{W}\Big) \leq 2.$$

Let $\hat{F}^{\mathbf{A}}$ denote the empirical spectral distribution function of a real symmetric matrix $\mathbf{A}$. By the rank inequality (e.g., see Theorem A.43 in [67]),

$$\sup_{\theta\in\mathbb{R}}\left|\hat{F}^{\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{X}_m}(\theta) - \hat{F}^{\mathbf{W}}(\theta)\right| \leq \frac{2}{m} \underset{p\to\infty}{\to} 0.$$

Since $\mathbf{W}$ is a standard Wishart matrix of size $m \times m$ with aspect ratio $\frac{m}{n_l} \to \gamma_{\text{eff}}$, by the Marchenko–Pastur theorem [66],

$$\hat{F}^{\mathbf{W}} \Rightarrow F_{\gamma_{\text{eff}}} \qquad \text{almost surely, conditional on } (\mathbf{U}_m)_{p\geq 1}.$$

Let $A := \left\{ F^{\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m} \Rightarrow F_{\gamma_{\text{eff}}} \right\}$. Taken together, the arguments above shows that:

$$\mathbb{P}(A \,|\, (\mathbf{U}_m)_{p\geq 1}) = 1.$$

We now remove the conditioning on $(\mathbf{U}_m)_{p\geq 1}$. By the law of total probability,

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{P}(A \,|\, (\mathbf{U}_m)_{p\geq 1})] = \mathbb{E}[1] = 1.$$

This proves that

$$\hat{F}^{\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m} \Rightarrow F_{\gamma_{\text{eff}}} \qquad \text{almost surely.}$$

□

Remark 1 (Fixed projected dimension). The assumption $m \to \infty$ ensures that the population spike has vanishing weight in the empirical spectral distribution. For fixed m and $n_l \to \infty$, one instead uses consistency of the sample covariance:

$$\left\| \frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m - \mathbf{\Sigma}_{mm} \right\|_{\text{op}} \xrightarrow{a.s.} 0.$$

Consequently, if

$$\tilde{\lambda}_l := 1 + (\lambda_l - 1)\|\mathbf{U}_m^\top \mathbf{q}\|^2 \xrightarrow{\text{a.s.}} \bar{\lambda}_l$$

then

$$\hat{F}^{\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m} \Rightarrow \frac{m-1}{m}\delta_1 + \frac{1}{m}\delta_{\bar{\lambda}_l}.$$

Thus, for $m > 1$, the non-spike part of the spectrum converges to 1, consistently with $F_0 = \delta_1$. For $m = 1$, there is only the spiked eigenvalue and no bulk.

Lemma 2 (Deterministic equivalent for the pseudoinverse trace)**.** *Consider the setup of Section 3, and define*

$$\mathbf{X}_m := \mathbf{X}_l \mathbf{U}_m.$$

Then

$$\text{Tr}\big[(\mathbf{X}_m^\top \mathbf{X}_m)^\dagger\big] \xrightarrow{a.s.} \begin{cases} \dfrac{\gamma_{\text{eff}}}{1-\gamma_{\text{eff}}}, & \gamma_{\text{eff}} < 1, \\ \dfrac{1}{\gamma_{\text{eff}} - 1}, & \gamma_{\text{eff}} > 1. \end{cases}$$

Proof. We have

$$\text{Tr}\big[(\mathbf{X}_m^\top \mathbf{X}_m)^\dagger\big] = \frac{1}{n_l}\text{Tr}\Big((\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m)^\dagger\Big) = \frac{m}{n_l}\cdot\frac{1}{m}\text{Tr}\Big((\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m)^\dagger\Big). \tag{31}$$

By Lemma 1, the empirical spectral distribution of $\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m$ converges weakly almost surely to the Marcenko–Pastur law with aspect ratio γ_{eff}. If $\gamma_{\text{eff}} < 1$, its support is bounded away from zero, and the claim follows by continuity of $x \mapsto 1/x$. If $\gamma_{\text{eff}} > 1$, there is an additional atom at zero; however, the pseudo-inverse removes this null space contribution, and the nonzero spectrum remains bounded away from zero, so the same argument applies on the effective support. Therefore,

$$\frac{1}{m}\text{Tr}\Big(\Big(\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m\Big)^\dagger\Big) = \int_{(0,\infty)} \frac{1}{x}\, dF^{\frac{1}{n_l}\mathbf{X}_m^\top \mathbf{X}_m}(x) \xrightarrow{a.s.} \int_{(0,\infty)} \frac{1}{x}\, dF_{\gamma_{\text{eff}}}(x).$$

Recall that

$$\int \frac{1}{x}\, dF_\gamma = \begin{cases} \dfrac{1}{1-\gamma}, & \gamma < 1, \\ \dfrac{1}{\gamma(\gamma-1)}, & \gamma > 1. \end{cases}$$

Substituting into (31) and using $\gamma = \gamma_{\text{eff}}$ gives

$$\mathrm{Tr}\big[(\mathbf{X}_m^\top \mathbf{X}_m)^\dagger\big] \xrightarrow{a.s.} \begin{cases} \dfrac{\gamma_{\text{eff}}}{1-\gamma_{\text{eff}}}, & \gamma_{\text{eff}} < 1, \\ \dfrac{1}{\gamma_{\text{eff}}-1}, & \gamma_{\text{eff}} > 1, \end{cases}$$

which concludes the proof. □

Lemma 3 (Spectral measure of the overlap)**.** *Consider the setup of Section 3, and let $(\mathbf{u}_i)_{i=1}^p$ be the eigenvectors of $\mathbf{S}_u = \frac{1}{n_u}\mathbf{X}_u^\top \mathbf{X}_u$. Let $\mathbf{a} \in \mathbb{R}^p$ be a fixed vector with population mass distribution $G_{\mathbf{a}}$ on the eigenvectors of $\mathbf{\Sigma}_u$, i.e., $G_{\mathbf{a}} = \frac{(\mathbf{a}^\top \mathbf{v})^2}{\|\mathbf{a}\|^2}\delta_{\lambda_u} + \big(1 - \frac{(\mathbf{a}^\top \mathbf{v})^2}{\|\mathbf{a}\|^2}\big)\delta_1$. Define the empirical CDF*

$$\hat{G}_{\mathbf{a}}^{\text{sample}}(\theta) = \frac{1}{\|\mathbf{a}\|^2}\sum_{i=1}^{p} (\mathbf{a}^\top \mathbf{u}_i)^2 \, \mathbf{1}\{\lambda_i(\mathbf{S}_u) \le \theta\}.$$

Then, $\hat{G}_{\mathbf{a}}^{\text{sample}} \Rightarrow G_{\mathbf{a}}^{\text{sample}}$ almost surely, where $G_{\mathbf{a}}^{\text{sample}}$ is such that (i) *its support satisfies $\text{supp}(G_{\mathbf{a}}^{\text{sample}}) \subseteq \mathcal{S}_{\gamma_u} \cup \{\theta^*\} \cup \{0\}$ with $\mathcal{S}_{\gamma_u}$ being the support of the Marcenko-Pastur density with aspect ratio γ_u;* (ii) *it has a continuous density $g_{\mathbf{a}}^{\text{sample}}(\theta)$ on the interior of $\mathcal{S}_{\gamma_u}$*

$$g_{\mathbf{a}}^{\text{sample}}(\theta) = \left(\int \frac{\tau\, \gamma_u}{\tau^2 - \tau(1-\gamma_u+\theta) + \theta}\, dG_{\mathbf{a}}(\tau) \right) f_{\gamma_u}(\theta);$$

(iii) *it may have an outlier atom at $\{\theta^*\}$ where $\theta^* = \dfrac{\lambda_u(\lambda_u - 1 + \gamma_u)}{\lambda_u - 1}$ is the location of the spike λ in the sample covariance spectrum*

$$G_{\mathbf{a}}^{\text{sample}}(\{\theta^*\}) = \begin{cases} \frac{1 - \frac{\gamma_u}{(\lambda_u - 1)^2}}{1 + \frac{\gamma_u}{(\lambda_u - 1)}} G_{\mathbf{a}}(\{\lambda_u\}) & \lambda_u > 1 + \sqrt{\gamma_u}, \\ 0 & \lambda_u \le 1 + \sqrt{\gamma_u} \end{cases};$$

and (iv) *it may have an atom at 0 of weight*

$$G_{\mathbf{a}}^{\text{sample}}(\{0\}) = \begin{cases} 0, & \gamma_u \le 1, \\ \displaystyle\int \frac{\gamma_u - 1}{\gamma_u - 1 + \tau}\, dG_{\mathbf{a}}(\tau), & \gamma_u > 1. \end{cases}$$

Proof. The result follows by specializing Lemma 3 in [45] to the spiked covariance model of Assumption 2. □

Lemma 4 (Deterministic limit of $\|\mathcal{P}_m \mathbf{a}\|^2, \|\mathcal{P}_\perp \mathbf{a}\|^2$)**.** *Consider the setup of Section 3, and let $\mathbf{a} \in \mathbb{R}^p$ be a fixed vector with population mass distribution $G_{\mathbf{a}}$ on the eigenvectors of $\mathbf{\Sigma}_u$. Let $\mathcal{P}_m$ be the projection onto the top m eigenvectors of the prior sample covariance $\mathbf{S}_u$, with $m/p \to \alpha$. Let $G_{\mathbf{a}}^{\text{sample}}$ denote the deterministic limiting weighted spectral measure defined in Lemma 3. When $0 < \alpha < \min\{1, 1/\gamma_u\}$, let λ_t be the unique bulk threshold satisfying*

$$F_{\gamma_u}(\lambda_t) = 1 - \alpha.$$

For $\gamma_u \le 1$ and $\alpha = 1$, set $\lambda_t = \lambda_-$. For $\alpha = 0, \lambda_t = \lambda_+$ When $\gamma_u > 1$ and $\alpha \ge 1/\gamma_u$, no bulk threshold is needed: all positive-eigenvalue directions are retained, and the nullspace-completion rule determines the discarded mass. Then, almost surely,

$$\frac{\|\mathcal{P}_\perp \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \begin{cases} \displaystyle\int_{\lambda_-}^{\lambda_t} g_{\mathbf{a}}^{\text{sample}}(\lambda) d\lambda, & \gamma_u \le 1, \\ G_{\mathbf{a}}^{\text{sample}}(\{0\}) + \displaystyle\int_{\lambda_-}^{\lambda_t} g_{\mathbf{a}}^{\text{sample}}(\lambda) d\lambda, & \gamma_u > 1,\ \alpha < \frac{1}{\gamma_u}, \\ \dfrac{1-\alpha}{1 - \frac{1}{\gamma_u}}\, G_{\mathbf{a}}^{\text{sample}}(\{0\}), & \gamma_u > 1,\ \alpha \ge \frac{1}{\gamma_u}, \end{cases} \tag{32}$$

where $\lambda_- = (1 - \sqrt{\gamma_u})^2$ is the infimum of the support of the Marcenko-Pastur distribution as defined in Equation (7) *and*

$$\|\mathcal{P}_m \mathbf{a}\|^2 = \|\mathbf{a}\|^2 - \|\mathcal{P}_\perp \mathbf{a}\|^2. \tag{33}$$

Proof. Let

$$\hat{G}_{\mathbf{a}}^{\text{sample}}(\theta) = \frac{1}{\|\mathbf{a}\|^2} \sum_{i=1}^{p} (\mathbf{a}^\top \mathbf{u}_i)^2 \, \mathbf{1}\{\lambda_i(\mathbf{S}_u) \le \theta\}$$

be the weighted empirical spectral measure of $\mathbf{S}_u$. By Lemma 3, $\hat{G}_{\mathbf{a}}^{\text{sample}}$ converges almost surely to the deterministic measure $G_{\mathbf{a}}^{\text{sample}}$.

Since $\|\mathcal{P}_\perp \mathbf{a}\|^2/\|\mathbf{a}\|^2$ is exactly the mass of $\hat{G}_{\mathbf{a}}^{\text{sample}}$ on the discarded eigenspaces, its limit is obtained by integrating $G_{\mathbf{a}}^{\text{sample}}$ over the discarded part of the spectrum.

If $\gamma_u \le 1$, the limiting weighted spectral measure has no atom at zero, so the discarded mass is given by the bulk integral. This gives

$$\frac{\|\mathcal{P}_\perp \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \int_{\lambda_-}^{\lambda_t} g_{\mathbf{a}}^{\text{sample}}(\lambda) d\lambda.$$

If $\gamma_u > 1$ and $\alpha < 1/\gamma_u$, fewer than all positive-eigenvalue directions are retained, so the entire nullspace is discarded as well. Hence

$$\frac{\|\mathcal{P}_\perp \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} G_{\mathbf{a}}^{\text{sample}}(\{0\}) + \int_{\lambda_-}^{\lambda_t} g_{\mathbf{a}}^{\text{sample}}(\lambda) d\lambda.$$

Finally, if $\gamma_u > 1$ and $\alpha \ge 1/\gamma_u$, all positive-eigenvalue directions are retained, and the retained subspace is completed by adding

$$m - \operatorname{rank}(\mathbf{S}_u) = m - n_u$$

orthonormal directions chosen uniformly at random in the nullspace, independently of $\mathbf{a}$. Therefore only the fraction

$$\frac{p-m}{p-n_u} \xrightarrow{a.s.} \frac{1-\alpha}{1-\frac{1}{\gamma_u}}$$

of the nullspace remains discarded. Since the total asymptotic mass of $\mathbf{a}$ in the nullspace is $G_{\mathbf{a}}^{\text{sample}}(\{0\})$, we obtain

$$\frac{\|\mathcal{P}_\perp \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \frac{1-\alpha}{1-\frac{1}{\gamma_u}} \, G_{\mathbf{a}}^{\text{sample}}(\{0\}).$$

The identity $\|\mathcal{P}_m \mathbf{a}\|^2 = \|\mathbf{a}\|^2 - \|\mathcal{P}_\perp \mathbf{a}\|^2$ is immediate. □

Lemma 5 (Deterministic limit of $\mathbf{a}^\top \mathcal{P}_m \mathbf{b}, \mathbf{a}^\top \mathcal{P}_\perp \mathbf{b}$)**.** *Consider the setup of Section 3, let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ be fixed vectors and let $\mathcal{P}_m$ denote the projection onto the top m principal components of $\mathbf{S}_u$, with $m/p \to \alpha$. Define the signed* population *spectral measure*

$$G_{\mathbf{a},\mathbf{b}}(\tau) = \frac{1}{\|\mathbf{a}\|\|\mathbf{b}\|} \sum_j (\mathbf{a}^\top \mathbf{v}_j)(\mathbf{b}^\top \mathbf{v}_j) \, \mathbf{1}\{\lambda_{u,j} \le \tau\}, \tag{34}$$

where $(\mathbf{v}_j, \lambda_{u,j})$ are the population eigenpairs of $\mathbf{\Sigma}_u$. Let $G_{\mathbf{a},\mathbf{b}}^{\text{sample}}$ denote the corresponding deterministic limiting signed sample *spectral measure onto the eigenpairs of $\mathbf{S}_u$. Use the same bulk-threshold convention and Haar-uniform nullspace-completion rule as in Lemma 4. Then,*

$$\mathbf{a}^\top \mathcal{P}_m \mathbf{b} = \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \mathcal{P}_\perp \mathbf{b}, \tag{35}$$

and, almost surely,

$$\frac{\mathbf{a}^\top \mathcal{P}_\perp \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \xrightarrow{a.s.} \begin{cases} \displaystyle\int_{\lambda_-}^{\lambda_t} g_{\mathbf{a},\mathbf{b}}^{\text{sample}}(\lambda) d\lambda, & \gamma_u \le 1, \\ \displaystyle G_{\mathbf{a},\mathbf{b}}^{\text{sample}}(\{0\}) + \int_{\lambda_-}^{\lambda_t} g_{\mathbf{a},\mathbf{b}}^{\text{sample}}(\lambda) d\lambda, & \gamma_u > 1, \ \alpha < 1/\gamma_u, \\ \displaystyle\frac{1-\alpha}{1-\frac{1}{\gamma_u}} \, G_{\mathbf{a},\mathbf{b}}^{\text{sample}}(\{0\}), & \gamma_u > 1, \ \alpha \ge 1/\gamma_u. \end{cases} \tag{36}$$

Moreover, $G_{\mathbf{a},\mathbf{b}}^{\text{sample}}(\{0\})$ and the bulk density of $g_{\mathbf{a},\mathbf{b}}^{\text{sample}}$ are given by the formulas of Lemma 3 with $dG_{\mathbf{a}}$ replaced by the signed measure $dG_{\mathbf{a},\mathbf{b}}$.

Proof. Define the signed mass density onto the *sample* eigenvectors of $\mathbf{S}_u$

$$\hat{G}^{\text{sample}}_{\mathbf{a},\mathbf{b}}(\theta) := \frac{1}{\|\mathbf{a}\|\|\mathbf{b}\|} \sum_{i=1}^{p} (\mathbf{a}^\top \mathbf{u}_i)(\mathbf{b}^\top \mathbf{u}_i)\, \mathbf{1}\{\lambda_i(\mathbf{S}_u) \le \theta\}, \tag{37}$$

so that its Stieltjes transform is

$$m_{\hat{G}^{\text{sample}}_{\mathbf{a},\mathbf{b}}}(z) = \int \frac{1}{\lambda - z}\, d\hat{G}^{\text{sample}}_{\mathbf{a},\mathbf{b}}(\lambda) = \frac{\mathbf{a}^\top (\mathbf{S}_u - zI)^{-1}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\mathrm{Tr}\big[\mathbf{a}\mathbf{b}^\top (\mathbf{S}_u - zI)^{-1}\big]}{\|\mathbf{a}\|\|\mathbf{b}\|}. \tag{38}$$

By Rubio and Mestre [100, Theorem 1], for any deterministic matrix Θ with bounded trace norm (i.e., $\mathrm{Tr}\big[(\Theta^\top\Theta)^{1/2}\big] = \|\Theta\|_{\mathrm{tr}}$),

$$\mathrm{Tr}\big[\Theta(\mathbf{S}_u - zI)^{-1}\big] - \delta_\Theta(z) \to 0 \qquad \text{a.s.}$$

where $\delta_\Theta(z)$ is a deterministic function of z. Applying this with $\Theta = \mathbf{a}\mathbf{b}^\top/\|\mathbf{a}\|\|\mathbf{b}\|$ and repeating the argument from [45] (in their paper, see Lemma 1 proven in Appendix A) yields the following almost sure weak convergence

$$\hat{G}^{\text{sample}}_{\mathbf{a},\mathbf{b}} \Rightarrow G^{\text{sample}}_{\mathbf{a},\mathbf{b}},$$

where the Stieltjes transform of $G^{\text{sample}}_{\mathbf{a},\mathbf{b}}$ is

$$m_{G^{\text{sample}}_{\mathbf{a},\mathbf{b}}}(z) = -\frac{1}{z} \int \frac{1}{1 + \tau\, \underline{m}_{\gamma_u}(z)}\, dG_{\mathbf{a},\mathbf{b}}(\tau). \tag{39}$$

Here, the companion Stieltjes transform is defined as $\underline{m}_{\gamma_u}(z) = \gamma_u m_{\gamma_u}(z) - (1-\gamma_u)\frac{1}{z}$, and m_{γ_u} is a Stieltjes transform of a Marcenko-Pastur density f_{γ_u}. This is exactly the same expression as in Lemma 1 in [45] with the population measure $G_{\mathbf{a}}$ replaced by the signed measure $G_{\mathbf{a},\mathbf{b}}$. Hence, the atom at 0 and the bulk density of $g^{\text{sample}}_{\mathbf{a},\mathbf{b}}(\{0\})$ are obtained from the formulas of Lemma 3 by the same substitution. Now the mass of $\hat{G}^{\text{sample}}_{\mathbf{a},\mathbf{b}}$ on the discarded eigenspaces is

$$\mathbf{a}^\top \mathcal{P}_\perp \mathbf{b} = \sum_{i=m+1}^{p} (\mathbf{a}^\top \mathbf{u}_i)(\mathbf{b}^\top \mathbf{u}_i) \tag{40}$$

noting that the eigenvectors $\mathbf{u}_i$ are sorted according to their eigenvalues, such that the set $\{m+1, \ldots, p\}$ contains the indices of the $p-m$ smallest eigenvalues. Therefore, the same geometric argument as in Lemma 4 gives the three regimes above: discarded bulk only when $\gamma_u \le 1$; bulk plus the full atom at 0 when $\gamma_u > 1$ and $\alpha < 1/\gamma_u$; and, when $\gamma_u > 1$ and $\alpha \ge 1/\gamma_u$, only the fraction

$$\frac{1-\alpha}{1-\frac{1}{\gamma_u}}$$

of the nullspace mass remains discarded, by the same random nullspace-completion argument. Finally,

$$\mathbf{a}^\top \mathcal{P}_m \mathbf{b} = \mathbf{a}^\top \mathbf{b} - \mathbf{a}^\top \mathcal{P}_\perp \mathbf{b},$$

which proves the result. □

Lemma 6 (Deterministic limit of the effective projection $\|\Pi\mathbf{a}\|^2$)**.** *Consider the setup of Section 3, let $\mathbf{X}_m = \mathbf{X}_l \mathbf{U}_m \in \mathbb{R}^{n_l \times m}$ and define*

$$\Pi_m := \mathbf{X}_m^\dagger \mathbf{X}_m \in \mathbb{R}^{m\times m}, \qquad \Pi := \mathbf{U}_m \Pi_m \mathbf{U}_m^\top \in \mathbb{R}^{p\times p}.$$

Let the eigendecomposition of $\mathbf{\Sigma}_{mm}$ be $\mathbf{\Sigma}_{mm} := \mathbf{U}_m^\top \mathbf{\Sigma}_l \mathbf{U}_m$ with $(\tilde{\mathbf{q}}_i, \tilde{\lambda}_i)$ denoting the eigenpairs. For any fixed vector $\mathbf{a} \in \mathbb{R}^p$ with $\mathcal{P}_m \mathbf{a} \neq 0$, define

$$\tilde{\mathbf{a}} := \frac{\mathbf{U}_m^\top \mathbf{a}}{\|\mathcal{P}_m \mathbf{a}\|} \in \mathbb{R}^m,$$

and let $\tilde{G}_{\tilde{\mathbf{a}}}$ be the mass distribution of $\tilde{\mathbf{a}}$ onto the eigenvectors of $\mathbf{\Sigma}_{mm}$:

$$\tilde{G}_{\tilde{\mathbf{a}}}(\tilde{\tau}) = \sum_i (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}}_i)^2\, \mathbf{1}\{\tilde{\lambda}_i \le \tilde{\tau}\}.$$

Then, it holds almost surely

$$\tilde{G}_{\tilde{\mathbf{a}}}(\tau) \Rightarrow \left(1 - \frac{\bar{p}^2_{\mathbf{a},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}\bar{p}_{\mathbf{q},\mathbf{q}}}\right)\delta_1 + \frac{\bar{p}^2_{\mathbf{a},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}\bar{p}_{\mathbf{q},\mathbf{q}}}\delta_{\bar{\lambda}_l},$$

and

$$\frac{\mathbf{a}^\top \Pi \mathbf{a}}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \bar{p}_{\mathbf{a},\mathbf{a}} \cdot \begin{cases} 1, & \gamma_{\text{eff}} < 1, \\ 1 - \displaystyle\int \frac{\gamma_{\text{eff}} - 1}{\gamma_{\text{eff}} - 1 + \tilde{\tau}}\, d\tilde{G}_{\tilde{\mathbf{a}}}(\tilde{\tau}), & \gamma_{\text{eff}} > 1. \end{cases} \tag{41}$$

with $\bar{p}_{\mathbf{a},\mathbf{a}}, \bar{p}_{\mathbf{q},\mathbf{q}}, \bar{p}_{\mathbf{a},\mathbf{q}}$ defined via Equations (32) *and* (36)*.*

Proof. We start by computing explicitly $\tilde{G}_{\tilde{\mathbf{a}}}$. Recall that

$$\mathbf{\Sigma}_{mm} = \mathbf{U}_m^\top \mathbf{\Sigma}_l \mathbf{U}_m = \mathbf{I}_m + (\lambda_l - 1)(\mathbf{U}_m^\top \mathbf{q})(\mathbf{U}_m^\top \mathbf{q})^\top.$$

Then, $\mathbf{\Sigma}_{mm}$ has $m-1$ eigenvalues equal to 1 and the remaining one is

$$\tilde{\lambda}_l = 1 + (\lambda_l - 1)\|\mathcal{P}_m \mathbf{q}\|^2 \quad \Rightarrow \quad \bar{\lambda}_l = 1 + (\lambda_l - 1)\bar{p}_{\mathbf{q},\mathbf{q}},$$

where the convergence occurs almost surely. The corresponding unit eigenvector is $\tilde{\mathbf{q}} = \frac{\mathbf{U}_m^\top \mathbf{q}}{\|\mathbf{U}_m^\top \mathbf{q}\|} = \frac{\mathbf{U}_m^\top \mathbf{q}}{\|\mathcal{P}_m \mathbf{q}\|}$. Therefore the spectral measure $\tilde{G}_{\tilde{\mathbf{a}}}$ of

$$\tilde{\mathbf{a}} = \frac{\mathbf{U}_m^\top \mathbf{a}}{\|\mathcal{P}_m \mathbf{a}\|}$$

with respect to the eigenbasis of $\mathbf{\Sigma}_{mm}$ is supported on $\{1, \tilde{\lambda}_l\}$, and hence

$$\tilde{G}_{\tilde{\mathbf{a}}} = \big(1 - (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})^2\big)\delta_1 + (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})^2 \delta_{\tilde{\lambda}_l},$$

with

$$(\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})^2 = \frac{((\mathbf{U}_m^\top \mathbf{a})^\top (\mathbf{U}_m^\top \mathbf{q}))^2}{\|\mathcal{P}_m \mathbf{a}\|^2 \, \|\mathcal{P}_m \mathbf{q}\|^2} = \frac{(\mathbf{a}^\top \mathbf{U}_m \mathbf{U}_m^\top \mathbf{q})^2}{\|\mathcal{P}_m \mathbf{a}\|^2 \, \|\mathcal{P}_m \mathbf{q}\|^2} = \frac{(\mathbf{a}^\top \mathcal{P}_m \mathbf{q})^2}{\|\mathcal{P}_m \mathbf{a}\|^2 \, \|\mathcal{P}_m \mathbf{q}\|^2}.$$

Thus, $\tilde{G}_{\tilde{\mathbf{a}}}$ is a two-point measure, with the original atom at λ replaced by an atom at $\tilde{\lambda}_l$, and with weight determined by the projected overlap. Almost surely, it holds that:

$$\tilde{G}_{\tilde{\mathbf{a}}}(\tau) \Rightarrow \left(1 - \frac{\bar{p}^2_{\mathbf{a},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}\bar{p}_{\mathbf{q},\mathbf{q}}}\right)\delta_1 + \frac{\bar{p}^2_{\mathbf{a},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}\bar{p}_{\mathbf{q},\mathbf{q}}}\delta_{\bar{\lambda}_l}.$$

Case $m < n_l$. In this regime, $\mathbf{X}_m \in \mathbb{R}^{n_l \times m}$ has full column rank almost surely, so $\Pi_m = \mathbf{I}_m$. Hence $\tilde{\mathbf{a}}^\top \Pi_m \tilde{\mathbf{a}} = 1$, giving

$$\frac{\mathbf{a}^\top \Pi \mathbf{a}}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \bar{p}_{\mathbf{a},\mathbf{a}},$$

as $\frac{\|\mathcal{P}_m \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \bar{p}_{\mathbf{a},\mathbf{a}}$ from Lemma 4.

Case $m > n_l$. In this regime, $\mathbf{X}_m$ has a nullspace of dimension $m - n_l$, and

$$\tilde{\mathbf{a}}^\top \Pi_m \tilde{\mathbf{a}} = 1 - \tilde{\mathbf{a}}^\top \Pi_{m,\text{null}} \tilde{\mathbf{a}}.$$

Conditioned on $\mathbf{U}_m$, the rows of $\mathbf{X}_m$ are Gaussian with covariance $\mathbf{\Sigma}_{mm} = \mathbf{U}_m^\top \mathbf{\Sigma}_l \mathbf{U}_m$. Therefore, by the nullspace mass formula from Lemma 3 applied to the m-dimensional model with aspect ratio $\gamma_{\text{eff}} = m/n_l > 1$,

$$\tilde{\mathbf{a}}^\top \Pi_{m,\text{null}} \tilde{\mathbf{a}} \xrightarrow{a.s.} \int \frac{\gamma_{\text{eff}} - 1}{\gamma_{\text{eff}} - 1 + \tilde{\tau}}\, d\tilde{G}_{\tilde{\mathbf{a}}}(\tilde{\tau}). \tag{42}$$

Finally, noting that $\frac{\|\mathcal{P}_m \mathbf{a}\|^2}{\|\mathbf{a}\|^2} \xrightarrow{a.s.} \bar{p}_{\mathbf{a},\mathbf{a}}$ according to Lemma 4 and combining the two terms via the continuous mapping theorem gives the required result.

□

Lemma 7 (Deterministic limit of the bilinear effective projection $\mathbf{a}^\top \Pi \mathbf{b}$)**.** *Consider the setup of Section 3, let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$*, with* $\mathcal{P}_m \mathbf{a} \neq \mathbf{0}, \mathcal{P}_m \mathbf{b} \neq \mathbf{0}$*, and define*

$$\tilde{\mathbf{a}} := \frac{\mathbf{U}_m^\top \mathbf{a}}{\|\mathcal{P}_m \mathbf{a}\|} \in \mathbb{R}^m, \qquad \tilde{\mathbf{b}} := \frac{\mathbf{U}_m^\top \mathbf{b}}{\|\mathcal{P}_m \mathbf{b}\|} \in \mathbb{R}^m.$$

Let $\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}$ *be the signed joint mass distribution of* $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}$ *onto the eigenvectors* $\tilde{\mathbf{q}}_i$ *of* $\mathbf{\Sigma}_{mm}$*:*

$$\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}(\tilde{\tau}) = \sum_i (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}}_i)(\tilde{\mathbf{b}}^\top \tilde{\mathbf{q}}_i)\, \mathbf{1}\{\tilde{\lambda}_i \leq \tilde{\tau}\}.$$

Then, it holds almost surely

$$\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}(\tau) \Rightarrow \left(\frac{\bar{p}_{\mathbf{a},\mathbf{b}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2}} - \frac{\bar{p}_{\mathbf{a},\mathbf{q}} \bar{p}_{\mathbf{b},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2} \bar{p}_{\mathbf{q},\mathbf{q}}} \right) \delta_1 + \frac{\bar{p}_{\mathbf{a},\mathbf{q}} \bar{p}_{\mathbf{b},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2} \bar{p}_{\mathbf{q},\mathbf{q}}} \delta_{\tilde{\lambda}_l},$$

and

$$\frac{\mathbf{a}^\top \Pi \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \xrightarrow{a.s.} \begin{cases} \bar{p}_{\mathbf{a},\mathbf{b}}, & \gamma_{\text{eff}} < 1, \\ \bar{p}_{\mathbf{a},\mathbf{b}} - \bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2} \displaystyle\int \frac{\gamma_{\text{eff}} - 1}{\gamma_{\text{eff}} - 1 + \tilde{\tau}}\, d\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}(\tilde{\tau}), & \gamma_{\text{eff}} > 1. \end{cases} \tag{43}$$

with $\bar{p}_{\mathbf{a},\mathbf{a}}, \bar{p}_{\mathbf{b},\mathbf{b}}, \bar{p}_{\mathbf{q},\mathbf{q}}, \bar{p}_{\mathbf{a},\mathbf{q}}, \bar{p}_{\mathbf{b},\mathbf{q}}$ *defined via Eq.* (32) *and* (36)*.*

Proof. The proof follows the steps of the proof of Lemma 6. The mass distribution $\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}$ can be written as

$$\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}} = \big(\tilde{\mathbf{a}}^\top \tilde{\mathbf{b}} - (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})(\tilde{\mathbf{b}}^\top \tilde{\mathbf{q}})\big)\delta_1 + (\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})(\tilde{\mathbf{b}}^\top \tilde{\mathbf{q}})\delta_{\tilde{\lambda}_l},$$

with

$$(\tilde{\mathbf{a}}^\top \tilde{\mathbf{q}})(\tilde{\mathbf{b}}^\top \tilde{\mathbf{q}}) = \frac{((\mathbf{U}_m^\top \mathbf{a})^\top (\mathbf{U}_m^\top \mathbf{q}))((\mathbf{U}_m^\top \mathbf{b})^\top (\mathbf{U}_m^\top \mathbf{q}))}{\|\mathcal{P}_m \mathbf{a}\|\, \|\mathcal{P}_m \mathbf{b}\|\, \|\mathcal{P}_m \mathbf{q}\|^2} = \frac{(\mathbf{a}^\top \mathcal{P}_m \mathbf{q})(\mathbf{b}^\top \mathcal{P}_m \mathbf{q})}{\|\mathcal{P}_m \mathbf{a}\|\, \|\mathcal{P}_m \mathbf{b}\|\, \|\mathcal{P}_m \mathbf{q}\|^2},$$

which converges almost surely to $\frac{\bar{p}_{\mathbf{a},\mathbf{q}} \bar{p}_{\mathbf{b},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2} \bar{p}_{\mathbf{q},\mathbf{q}}}$ and

$$\tilde{\mathbf{a}}^\top \tilde{\mathbf{b}} = \frac{(\mathbf{U}_m^\top \mathbf{a})^\top (\mathbf{U}_m^\top \mathbf{b})}{\|\mathcal{P}_m \mathbf{a}\| \|\mathcal{P}_m \mathbf{b}\|} \xrightarrow{a.s.} \frac{\bar{p}_{\mathbf{a},\mathbf{b}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{\mathbf{b},\mathbf{b}}^{1/2}}$$

This readily gives the almost sure limit of $\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}(\tau)$. Next, observe

$$\frac{\mathbf{a}^\top \Pi \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{1}{\|\mathbf{a}\| \|\mathbf{b}\|} (\mathbf{U}_m^\top \mathbf{a})^\top \Pi_m (\mathbf{U}_m^\top \mathbf{b}) = \frac{\|\mathcal{P}_m \mathbf{a}\|}{\|\mathbf{a}\|} \frac{\|\mathcal{P}_m \mathbf{b}\|}{\|\mathbf{b}\|}\, \tilde{\mathbf{a}}^\top \Pi_m \tilde{\mathbf{b}}. \tag{44}$$

Combining this with the almost sure limits from Lemma 4, with the same steps as in the proof of Lemma 6 and with the mass distribution of $\tilde{G}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}$ proves the lemma. □

B Phase transitions

B.1 Phase transition for $\gamma_l > 1$

Here, we derive parametric equations for the two phase transitions we observe in the phase diagrams for α in Figure 3. For small downstream tasks, $\gamma_l > 1$, there is a sharp phase transition due to competing local minima at different sides of the double descent (local minimum at large vs small α), which we tackle first. We find a proxy of the parametric curve of this phase transition by comparing the values of the error at the endpoints in α, and check when $E_\infty^{\text{gen}}(\alpha \to 0) = E_\infty^{\text{gen}}(\alpha = 1)$. Note that the left endpoint is not exactly 0, as we always retain the first principal component of the data, but in the high-dimensional regime $p \to \infty$, α approaches 0.

We denote signed normalized overlaps between deterministic vectors $\mathbf{a}, \mathbf{b}$ as

$$\rho_{\mathbf{a},\mathbf{b}} := \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\, \|\mathbf{b}\|}.$$

Note that the signs of the corresponding ρ's must be retained. Once we specify vectors $\mathbf{a}, \mathbf{b}$ to be one of the $\{\mathbf{w}^\star, \mathbf{v}, \mathbf{q}\}$, we use the previous notation r, r_v, r_q for the overlaps.

Define the limiting value of the population and sample spike eigenvector overlap $(\mathbf{v}^\top \mathbf{u}_1)^2 \to c$, which depends on whether the spike is below or above the BBP transition [45]:

$$c := \begin{cases} \dfrac{1 - \frac{\gamma_u}{(\lambda_u - 1)^2}}{1 + \frac{\gamma_u}{\lambda_u - 1}} & \gamma_u < (\lambda_u - 1)^2 \\ 0 & \gamma_u \geq (\lambda_u - 1)^2 \end{cases}.$$

Additionally, the squared overlap between the leading empirical principal component and any other vector $\mathbf{a} \perp \mathbf{v}$ is 0.

Endpoint values for the eigenvalues and projections. At the left endpoint, meaning $m = 1$ and hence $\alpha \to 0$, the retained direction is the empirical spike direction. Therefore, the limiting value of projections for any fixed vectors $\mathbf{a}, \mathbf{b}$ are given by

$$\begin{aligned} \bar{p}_{\mathbf{a},\mathbf{a}}(0) &= \rho_{\mathbf{a},\mathbf{v}}^2 c, & \bar{p}_{\perp,\mathbf{a},\mathbf{a}}(0) &= 1 - \rho_{\mathbf{a},\mathbf{v}}^2 c, \\ \bar{p}_{\mathbf{a},\mathbf{b}}(0) &= \rho_{\mathbf{a},\mathbf{v}} \rho_{\mathbf{b},\mathbf{v}} c, & \bar{p}_{\perp,\mathbf{a},\mathbf{b}}(0) &= \rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{v}} \rho_{\mathbf{b},\mathbf{v}} c. \end{aligned}$$

At the right endpoint, $\alpha = 1$, the projection is the identity, $\mathcal{P}_m = \mathbf{I}_p$. Hence, we have

$$\begin{aligned} \bar{p}_{\mathbf{a},\mathbf{a}}(1) &= 1, & \bar{p}_{\perp,\mathbf{a},\mathbf{a}}(1) &= 0 \\ \bar{p}_{\mathbf{a},\mathbf{b}}(1) &= \rho_{\mathbf{a},\mathbf{b}}, & \bar{p}_{\perp,\mathbf{a},\mathbf{b}}(1) &= 0. \end{aligned}$$

Next, let us analyse the effective limiting projections $\bar{\pi}$. At $\alpha = 0$, the effective aspect ratio satisfies $\gamma_{\text{eff}} = 0$. Hence the compressed task design is overdetermined, and the task projection is exact inside the retained subspace:

$$\begin{aligned} \bar{\pi}_{\mathbf{a},\mathbf{a}}(0) &= \bar{p}_{\mathbf{a},\mathbf{a}}(0) = \rho_{\mathbf{a},\mathbf{v}}^2 c, \\ \bar{\pi}_{\mathbf{a},\mathbf{b}}(0) &= \bar{p}_{\mathbf{a},\mathbf{b}}(0) = \rho_{\mathbf{a},\mathbf{v}} \rho_{\mathbf{b},\mathbf{v}} c. \end{aligned}$$

At $\alpha = 1$, the effective downstream eigenvalue is λ_l and $\gamma_{\text{eff}} = \gamma_l$. Therefore,

$$\bar{\pi}_{\mathbf{a},\mathbf{b}}(1) = \begin{cases} \rho_{\mathbf{a},\mathbf{b}}, & \gamma_l < 1, \\ \rho_{\mathbf{a},\mathbf{q}} \rho_{\mathbf{b},\mathbf{q}} \dfrac{\lambda_l}{\gamma_l - 1 + \lambda_l} + \dfrac{\rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{q}} \rho_{\mathbf{b},\mathbf{q}}}{\gamma_l}, & \gamma_l > 1. \end{cases}$$

At the compressed endpoint, the effective downstream spike eigenvalue is

$$\bar{\lambda}_{l,0} = 1 + (\lambda_l - 1) r^2 c.$$

Endpoint values for the generalisation error. At $\alpha = 1$, we have $P_m = I_p$, hence

$$\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star} = 0, \qquad \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}} = 0, \qquad \bar{\sigma}_{\text{eff}}^2 = 0.$$

Therefore, for $\gamma_l > 1$,

$$E_\infty^{\text{gen}}(1) = \bar{w}^* \big(1 - \bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}(1)\big) + \frac{\sigma^2}{\gamma_l - 1} + \sigma^2 + \bar{w}^* (\lambda_l - 1) \left[r_q - \bar{\pi}_{\mathbf{w}^\star,\mathbf{q}}(1)\right]^2.$$

Substituting the endpoint projections gives

$$\begin{aligned} B_1(\gamma_l) &:= 1 - r_q^2 \frac{\lambda_l}{\gamma_l - 1 + \lambda_l} - \frac{1 - r_q^2}{\gamma_l} + (\lambda_l - 1) \left[r_q - r_q \frac{\lambda_l}{\gamma_l - 1 + \lambda_l}\right]^2 \\ &= (1 - r_q^2) \frac{\gamma_l - 1}{\gamma_l} + r_q^2 \frac{\lambda_l \gamma_l (\gamma_l - 1)}{(\gamma_l - 1 + \lambda_l)^2}. \end{aligned}$$

Thus,

$$E_\infty^{\text{gen}}(1) = \bar{w}^* B_1(\gamma_l) + \frac{\sigma^2}{\gamma_l - 1} + \sigma^2.$$

At $\alpha = 0$, the effective aspect ratio vanishes, so the variance term in the estimation error vanishes. Moreover,

$$\begin{aligned}\bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}(0) &= r_v^2 c,\\ \bar{\pi}_{\mathbf{q},\mathbf{q}}(0) &= r^2 c,\\ \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}(0) &= r_q - r_v r c.\end{aligned}$$

Substitution into the estimation-error formula gives

$$\frac{E_\infty^{\mathrm{est}}(0)}{\bar{w}^*} = 1 - r_v^2 c + \frac{(\lambda_l - 1)^2 r^2 c}{\bar{\lambda}_{l,0}^2} \left(r_q - r_v r c\right)^2 .$$

The generalisation error residual simplifies to

$$\begin{aligned}&(\lambda_l - 1)\left(r_q - \bar{\pi}_{\mathbf{w}^\star,\mathbf{q}}(0) - \frac{\lambda_l - 1}{\bar{\lambda}_{l,0}} \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}(0)\bar{\pi}_{\mathbf{q},\mathbf{q}}(0)\right)^2\\ &\quad = (\lambda_l - 1)\frac{(r_q - r_v r c)^2}{\bar{\lambda}_{l,0}^2}\end{aligned}$$

Consequently, define

$$\begin{aligned}B_0(\gamma_u) &:= 1 - r_v^2 c + \frac{(\lambda_l - 1)^2 r^2 c}{\bar{\lambda}_{l,0}^2} \left(r_q - r_v r c\right)^2 + \frac{(\lambda_l - 1)}{\bar{\lambda}_{l,0}^2} \left(r_q - r_v r c\right)^2\\ &= 1 - r_v^2 c + \frac{(\lambda_l - 1)}{\bar{\lambda}_{l,0}} \left(r_q - r_v r c\right)^2\end{aligned}$$

Then

$$E_\infty^{\mathrm{gen}}(0) = \bar{w}^* B_0(\gamma_u) + \sigma^2.$$

Let $S := \bar{w}^*/\sigma^2$. The phase boundary in the (γ_u, γ_l)-plane is therefore approximated by

$$B_0(\gamma_u) = B_1(\gamma_l) + \frac{1}{S(\gamma_l - 1)}, \qquad \gamma_l > 1.$$

Here the dependence of B_0 on γ_u is through c. This endpoint comparison approximates the transition between the actual minima when they lie close to the endpoints.

Specialization to $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}_u$. Set $\mathbf{q} = \mathbf{v}$ and $\lambda_u = \lambda_l = \lambda$. Then

$$\begin{aligned}B_0 &= 1 - r_v^2 c + \frac{(\lambda - 1) r_v^2 (1 - c)^2}{1 + (\lambda - 1)c}\\ &= 1 + r_v^2(\lambda - 1) - \frac{r_v^2 \lambda^2 c}{1 + (\lambda - 1)c},\end{aligned}$$

while

$$B_1 = (1 - r_v^2)\frac{\gamma_l - 1}{\gamma_l} + r_v^2 \frac{\lambda \gamma_l (\gamma_l - 1)}{(\gamma_l - 1 + \lambda)^2}.$$

Specialization to $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}_u$ and $\mathbf{w}^\star = \sqrt{\bar{w}^*}\mathbf{v}$. Now everything simplifies even further:

$$B_0 = \frac{\lambda(1 - c)}{1 + (\lambda - 1)c}, \qquad B_1 = \frac{\lambda \gamma_l (\gamma_l - 1)}{(\gamma_l - 1 + \lambda)^2}.$$

Due to the two regimes of the empirical spike overlap, the proxy transition splits into two components:

$$\frac{\lambda \gamma_l (\gamma_l - 1)}{(\gamma_l - 1 + \lambda)^2} + \frac{1}{S(\gamma_l - 1)} = \begin{cases} \lambda, & \gamma_u \geq (\lambda - 1)^2,\\ \dfrac{\lambda \gamma_u}{(\lambda - 1)^2}, & \gamma_u < (\lambda - 1)^2. \end{cases}$$

B.2 Phase transition for $\gamma_l < 1$

Next, we study the regime of larger downstream tasks, $\gamma_l < 1$, and find a transition between the phases $\alpha = 1$ and $\alpha < 1$ by finding a parametric curve describing the loss of stability of $\alpha = 1$, i.e., when

$$\frac{d}{d\alpha} E_\infty^{\text{gen}}(\alpha)\Big|_{\alpha=1^-} = 0.$$

Derivatives of the pretraining projection at $\alpha = 1$. Let $\lambda_\pm = (1 \pm \sqrt{\gamma_u})^2$, and define

$$H_{\mathbf{a}}(\theta) := (1 - \rho_{\mathbf{a},\mathbf{v}}^2) + \rho_{\mathbf{a},\mathbf{v}}^2 \frac{\lambda_u \gamma_u}{\lambda_u^2 - \lambda_u(1 - \gamma_u + \theta) + \theta}.$$

Since

$$g_{\mathbf{a}}^{\text{sample}}(\theta) = H_{\mathbf{a}}(\theta) f_{\gamma_u}(\theta),$$

the derivative of $\bar{p}_{\perp,\mathbf{a},\mathbf{a}}$ at $\alpha = 1$ is

$$\frac{d}{d\alpha} \bar{p}_{\perp,\mathbf{a},\mathbf{a}}(\alpha)\Big|_{\alpha=1^-} = \begin{cases} -H_{\mathbf{a}}(\lambda_-), & \gamma_u < 1, \\ -\dfrac{G_{\mathbf{a}}^{\text{sample}}(\{0\})}{1 - \gamma_u^{-1}}, & \gamma_u > 1. \end{cases}$$

Equivalently,

$$\frac{d}{d\alpha} \bar{p}_{\mathbf{a},\mathbf{a}}(\alpha)\Big|_{\alpha=1^-} = (1 - \rho_{\mathbf{a},\mathbf{v}}^2) + \rho_{\mathbf{a},\mathbf{v}}^2 D_{\mathbf{v}}(\gamma_u; \lambda_u),$$

where

$$D_{\mathbf{v}}(\gamma_u; \lambda_u) = \begin{cases} \dfrac{\lambda_u \gamma_u}{(\lambda_u - 1 + \sqrt{\gamma_u})^2}, & \gamma_u < 1, \\ \dfrac{\gamma_u}{\gamma_u - 1 + \lambda_u}, & \gamma_u > 1. \end{cases}$$

The value at $\gamma_u = 1$ follows from the limiting spectral density at zero, or equivalently by continuity of the displayed expressions.

For signed bilinear projections, replace $\rho_{\mathbf{a},\mathbf{v}}^2$ by $\rho_{\mathbf{a},\mathbf{v}} \rho_{\mathbf{b},\mathbf{v}}$, and replace the constant bulk contribution $1 - \rho_{\mathbf{a},\mathbf{v}}^2$ by $\rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{v}} \rho_{\mathbf{b},\mathbf{v}}$.

Derivative of the effective projection at $\alpha = 1$. For $\gamma_l < 1$, the compressed task design remains overdetermined in a neighbourhood of $\alpha = 1$. Thus,

$$\bar{\pi}_{\mathbf{a},\mathbf{b}}(\alpha) = \bar{p}_{\mathbf{a},\mathbf{b}}(\alpha),$$

and hence

$$\frac{d}{d\alpha} \bar{\pi}_{\mathbf{a},\mathbf{b}}(\alpha)\Big|_{\alpha=1^-} = \frac{d}{d\alpha} \bar{p}_{\mathbf{a},\mathbf{b}}(\alpha)\Big|_{\alpha=1^-}.$$

Stability of $\alpha = 1$ for $\gamma_l < 1$. Assume $\gamma_l < 1$ and let $\alpha \to 1^-$. Define

$$\begin{aligned} D_{\mathbf{w}^\star} &:= \frac{d}{d\alpha} \bar{p}_{\mathbf{w}^\star,\mathbf{w}^\star}(\alpha)\Big|_{\alpha=1^-} \\ &= (1 - r_v^2) + r_v^2 D_{\mathbf{v}}(\gamma_u; \lambda_u). \end{aligned}$$

At $\alpha = 1$, $\bar{p}_{\perp,\mathbf{a},\mathbf{b}} = 0$ and $\bar{\sigma}_{\text{eff}}^2 = 0$, thus simplifying

$$\frac{d}{d\alpha} E_\infty^{\text{gen}}(\alpha)\Big|_{\alpha=1^-} = \frac{\gamma_l \sigma^2}{(1 - \gamma_l)^2} - \frac{\bar{w}^*}{1 - \gamma_l} D_{\mathbf{w}^\star}.$$

Consequently, the derivative is negative when

$$\gamma_l < \frac{S D_{\mathbf{w}^\star}}{1 + S D_{\mathbf{w}^\star}}, \qquad S := \frac{\bar{w}^*}{\sigma^2},$$

so sufficiently small decreases in α increase the error. The derivative is positive when the inequality is reversed, in which case compression locally improves generalisation. Equality gives the marginal stability boundary

$$\gamma_l^{\mathrm{crit}} = \frac{SD_{\mathbf{w}^\star}}{1 + SD_{\mathbf{w}^\star}}.$$

Writing in terms of $D_{\mathbf{v}}$:

$$\gamma_l^{\mathrm{crit}} = \frac{S[(1 - r_v^2) + r_v^2 D_{\mathbf{v}}(\gamma_u; \lambda_u)]}{1 + S[(1 - r_v^2) + r_v^2 D_{\mathbf{v}}(\gamma_u; \lambda_u)]}.$$

C Infinite data limits

We consider what happens to the errors and optimal α when $\gamma_u \to 0$ or $\gamma_l \to 0$. We specify this to the case of $\mathbf{\Sigma}_u = \mathbf{\Sigma}_l$, i.e. $\lambda_u = \lambda_l = \lambda$, and $\mathbf{q} = \mathbf{v}$. The more general case is easy to obtain as well, but the resulting expressions are not as readable.

Prior data limit $\gamma_u \to 0$. In this limit, the empirical spike aligns with the population spike, and the retained subspace contains the spike direction together with an α-fraction of the isotropic bulk.

For any fixed vectors $\mathbf{a}, \mathbf{b}$, we have

$$\begin{aligned}
\bar{p}_{\mathbf{a},\mathbf{a}} &\to \rho_{\mathbf{a},\mathbf{v}}^2 + (1 - \rho_{\mathbf{a},\mathbf{v}}^2)\alpha, \\
\bar{p}_{\perp,\mathbf{a},\mathbf{a}} &\to (1 - \rho_{\mathbf{a},\mathbf{v}}^2)(1 - \alpha), \\
\bar{p}_{\mathbf{a},\mathbf{b}} &\to \rho_{\mathbf{a},\mathbf{v}}\rho_{\mathbf{b},\mathbf{v}} + \alpha\left(\rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{v}}\rho_{\mathbf{b},\mathbf{v}}\right), \\
\bar{p}_{\perp,\mathbf{a},\mathbf{b}} &\to (1 - \alpha)\left(\rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{v}}\rho_{\mathbf{b},\mathbf{v}}\right).
\end{aligned}$$

In particular,

$$\bar{p}_{\mathbf{v},\mathbf{v}} \to 1, \qquad \bar{p}_{\perp,\mathbf{v},\mathbf{v}} \to 0, \qquad \bar{\lambda} \to \lambda.$$

For the effective projections $\bar{\pi}$, the behaviour depends on γ_{eff}.

If $\gamma_{\mathrm{eff}} < 1$, then

$$\bar{\pi}_{\mathbf{a},\mathbf{a}} \to \bar{p}_{\mathbf{a},\mathbf{a}}, \qquad \bar{\pi}_{\mathbf{a},\mathbf{b}} \to \bar{p}_{\mathbf{a},\mathbf{b}}.$$

If $\gamma_{\mathrm{eff}} > 1$, define

$$A_\lambda := \frac{\lambda}{\gamma_{\mathrm{eff}} - 1 + \lambda}, \qquad A_1 := \frac{1}{\gamma_{\mathrm{eff}}},$$

then

$$\begin{aligned}
\bar{\pi}_{\mathbf{a},\mathbf{a}} &\to \rho_{\mathbf{a},\mathbf{v}}^2 A_\lambda + (1 - \rho_{\mathbf{a},\mathbf{v}}^2)\alpha A_1, \\
\bar{\pi}_{\mathbf{a},\mathbf{b}} &\to \rho_{\mathbf{a},\mathbf{v}}\rho_{\mathbf{b},\mathbf{v}} A_\lambda + \alpha\left(\rho_{\mathbf{a},\mathbf{b}} - \rho_{\mathbf{a},\mathbf{v}}\rho_{\mathbf{b},\mathbf{v}}\right) A_1.
\end{aligned}$$

Errors in the limit $\gamma_u \to 0$. In the infinite-prior-data limit, the empirical spike is recovered exactly. Hence

$$\bar{p}_{\mathbf{v},\mathbf{v}} \to 1, \qquad \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{v}} \to 0, \qquad \bar{\lambda} \to \lambda,$$

and

$$\bar{p}_{\mathbf{w}^\star,\mathbf{w}^\star} \to r_v^2 + (1 - r_v^2)\alpha, \qquad \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star} \to (1 - r_v^2)(1 - \alpha).$$

Therefore

$$\bar{\sigma}_{\mathrm{eff}}^2 \to (1 - r_v^2)(1 - \alpha).$$

Since the leak term vanishes,

$$\boxed{E_\infty^{\mathrm{est}} \to \bar{w}^*\left(1 - \bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}\right) + \frac{\min\{\gamma_{\mathrm{eff}}, 1\}}{|\gamma_{\mathrm{eff}} - 1|}\left[\bar{w}^*(1 - r_v^2)(1 - \alpha) + \sigma^2\right],}$$

where

$$\bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star} = \begin{cases} r_v^2 + (1 - r_v^2)\alpha, & \gamma_{\mathrm{eff}} < 1, \\ r_v^2 \dfrac{\lambda}{\gamma_{\mathrm{eff}} - 1 + \lambda} + (1 - r_v^2)\dfrac{1}{\gamma_l}, & \gamma_{\mathrm{eff}} > 1. \end{cases}$$

Moreover, we obtain

$$\boxed{E_\infty^{\text{gen}} \to E_\infty^{\text{est}} + \sigma^2 + \bar{w}^*(\lambda-1)r_v^2 \begin{cases} 0, & \gamma_{\text{eff}} < 1 \\ \left(\frac{\gamma_{\text{eff}}-1}{\gamma_{\text{eff}}-1+\lambda}\right)^2, & \gamma_{\text{eff}} > 1 \end{cases}.}$$

Finally, the training error satisfies

$$\boxed{E_\infty^{\text{train}} \to \left[\sigma^2 + \bar{w}^*(1-r_v^2)(1-\alpha)\right] \max\{0, 1-\gamma_{\text{eff}}\}.}$$

Downstream-data limit $\gamma_l \to 0$. For any fixed $\alpha > 0$, we have $\gamma_{\text{eff}} = \alpha\gamma_l \to 0$, hence

$$\bar{\pi}_{\mathbf{a},\mathbf{a}} \to \bar{p}_{\mathbf{a},\mathbf{a}}, \qquad \bar{\pi}_{\mathbf{a},\mathbf{b}} \to \bar{p}_{\mathbf{a},\mathbf{b}}.$$

Moreover, the variance factor satisfies

$$\frac{\min\{\gamma_{\text{eff}}, 1\}}{|\gamma_{\text{eff}} - 1|} = \frac{\alpha\gamma_l}{1-\alpha\gamma_l} \to 0,$$

so all downstream finite-sample variance contributions vanish.

The pretraining projections $\bar{p}$ remain determined by γ_u.

At $\alpha = 1$,

$$\bar{p}_{\mathbf{a},\mathbf{a}} = 1, \qquad \bar{p}_{\perp,\mathbf{a},\mathbf{a}} = 0, \qquad \bar{\pi}_{\mathbf{a},\mathbf{a}} \to 1,$$

and

$$\bar{p}_{\mathbf{a},\mathbf{b}} = \rho_{\mathbf{a},\mathbf{b}}, \qquad \bar{p}_{\perp,\mathbf{a},\mathbf{b}} = 0, \qquad \bar{\pi}_{\mathbf{a},\mathbf{b}} \to \rho_{\mathbf{a},\mathbf{b}}.$$

At the spike-retaining endpoint $\alpha \to 0$,

$$\bar{\pi}_{\mathbf{a},\mathbf{a}}(0) \to \bar{p}_{\mathbf{a},\mathbf{a}}(0), \qquad \bar{\pi}_{\mathbf{a},\mathbf{b}}(0) \to \bar{p}_{\mathbf{a},\mathbf{b}}(0),$$

with $\bar{p}$ determined by the prior overlap $c(\gamma_u)$.

Errors in the limit $\gamma_l \to 0$. For fixed $\alpha > 0$, $\gamma_{\text{eff}} = \alpha\gamma_l \to 0$. Hence

$$\bar{\pi}_{\mathbf{a},\mathbf{a}} \to \bar{p}_{\mathbf{a},\mathbf{a}}, \qquad \bar{\pi}_{\mathbf{a},\mathbf{b}} \to \bar{p}_{\mathbf{a},\mathbf{b}},$$

and

$$\frac{\min\{\gamma_{\text{eff}}, 1\}}{|\gamma_{\text{eff}} - 1|} \to 0.$$

Therefore the estimation error satisfies

$$\boxed{E_\infty^{\text{est}} \to \bar{w}^* \left[1 - \bar{p}_{\mathbf{w}^\star,\mathbf{w}^\star} + \frac{(\lambda-1)^2}{\bar{\lambda}^2} \bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{v}} \bar{p}_{\mathbf{v},\mathbf{v}}\right].}$$

The generalisation error satisfies

$$\boxed{E_\infty^{\text{gen}} \to E_\infty^{\text{est}} + \bar{w}^*(\lambda-1)\left(r_v - \bar{p}_{\mathbf{w}^\star,\mathbf{v}} - \frac{\lambda-1}{\bar{\lambda}} \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{v}} \bar{p}_{\mathbf{v},\mathbf{v}}\right)^2 + \sigma^2.} \tag{45}$$

The training error satisfies

$$\boxed{E_\infty^{\text{train}} \to \sigma^2 + \bar{w}^* \left[\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star} + \frac{\lambda-1}{\bar{\lambda}} \bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{v}}\right].} \tag{46}$$

Corollary 3. *Assume $\gamma_u = p/n_u \to 0$, so that the empirical spike direction converges to the population spike* $\mathbf{v}$. *Let $\gamma_{\text{eff}} = \alpha\gamma_l$, $S = \bar{w}^*/\sigma^2$. Then the estimation error becomes*

$$E_\infty^{\text{est}} \to \bar{w}^*(1 - \bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star}) + \frac{\min\{\gamma_{\text{eff}}, 1\}}{|\gamma_{\text{eff}} - 1|} \left[\bar{w}^*(1-r_v^2)(1-\alpha) + \sigma^2\right],$$

where

$$\bar{\pi}_{\mathbf{w}^\star,\mathbf{w}^\star} = \begin{cases} r_v^2 + (1-r_v^2)\alpha, & \gamma_{\text{eff}} < 1, \\ r_v^2 \dfrac{\lambda}{\gamma_{\text{eff}} - 1 + \lambda} + (1-r_v^2)\dfrac{\alpha}{\gamma_{\text{eff}}}, & \gamma_{\text{eff}} > 1. \end{cases}$$

For matched test covariance $\boldsymbol{\Sigma}_{\text{new}} = \boldsymbol{\Sigma}$*,*

$$E_\infty^{\text{gen}} \to E_\infty^{\text{est}} + \bar{w}^*(\lambda-1)r_v^2 \begin{cases} 0, & \gamma_{\text{eff}} < 1, \\ \left(\dfrac{\gamma_{\text{eff}}-1}{\gamma_{\text{eff}}-1+\lambda}\right)^2, & \gamma_{\text{eff}} > 1, \end{cases} + \sigma^2.$$

Finally,

$$E_\infty^{\text{train}} \to \left[\sigma^2 + \bar{w}^*(1-r_v^2)(1-\alpha)\right] \max\{0, 1-\gamma_{\text{eff}}\}.$$

Consequently, the phase transitions in this limit can either be derived by taking derivatives and limiting values α, or by substituting the limits directly in Corollaries 1 and 2. In this regime, the phase boundaries become independent of γ_u and depend only on r_v^2, SNR, λ, and γ_l.

Corollary 4 (Phase transitions for $\gamma_u \to 0$)**.** *In the limit* $\gamma_u \to 0$*, the spike direction is recovered exactly. Let* $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}_u, \mathbf{q} = \mathbf{v}$*. Hence,* $c(\gamma_u) \to 1$ *and* $D_{\mathbf{v}}(\gamma_u) \to 0$*. Thus, the transition curve when* $\gamma_l > 1$ *simplifies to*

$$(1-r_v^2)\frac{1}{\gamma_l} = r_v^2 \frac{\lambda\gamma_l(\gamma_l-1)}{(\gamma_l-1+\lambda)^2} + \frac{1}{S(\gamma_l-1)}.$$

Furthermore, when $\gamma_l < 1$*, the stability boundary reduces to*

$$\gamma_l^{\text{crit}} = \frac{S(1-r_v^2)}{1+S(1-r_v^2)}.$$

D Regularized regression

We consider the ridge estimator

$$\hat{\mathbf{w}}_\rho = \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n_l}\|\mathbf{y}_l - \mathbf{X}_l\mathcal{P}_m\mathbf{w}\|^2 + \rho\|\mathbf{w}\|^2, \qquad \rho > 0. \tag{47}$$

The loss is normalized by n_l so that ρ remains fixed in the high-dimensional limit.

Let $\mathbf{X}_m := \mathbf{X}_l\mathbf{U}_m$ and $\mathbf{S}_m := n_l^{-1}\mathbf{X}_m^\top\mathbf{X}_m$. As before, the minimizer lies in the $\mathcal{P}_m$ subspace and is

$$\hat{\mathbf{w}}_\rho = \mathbf{U}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{y}_l. \tag{48}$$

Define the regularized effective projection

$$\Pi_\rho := \mathbf{U}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\mathbf{S}_m\mathbf{U}_m^\top. \tag{49}$$

As $\rho \to 0$, this converges to the row-space projector $\Pi = \mathbf{U}_m\mathbf{X}_m^\dagger\mathbf{X}_m\mathbf{U}_m^\top$. For $\rho > 0$, Π_ρ is generally not idempotent, so the error analysis requires both Π_ρ and Π_ρ^2. We next state all lemmas and theorems that differ from the unregularized setup.

Spectrum of the regularized effective projection.

Lemma 8 (Spectrum of Π_ρ)**.** *Let*

$$\mathbf{S}_m = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top, \qquad \mathbf{B} = \operatorname{diag}(b_1, \ldots, b_m),$$

be the eigendecomposition of $\mathbf{S}_m$*. Then*

$$\Pi_\rho = \mathbf{U}_m\mathbf{Q}\operatorname{diag}\left(\frac{b_1}{b_1+\rho}, \ldots, \frac{b_m}{b_m+\rho}\right)\mathbf{Q}^\top\mathbf{U}_m^\top. \tag{50}$$

Moreover, the empirical spectral distribution of the retained block of Π_ρ *converges weakly almost surely to the pushforward of* $F_{\gamma_{\text{eff}}}$ *under the map*

$$h_\rho(x) = \frac{x}{x+\rho}.$$

That is,

$$\frac{1}{m}\sum_{i=1}^{m}\delta_{b_i/(b_i+\rho)} \Rightarrow (h_\rho)_{\#}F_{\gamma_{\text{eff}}} \qquad \textit{almost surely.}$$

Proof. Since $\mathbf{S}_m$ is symmetric positive semidefinite,

$$\mathbf{S}_m + \rho\mathbf{I}_m = \mathbf{Q}(\mathbf{B} + \rho\mathbf{I}_m)\mathbf{Q}^\top,$$

and hence

$$(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\mathbf{S}_m = \mathbf{Q}\,\mathrm{diag}\left(\frac{b_1}{b_1+\rho}, \ldots, \frac{b_m}{b_m+\rho}\right)\mathbf{Q}^\top.$$

Substitution into (49) gives (50). By Lemma 1, the empirical spectral distribution of $\mathbf{S}_m$ converges weakly almost surely to $F_{\gamma_{\text{eff}}}$. Since $h_\rho(x) = x/(x+\rho)$ is bounded and continuous on $[0,\infty)$ for every $\rho > 0$, the continuous mapping theorem gives the claimed pushforward limit. □

Regularized bilinear effective projections. We next state the ridge analogue of Lemmas 6 and 7. Since Π_ρ is not idempotent, the estimation error depends on both Π_ρ and Π_ρ^2.

As before, for deterministic vectors $\mathbf{a}, \mathbf{b}$, define the retained-subspace overlaps

$$\frac{\mathbf{a}^\top\mathcal{P}_m\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \to \bar{p}_{\mathbf{a},\mathbf{b}}, \qquad \bar{p}_{\mathbf{a},\mathbf{b}} = \rho_{\mathbf{a},\mathbf{b}} - \bar{p}_{\perp,\mathbf{a},\mathbf{b}}..$$

These quantities depend on the pretraining covariance $\mathbf{\Sigma}_u = \mathbf{I}_p + (\lambda_u - 1)\mathbf{v}\mathbf{v}^\top$ through the learned subspace $\mathcal{P}_m$.

Conditioned on $\mathbf{U}_m$, however, the downstream retained-coordinate covariance is

$$\mathbf{\Sigma}_{mm} = \mathbf{U}_m^\top\mathbf{\Sigma}_l\mathbf{U}_m = \mathbf{I}_m + (\lambda_l - 1)(\mathbf{U}_m^\top\mathbf{q})(\mathbf{U}_m^\top\mathbf{q})^\top.$$

Its effective spike eigenvalue converges to

$$\bar{\lambda}_l := 1 + (\lambda_l - 1)\bar{p}_{\mathbf{q},\mathbf{q}}. \tag{51}$$

Assume $\bar{p}_{\mathbf{a},\mathbf{a}} > 0$ and $\bar{p}_{\mathbf{b},\mathbf{b}} > 0$, and define

$$\tilde{\mathbf{a}} := \frac{\mathbf{U}_m^\top\mathbf{a}}{\|\mathcal{P}_m\mathbf{a}\|}, \qquad \tilde{\mathbf{b}} := \frac{\mathbf{U}_m^\top\mathbf{b}}{\|\mathcal{P}_m\mathbf{b}\|}.$$

The limiting signed population spectral measure of $(\tilde{\mathbf{a}}, \tilde{\mathbf{b}})$ with respect to $\mathbf{\Sigma}_{mm}$ is

$$\begin{aligned}
&\widetilde{G}_{\mathbf{a},\mathbf{b}} = (\omega_{\mathbf{a},\mathbf{b}} - \omega_{\mathbf{a},\mathbf{q}}\omega_{\mathbf{b},\mathbf{q}})\,\delta_1 + \omega_{\mathbf{a},\mathbf{q}}\omega_{\mathbf{b},\mathbf{q}}\delta_{\bar{\lambda}_l}, \\
&\omega_{\mathbf{a},\mathbf{b}} := \frac{\bar{p}_{\mathbf{a},\mathbf{b}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{-1/2}\bar{p}_{\mathbf{b},\mathbf{b}}^{-1/2}}, \qquad \omega_{\mathbf{a},\mathbf{q}} := \frac{\bar{p}_{\mathbf{a},\mathbf{q}}}{\bar{p}_{\mathbf{a},\mathbf{a}}^{-1/2}\bar{p}_{\mathbf{q},\mathbf{q}}^{-1/2}}, \qquad \omega_{\mathbf{b},\mathbf{q}} := \frac{\bar{p}_{\mathbf{b},\mathbf{q}}}{\bar{p}_{\mathbf{b},\mathbf{b}}^{-1/2}\bar{p}_{\mathbf{q},\mathbf{q}}^{-1/2}}.
\end{aligned} \tag{52}$$

Let $G^{l,\text{sample}}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}$ denote the corresponding limiting signed sample spectral measure after observing n_l downstream samples in m retained dimensions, obtained from $\widetilde{G}_{\mathbf{a},\mathbf{b}}$ through the same deterministic equivalent as in Lemma 3, but with aspect ratio γ_{eff}.

Lemma 9 (Regularized bilinear effective projection)**.** *For fixed $\rho > 0$ and deterministic vectors $\mathbf{a}, \mathbf{b}$ with $\bar{p}_{\mathbf{a},\mathbf{a}}, \bar{p}_{\mathbf{b},\mathbf{b}} > 0$, we have, for $k = 1, 2$,*

$$\frac{\mathbf{a}^\top\Pi_\rho^k\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \xrightarrow{\text{a.s.}} \pi^{(k)}_{\rho,\mathbf{a},\mathbf{b}} := \sqrt{\bar{p}_{\mathbf{a},\mathbf{a}}\bar{p}_{\mathbf{b},\mathbf{b}}}\int\left(\frac{x}{x+\rho}\right)^k dG^{l,\text{sample}}_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}(x). \tag{53}$$

Proof. Write $\mathbf{S}_m = n_l^{-1}(\mathbf{X}_l\mathbf{U}_m)^\top(\mathbf{X}_l\mathbf{U}_m)$. Since

$$\Pi_\rho = \mathbf{U}_m\mathbf{S}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\mathbf{U}_m^\top,$$

the result follows from convergence of the retained-subspace overlaps and the downstream signed spectral measures, applied to the bounded continuous functions $(x/(x+\rho))^k$. □

Regularized variance trace. The variance terms in the ridge estimator involve

$$\frac{1}{n_l}\mathrm{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right],$$

rather than the pseudoinverse trace appearing in the ridgeless case.

Lemma 10 (Regularized variance trace)**.** *Under the assumptions of Section 3, for fixed $\rho > 0$,*

$$\frac{1}{n_l}Tr\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right] \xrightarrow{\text{a.s.}} V_\rho(\gamma_{\text{eff}}), \tag{54}$$

where

$$V_\rho(\gamma_{\text{eff}}) := \gamma_{\text{eff}} \int \frac{x}{(x+\rho)^2}\, dF_{\gamma_{\text{eff}}}(x). \tag{55}$$

Proof. We write

$$\frac{1}{n_l}\mathrm{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right] = \frac{m}{n_l}\frac{1}{m}\mathrm{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right].$$

By Lemma 1, the empirical spectral distribution of $\mathbf{S}_m$ converges weakly almost surely to $F_{\gamma_{\text{eff}}}$. Since $x \mapsto x/(x+\rho)^2$ is bounded and continuous on $[0,\infty)$, weak convergence gives

$$\frac{1}{m}\mathrm{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right] \xrightarrow{\text{a.s.}} \int \frac{x}{(x+\rho)^2}\, dF_{\gamma_{\text{eff}}}(x).$$

Multiplying by $m/n_l \to \gamma_{\text{eff}}$ gives the claim. □

D.1 Decomposition of the ridge estimator

We now repeat the estimator decomposition from Appendix A.2, with the hard projector Π replaced by Π_ρ. Recall from the conditional Gaussian decomposition that

$$\mathbf{X}_\perp = \mathbf{X}_m\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp} + \mathbf{E},$$

where $\mathbf{E}$ is independent of $\mathbf{X}_m$. Starting from (48), substitute

$$\mathbf{y}_l = \mathbf{X}_l\mathbf{w}^\star + \xi = \mathbf{X}_m\mathbf{U}_m^\top\mathbf{w}^\star + \mathbf{X}_\perp\mathbf{U}_\perp^\top\mathbf{w}^\star + \xi.$$

Using

$$\mathbf{X}_\perp = \mathbf{X}_m\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp} + \mathbf{E}$$

and

$$\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{\Sigma}_{m\perp}\mathbf{U}_\perp^\top\mathbf{w}^\star = A\mathbf{U}_m^\top\mathbf{q},$$

where

$$A := \frac{(\lambda_l-1)\mathbf{w}^{\star\top}\mathcal{P}_\perp\mathbf{q}}{\tilde{\lambda}_l}, \qquad \tilde{\lambda}_l := 1+(\lambda_l-1)\|\mathcal{P}_m\mathbf{q}\|^2.$$

We obtain

$$\begin{aligned}\hat{\mathbf{w}}_\rho &= \mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\mathbf{S}_m\mathbf{U}_m^\top\mathbf{w}^\star + A\mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\mathbf{S}_m\mathbf{U}_m^\top\mathbf{q}\\ &\quad + \mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star + \mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\xi.\end{aligned}$$

The first two terms are precisely $\Pi_\rho\mathbf{w}^\star$ and $A\Pi_\rho\mathbf{q}$, thus the final decomposition is:

$$\begin{aligned}\hat{\mathbf{w}}_\rho &= \Pi_\rho\mathbf{w}^\star + A\Pi_\rho\mathbf{q}\\ &\quad + \mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star + \mathbf{U}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\xi.\end{aligned} \tag{56}$$

D.2 Asymptotic estimation error for ridge regression

The ridge estimator decomposition from (56) can be written as

$$\hat{\mathbf{w}}_\rho = \Pi_\rho(\mathbf{w}^\star + A\mathbf{q}) + \mathbf{z}_E + \mathbf{z}_\xi, \tag{57}$$

where

$$\mathbf{z}_E := \mathbf{U}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\mathbf{E}\mathbf{U}_\perp^\top\mathbf{w}^\star, \qquad \mathbf{z}_\xi := \mathbf{U}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-1}\frac{1}{n_l}\mathbf{X}_m^\top\xi.$$

Thus the deterministic part of the estimator is the regularized projection of the effective signal $\mathbf{w}^\star + A\mathbf{q}$.

Define

$$\mathbf{e}_\star := \frac{\mathbf{w}^\star}{\|\mathbf{w}^\star\|}, \qquad c_p := \frac{A}{\|\mathbf{w}^\star\|} = \frac{\lambda_l - 1}{\bar{\lambda}_l}\frac{\mathbf{w}^{\star\top}\mathcal{P}_\perp\mathbf{q}}{\|\mathbf{w}^\star\|}. \tag{58}$$

In the high-dimensional limit,

$$c_p \xrightarrow{\text{a.s.}} c_\star := \frac{\lambda_l - 1}{\bar{\lambda}_l}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}, \qquad \bar{\lambda}_l = 1 + (\lambda_l - 1)\bar{p}_{m,\mathbf{q},\mathbf{q}}. \tag{59}$$

We then define

$$\mathbf{u}_\star := \mathbf{e}_\star + c_\star\mathbf{q}, \qquad s_\star := \|\mathbf{u}_\star\|. \tag{60}$$

Equivalently,

$$s_\star^2 = 1 + c_\star^2 + 2c_\star r_q, \qquad r_q := \lim\frac{\mathbf{w}^{\star\top}\mathbf{q}}{\|\mathbf{w}^\star\|}. \tag{61}$$

Finally, define

$$\tau^2 := \bar{w}^*\bar{\sigma}^2_{\text{eff}} + \sigma^2, \qquad \bar{\sigma}^2_{\text{eff}} = \bar{p}_{\perp,\mathbf{w}^\star,\mathbf{w}^\star} + \frac{\lambda_l - 1}{\bar{\lambda}_l}\bar{p}^2_{\perp,\mathbf{w}^\star,\mathbf{q}}. \tag{62}$$

Theorem 4 (Asymptotic estimation error for ridge regression)**.** *Under the assumptions of Section 3, for fixed $\rho > 0$,*

$$\boxed{E^{\text{est}}_{\infty,\rho} = \bar{w}^*\left[1 - 2s_\star\pi^{(1)}_{\rho,\mathbf{e}_\star,\mathbf{u}_\star} + s_\star^2\pi^{(2)}_{\rho,\mathbf{u}_\star,\mathbf{u}_\star}\right] + \tau^2 V_\rho(\gamma_{\text{eff}}).} \tag{63}$$

Proof. From (57),

$$\hat{\mathbf{w}}_\rho = \Pi_\rho(\mathbf{w}^\star + A\mathbf{q}) + \mathbf{z}_E + \mathbf{z}_\xi.$$

Conditionally on $\mathbf{X}_m$ and $\mathbf{U}_m$, the random vectors $\mathbf{z}_E$ and $\mathbf{z}_\xi$ are centred and independent of the deterministic bias term. Hence the corresponding cross terms vanish by the same conditional-centering and concentration arguments as in the proof of Theorem 1.

For the deterministic bias,

$$\mathbf{w}^\star + A\mathbf{q} = \|\mathbf{w}^\star\|(\mathbf{e}_\star + c_p\mathbf{q}),$$

and $c_p \to c_\star$ almost surely. Therefore,

$$\frac{1}{\|\mathbf{w}^\star\|^2}\left\|\mathbf{w}^\star - \Pi_\rho(\mathbf{w}^\star + A\mathbf{q})\right\|^2 \to \left\|\mathbf{e}_\star - \Pi_\rho\mathbf{u}_\star\right\|^2.$$

Expanding,

$$\left\|\mathbf{e}_\star - \Pi_\rho\mathbf{u}_\star\right\|^2 = 1 - 2\mathbf{e}_\star^\top\Pi_\rho\mathbf{u}_\star + \mathbf{u}_\star^\top\Pi_\rho^2\mathbf{u}_\star.$$

By Lemma 9,

$$\mathbf{e}_\star^\top\Pi_\rho\mathbf{u}_\star \xrightarrow{\text{a.s.}} s_\star\pi^{(1)}_{\rho,\mathbf{e}_\star,\mathbf{u}_\star},$$

and

$$\mathbf{u}_\star^\top\Pi_\rho^2\mathbf{u}_\star \xrightarrow{\text{a.s.}} s_\star^2\pi^{(2)}_{\rho,\mathbf{u}_\star,\mathbf{u}_\star}.$$

Using $\|\mathbf{w}^\star\|^2 \to \bar{w}^*$ gives the bias contribution in (63).

For the variance terms, conditionally on $\mathbf{X}_m, \mathbf{U}_m$,

$$\mathbb{E}_\xi\|\mathbf{z}_\xi\|^2 = \frac{\sigma^2}{n_l}\text{Tr}\left[\mathbf{S}_m(\mathbf{S}_m + \rho\mathbf{I}_m)^{-2}\right].$$

Similarly,

$$\mathbb{E}_E\|\mathbf{z}_E\|^2 = \frac{\|\mathbf{w}^\star\|^2\sigma_{\text{eff}}^2}{n_l}\text{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right],$$

where

$$\sigma_{\text{eff}}^2 \xrightarrow{\text{a.s.}} \bar{\sigma}_{\text{eff}}^2.$$

Thus

$$\mathbb{E}\|\mathbf{z}_E+\mathbf{z}_\xi\|^2 = \left(\|\mathbf{w}^\star\|^2\sigma_{\text{eff}}^2+\sigma^2\right)\frac{1}{n_l}\text{Tr}\left[\mathbf{S}_m(\mathbf{S}_m+\rho\mathbf{I}_m)^{-2}\right].$$

By Lemma 10, this converges almost surely to $\tau^2 V_\rho(\gamma_{\text{eff}})$. Combining the bias and variance terms gives (63). □

D.3 Asymptotic generalisation error for ridge regression

For a fresh test point drawn from the downstream distribution,

$$\mathbf{x}\sim\mathcal{N}(0,\mathbf{\Sigma}_l), \qquad \mathbf{\Sigma}_l = \mathbf{I}_p+(\lambda_l-1)\mathbf{q}\mathbf{q}^\top, \qquad y=\mathbf{x}^\top\mathbf{w}^\star+\xi,$$

the prediction error is

$$E_{p,\rho}^{\text{gen}} = (\hat{\mathbf{w}}_\rho-\mathbf{w}^\star)^\top\mathbf{\Sigma}_l(\hat{\mathbf{w}}_\rho-\mathbf{w}^\star)+\sigma^2. \tag{64}$$

Hence,

$$E_{p,\rho}^{\text{gen}} = \|\hat{\mathbf{w}}_\rho-\mathbf{w}^\star\|^2+(\lambda_l-1)\left[\mathbf{q}^\top(\hat{\mathbf{w}}_\rho-\mathbf{w}^\star)\right]^2+\sigma^2. \tag{65}$$

The first term is exactly the estimation error analysed in Theorem 4. It remains to compute the residual along the downstream spike.

Using the grouped ridge decomposition,

$$\hat{\mathbf{w}}_\rho = \Pi_\rho(\mathbf{w}^\star+A\mathbf{q})+\mathbf{z}_E+\mathbf{z}_\xi,$$

we obtain

$$\frac{\mathbf{q}^\top(\mathbf{w}^\star-\hat{\mathbf{w}}_\rho)}{\|\mathbf{w}^\star\|} = r_q-\frac{\mathbf{q}^\top\Pi_\rho\mathbf{w}^\star}{\|\mathbf{w}^\star\|}-\frac{A}{\|\mathbf{w}^\star\|}\mathbf{q}^\top\Pi_\rho\mathbf{q} - \frac{\mathbf{q}^\top\mathbf{z}_E}{\|\mathbf{w}^\star\|}-\frac{\mathbf{q}^\top\mathbf{z}_\xi}{\|\mathbf{w}^\star\|}.$$

The deterministic terms converge almost surely as

$$\frac{\mathbf{q}^\top\Pi_\rho\mathbf{w}^\star}{\|\mathbf{w}^\star\|}\to\pi^{(1)}_{\rho,\mathbf{w}^\star,\mathbf{q}}, \qquad \mathbf{q}^\top\Pi_\rho\mathbf{q}\to\pi^{(1)}_{\rho,\mathbf{q},\mathbf{q}},$$

and

$$\frac{A}{\|\mathbf{w}^\star\|}\to\frac{\lambda_l-1}{\bar{\lambda}_l}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}.$$

The random terms projected onto the fixed direction $\mathbf{q}$ vanish almost surely by the same argument used in the proof of Theorem 2. Therefore,

$$\frac{\mathbf{q}^\top(\mathbf{w}^\star-\hat{\mathbf{w}}_\rho)}{\|\mathbf{w}^\star\|}\xrightarrow{\text{a.s.}} r_q-\pi^{(1)}_{\rho,\mathbf{w}^\star,\mathbf{q}}-\frac{\lambda_l-1}{\bar{\lambda}_l}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}\pi^{(1)}_{\rho,\mathbf{q},\mathbf{q}}. \tag{66}$$

Theorem 5 (Asymptotic generalisation error for ridge regression)**.** *Under the same assumptions as above,*

$$\boxed{E_{\infty,\rho}^{\text{gen}} = E_{\infty,\rho}^{\text{est}}+\bar{w}^*(\lambda_l-1)\left[r_q-\pi^{(1)}_{\rho,\mathbf{w}^\star,\mathbf{q}}-\frac{\lambda_l-1}{\bar{\lambda}_l}\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{q}}\pi^{(1)}_{\rho,\mathbf{q},\mathbf{q}}\right]^2+\sigma^2.} \tag{67}$$

Proof. The decomposition above shows that the deterministic spike residual converges to (66). The terms involving $\mathbf{z}_E$ and $\mathbf{z}_\xi$ along the fixed direction $\mathbf{q}$ vanish almost surely by the same conditional-Gaussian concentration argument as in the proof of Theorem 2. Substitution into the decomposition of $E_{p,\rho}^{\text{gen}}$ gives the result. □

D.4 Asymptotic training error for ridge regression

For completeness, we also record the training error. This is no longer the simple interpolation expression from the ridgeless case because ridge shrinkage leaves residual error even in the retained signal subspace.

Define

$$\kappa_{\rho,\mathbf{a},\mathbf{b}} := \bar{p}_{m,\mathbf{a},\mathbf{a}}^{1/2} \bar{p}_{m,\mathbf{b},\mathbf{b}}^{1/2} \int \frac{\rho^2 x}{(x+\rho)^2}\, dG_{\tilde{\mathbf{a}},\tilde{\mathbf{b}}}^{l,\mathrm{sample}}(x), \tag{68}$$

and

$$T_\rho(\gamma_{\mathrm{eff}}) := 1 - 2\gamma_{\mathrm{eff}} \int \frac{x}{x+\rho}\, dF_{\gamma_{\mathrm{eff}}}(x) + \gamma_{\mathrm{eff}} \int \left(\frac{x}{x+\rho}\right)^2 dF_{\gamma_{\mathrm{eff}}}(x). \tag{69}$$

Theorem 6 (Asymptotic training error for ridge regression)**.** *Under the same assumptions as above, the ridge training error converges almost surely to*

$$E_{\infty,\rho}^{\mathrm{train}} = \bar{w}^* \left[\kappa_{\rho,\mathbf{w}^\star,\mathbf{w}^\star} + 2c_\star \kappa_{\rho,\mathbf{w}^\star,\mathbf{q}} + c_\star^2 \kappa_{\rho,\mathbf{q},\mathbf{q}}\right] + T_\rho(\gamma_{\mathrm{eff}}) \left(\bar{w}^* \bar{\sigma}_{\mathrm{eff}}^2 + \sigma^2\right). \tag{70}$$

Proof. Let

$$H_\rho := \frac{1}{n_l} \mathbf{X}_m (\mathbf{S}_m + \rho \mathbf{I}_m)^{-1} \mathbf{X}_m^\top.$$

Its nonzero eigenvalues are

$$\frac{b_i}{b_i + \rho},$$

where b_i are the eigenvalues of $\mathbf{S}_m$. The training residual is

$$\mathbf{y}_l - \mathbf{X}_l \mathcal{P}_m \hat{\mathbf{w}}_\rho = (\mathbf{I}_{n_l} - H_\rho)\mathbf{y}_l.$$

Using the same conditional decomposition as before,

$$\mathbf{y}_l = \mathbf{X}_m \mathbf{U}_m^\top \mathbf{w}^\star + A \mathbf{X}_m \mathbf{U}_m^\top \mathbf{q} + \mathbf{E}\mathbf{U}_\perp^\top \mathbf{w}^\star + \xi.$$

For the deterministic retained signal part, set

$$\mathbf{s} := \mathbf{U}_m^\top \mathbf{w}^\star + A \mathbf{U}_m^\top \mathbf{q}.$$

Then

$$(\mathbf{I}_{n_l} - H_\rho)\mathbf{X}_m \mathbf{s} = \rho \mathbf{X}_m (\mathbf{S}_m + \rho \mathbf{I}_m)^{-1} \mathbf{s}.$$

Therefore

$$\frac{1}{n_l} \left\|(\mathbf{I}_{n_l} - H_\rho)\mathbf{X}_m \mathbf{s}\right\|^2 = \mathbf{s}^\top \rho^2 \mathbf{S}_m (\mathbf{S}_m + \rho \mathbf{I}_m)^{-2} \mathbf{s}.$$

Applying the same regularized signed spectral-measure result with the function $x \mapsto \rho^2 x/(x+\rho)^2$, and using $A/\|\mathbf{w}^\star\| \to c_\star$, gives the first line of (70).

For the independent model-mismatch and label-noise terms, conditionally on $\mathbf{X}_m, \mathbf{U}_m$,

$$\frac{1}{n_l} \mathbb{E}\left[\left\|(\mathbf{I}_{n_l} - H_\rho)(\mathbf{E}\mathbf{U}_\perp^\top \mathbf{w}^\star + \xi)\right\|^2\right] = \left(\|\mathbf{w}^\star\|^2 \sigma_{\mathrm{eff}}^2 + \sigma^2\right) \frac{1}{n_l} \mathrm{Tr}[(\mathbf{I}_{n_l} - H_\rho)^2].$$

Since

$$\frac{1}{n_l} \mathrm{Tr}[(\mathbf{I}_{n_l} - H_\rho)^2] = 1 - \frac{2}{n_l} \mathrm{Tr} H_\rho + \frac{1}{n_l} \mathrm{Tr} H_\rho^2,$$

and the nonzero eigenvalues of H_ρ are $b_i/(b_i + \rho)$,

$$\frac{1}{n_l} \mathrm{Tr}[(\mathbf{I}_{n_l} - H_\rho)^2] \xrightarrow{\text{a.s.}} T_\rho(\gamma_{\mathrm{eff}}).$$

Finally, $\sigma_{\mathrm{eff}}^2 \to \bar{\sigma}_{\mathrm{eff}}^2$ almost surely, which proves the claim. □

E Theory-experiment validation

Below, we present additional plots validating the theoretical predictions through numerical experiments. We introduce an overlap measure for the task-spike alignment θ, which can be defined through the overlap r_v^2 as

$$\frac{(\mathbf{w}^{\star\top} \mathbf{v})^2}{\|\mathbf{w}^\star\|^2} = r_v^2 = (\cos\theta)^2.$$

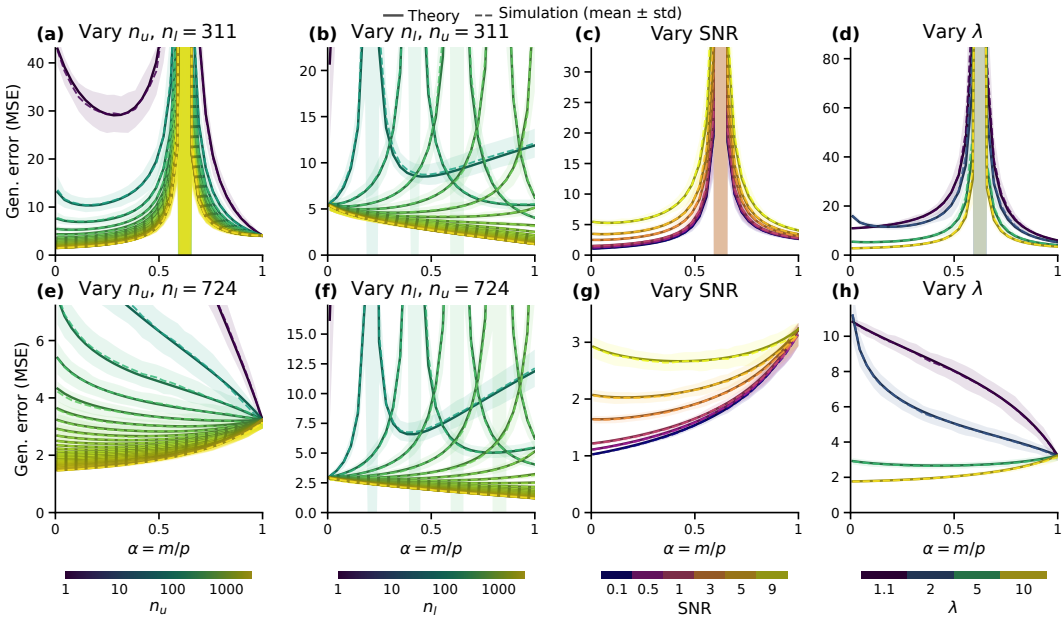


Figure 5: Generalisation error of PCR as a function of the number of retained components $\alpha = \frac{m}{p}$. Solid lines show the theoretical prediction; dashed lines show the empirical mean $\pm$ one standard deviation across trials. Each column varies one parameter while holding the others fixed: **(a,e)** varying the pre-training dataset size n_u with n_l fixed; **(b,f)** varying the downstream dataset size n_l with n_u fixed; **(c,g)** varying the signal-to-noise ratio SNR; **(d,h)** varying the spike eigenvalue λ. The two rows correspond to two regimes: smaller (top) and larger (bottom) values of n_u and n_l. Results are shown for a rank-1 spiked covariance model with perfectly aligned signal ($r_v^2 = 1$).

E.1 Error curves

Figure 5 illustrates the range of generalisation error curves across parameter settings. When the downstream sample size is limited ($n_l < p$), the error exhibits a double descent profile, with the peak occurring at $\alpha \approx 1/\gamma_l$. Increasing the size of the pretraining dataset n_u uniformly reduces the overall error scale (panels **(a,e)**), while varying n_l shifts the location of the interpolation threshold (panels **(b,f)**). Increasing the SNR raises the generalisation error (panels **(c,g)**), as it amplifies the contribution of signal-dependent bias. This increase is non-uniform across α: smaller values of α are affected more strongly, since a larger fraction of the signal lies outside the retained subspace. A similar non-uniform effect is observed when varying the spike strength λ (panels **(d,h)**).

We further decompose these effects by examining the estimation error components in Figure 6. The bias due to missing signal is determined by the downstream sample regime. In the overparametrized regime ($n_l < p$), this bias admits a strictly positive lower bound (panel **(a)**), reflecting the inability to fully reconstruct the signal. In contrast, when $n_l > p$, this component can vanish (panel **(e)**). Independently of this constraint, the bias decreases with α, since increasing α enlarges the subspace $\mathcal{P}_m$ and thus increases the fraction of $\mathbf{w}^\star$ that can be represented. The leak-induced bias follows a similar decreasing trend in α, as it is controlled by $\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{v}}$, which shrinks as the orthogonal complement $\mathcal{P}_\perp$ becomes smaller. However, its dependence on λ is non-monotonic (panels **(b,f)**): it is negligible for weak spikes (e.g. $\lambda = 1.1$), increases at intermediate values (e.g. $\lambda = 2, 5$), and decreases again for strong spikes (e.g. $\lambda = 10$). This reflects a trade-off between decreasing leak size $\bar{p}_{\perp,\mathbf{w}^\star,\mathbf{v}}$ and increasing the effective component of the leak vector $\mathbf{v}$ in the subspace $\mathbf{X}_l\mathcal{P}_m$ through $\bar{\pi}_{\mathbf{v}}$. For the parameter ranges considered, the magnitude of the leak term remains small relative to the other components. The variance behaves as expected: in the regime $n_l < p$, it decreases monotonically with λ, reflecting improved signal-to-noise separation (panels **(c,g)**).

Finally, the generalisation error (Figure 7) includes two additional contributions relative to the estimation error. The first is the additive noise variance σ^2 (panels **(c,g)**). The second is an additional bias term induced by the spike structure in the test distribution. This term combines a residual missing

signal component with a correction from the leak, and it is amplified by λ. Its structure leads to a discontinuity at $\alpha = \gamma_l$ (panel **(b)**), inherited from the behaviour of the "missing signal" bias. Overall, the generalisation error closely tracks the estimation error, with the main differences appearing at small α, where these additional bias contributions are most pronounced (panels **(e,f)**).

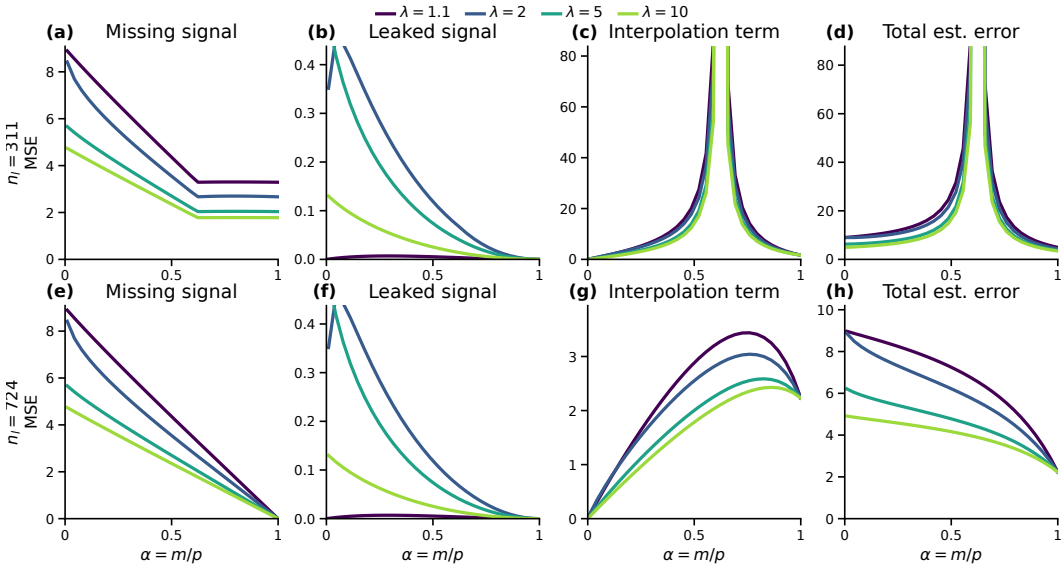


Figure 6: Components of the theoretical estimation error as a function of α for different values of λ. The total error (panels **(d, h)**) is decomposed into the missing signal term (panels **(a, e)**), the leak term (panels **(b, f)**), the variance term (panels **(c, g)**). All plots assume fixed parameters $n_u = 311$, $n_l = [311, 724]$, $\mathrm{SNR} = \frac{\|\mathbf{w}^\star\|^2}{\sigma^2}$, and a target signal misaligned with the spike eigenvector ($r_v^2 = 0.75$). The number of features is set to $p = 500$.

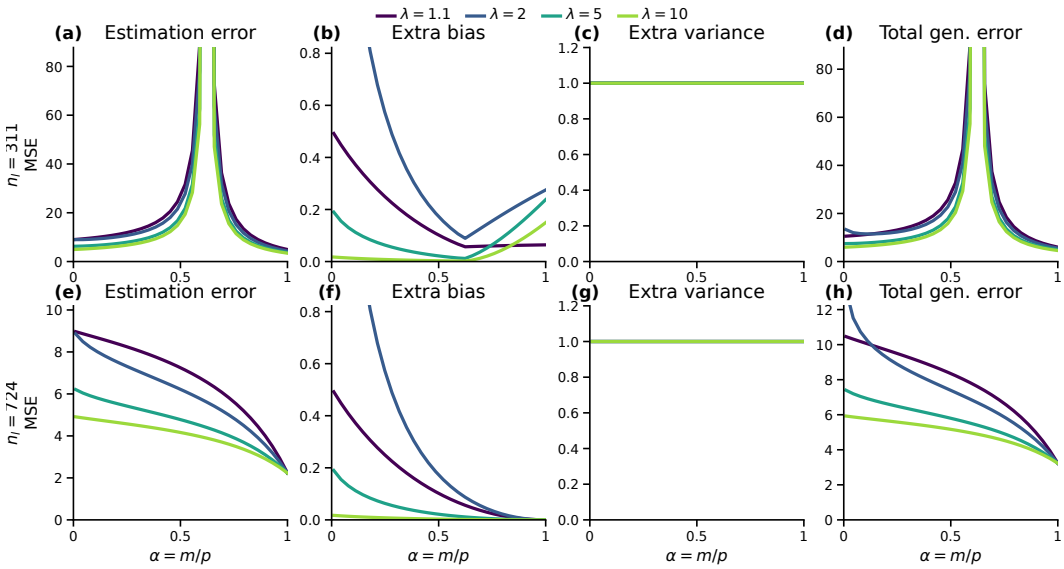


Figure 7: Components of the theoretical generalisation error as a function of α for different values of λ. The total error (panels **(d, h)**) is decomposed into the estimation error (panels **(a, e)**), the missing signal (panels **(b, f)**), and the leak term **(c, g)**. All plots assume fixed parameters $n_u = 311$, $n_l = [311, 724]$, $\mathrm{SNR} = \frac{\|\mathbf{w}^\star\|^2}{\sigma^2}$, and a target signal misaligned with the spike eigenvector ($r_v^2 = 0.75$). The number of features is set to $p = 500$.

E.2 Optimal dimensionality

In Figures 8, 9, 10,11, and 12 we present various phase portraits in α, reflecting their dependence on the signal-to-noise ratio (SNR), the spike strength (λ), the task overlap r_v and the spike overlap r. For the optimal α minimizing the training error (Figures 9 and 11), the SNR, task overlap, and spike strength have little effect. In contrast, for the optimal α minimizing the generalisation error, the phase-transition boundaries and the size of the bulk region are highly sensitive to the task overlap, SNR, and spike strength (λ). Figures 8 and 10 show that the phase-transition boundary flattens into a horizontal line at a fixed n_l *(i)* as the task-spike overlap decreases, *(ii)* as the SNR increases or *(iii)* as the spike strength decreases. At the same time, the bulk region expands along the n_u axis, thereby reducing the region of parameter space in which compression is used. This behaviour can be explained by the fact that a weaker signal is harder to detect, leading to a less accurate estimate of the prior. Consequently, exploiting this poorly-estimated prior through compression is not advantageous across many regimes, even when the amount of new data is large.

Figure 13 shows the generalization-optimal α^* for the ridge extension of Section D. We omit the training-optimal plots: with a nonzero ridge parameter the training error decreases monotonically in the number of retained components, so $\alpha^* = 1$ everywhere. For generalization, increasing the ridge parameter shrinks the region of the (n_u, n_l) plane in which compression is optimal. Ridge also regularizes the downstream regression near its interpolation threshold $n_l = p$, which removes the error spike that compression otherwise mitigates there. As a result, the band of compression ($\alpha^* < 1$) that separates the low-n_u, low-n_l "no-compression" corner from the large-n_l no-compression region disappears for sufficiently large ridge. The two regions merge, and α^* no longer varies non-monotonically with n_l.

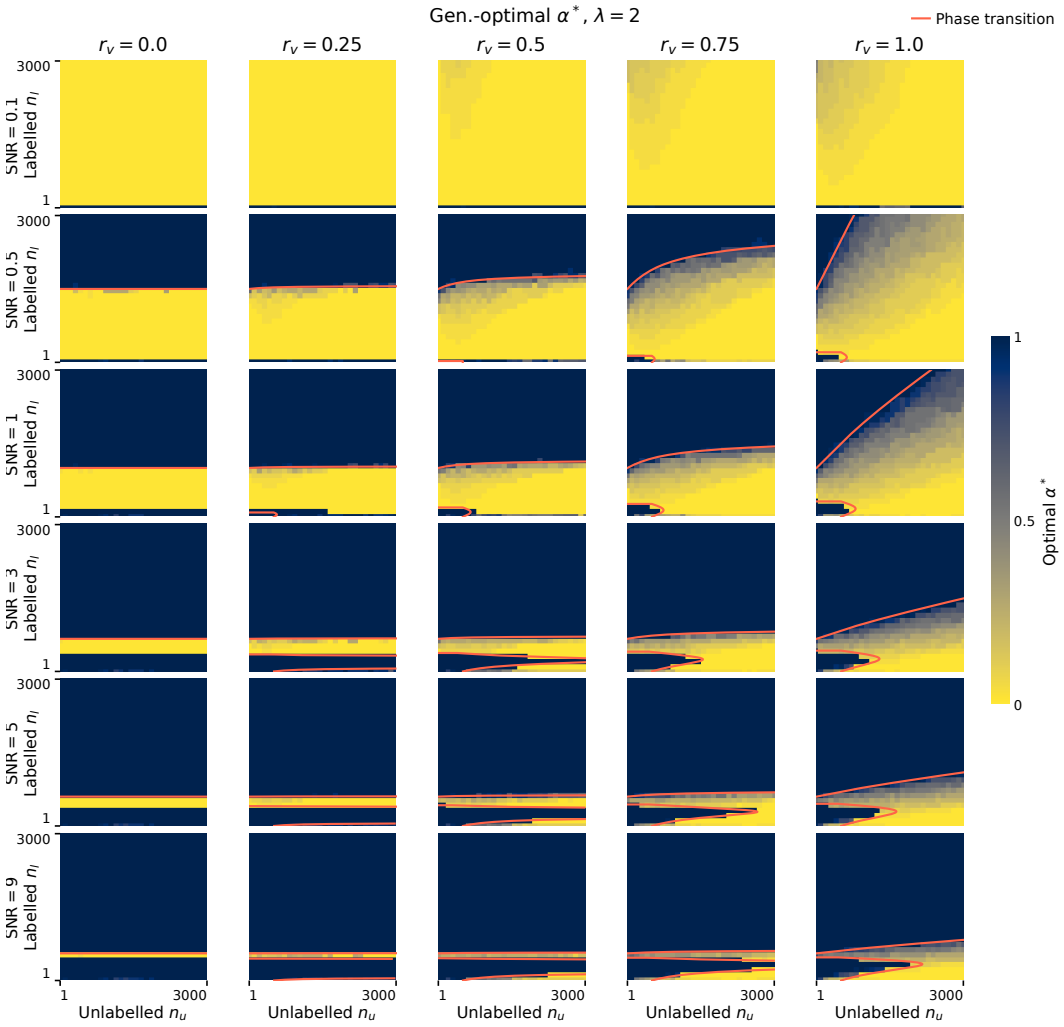


Figure 8: Optimal value of α that minimizes the generalisation error, shown as a function of SNR and task alignment r_v. We fix $\lambda = 2$. Red lines indicate the approximate phase transitions.

F Additional experiments

We provide additional figures from the experiments outlined in the main text in Section 5.

F.1 Additional autoencoder experiments

Figure 14 extends the comparison in panels **(a-d)** of Figure 4. The top row shows simulations that directly match our theoretical setting directly :PCA followed by regression. In the middle and bottom rows, PCA is replaced by unsupervised representation learning with linear and a nonlinear one-hidden-layer autoencoder, respectively. In both the spiked identity and general covariance settings, the linear autoencoder closely reproduces the theoretical predictions, consistent with its recovering the same subspace as PCA.

Figure 15 shows the corresponding results at lower SNR. The phase boundaries shift in the same qualitative way as predicted by the theory: the optimal α is low across a larger part of the (n_u, n_l) plane, indicating that the theory captures how SNR affects the benefits of compression beyond the exact PCA setting.

To examine where autoencoders depart from PCA, Figure 16 compares the theoretical overlap $\mathcal{P}_m \mathbf{v}$ with empirical overlaps obtained by passing $\mathbf{v}$ through trained linear and nonlinear autoencoders

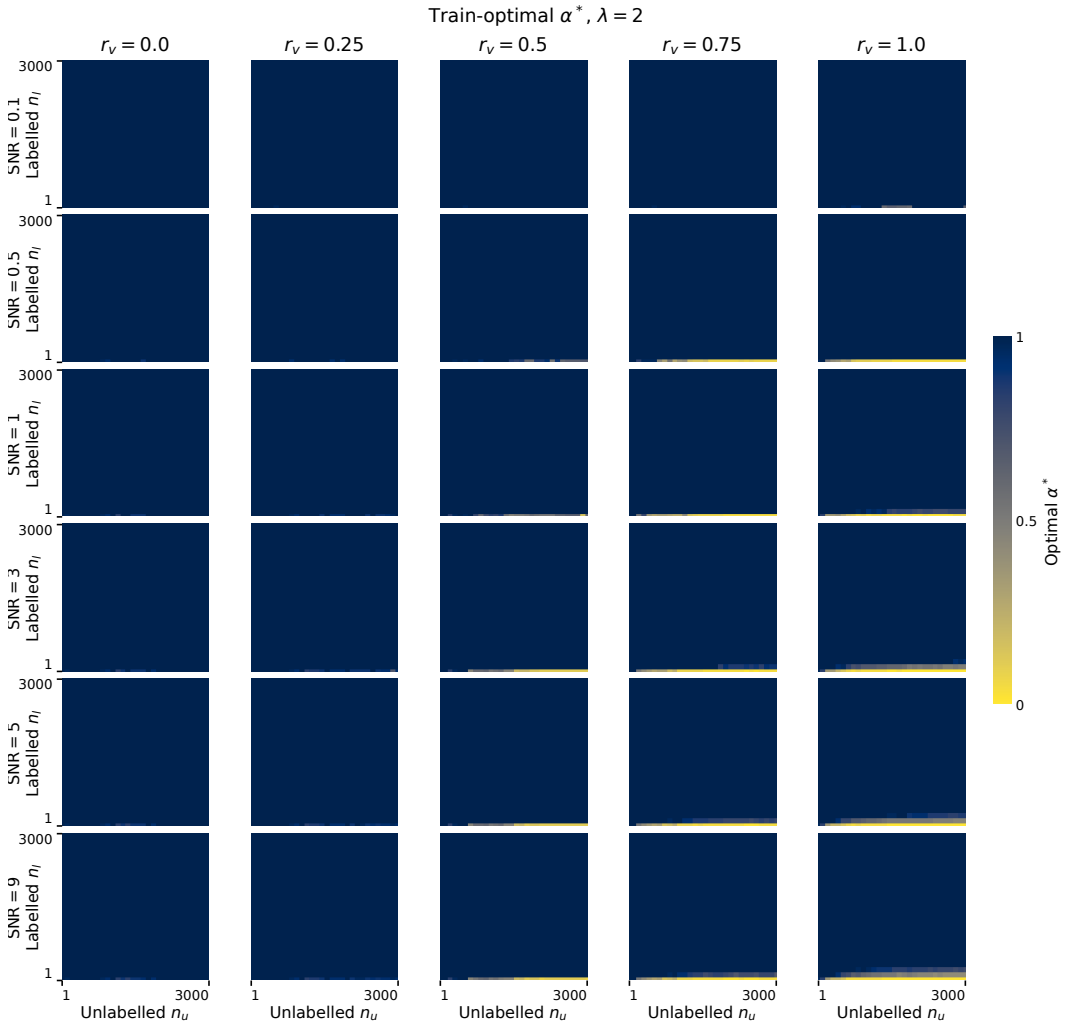


Figure 9: Optimal value of α that minimizes the training error, shown as a function of SNR and task alignment r_v. We fix $\lambda = 2$.

with varying bottleneck dimension. The linear autoencoder deviates from PCA only when $n_u < p$, whereas the nonlinear autoencoder shows larger discrepancies across a broader range of n_u. One source of these discrepancies is the "yellow band" in the plots, which arises from a double-descent effect during autoencoder pretraining. In this region, gradient descent can reduce training loss by driving the encoder's pre-activations to extreme values. The bounded sigmoid outputs then approach binary codes, allowing the model to effectively memorize a small training set while saturation destroys continuous, graded information. The linear autoencoder does not exhibit this behavior: its optimum is exact PCA, a bounded-norm solution. In our experiments, regularizing the pre-activations to keep them small resolves this effect.

Despite these deviations, when $n_u < p$, the optimal bottleneck dimension that minimizes both training and generalisation error remains close to $\alpha \approx 1$. Accordingly, the theory continues to predict the performance of linear autoencoder-based pipelines well across all values of n_u. The nonlinear autoencoder, by contrast, shows a persistent mismatch when $n_u > p$, which diminishes only with sufficiently large pretraining datasets (empirically around $n_u \approx 3p$). Its magnitude also depends on SNR, as the comparison between Figures 14 and 15 shows.

To quantify the agreement, we compare the theoretically optimal representation dimensionality with the empirical optimum for PCA followed by regression, a linear autoencoder, and a nonlinear autoencoder. The average discrepancy from the PCR theory is 7.2% for the linear autoencoder,

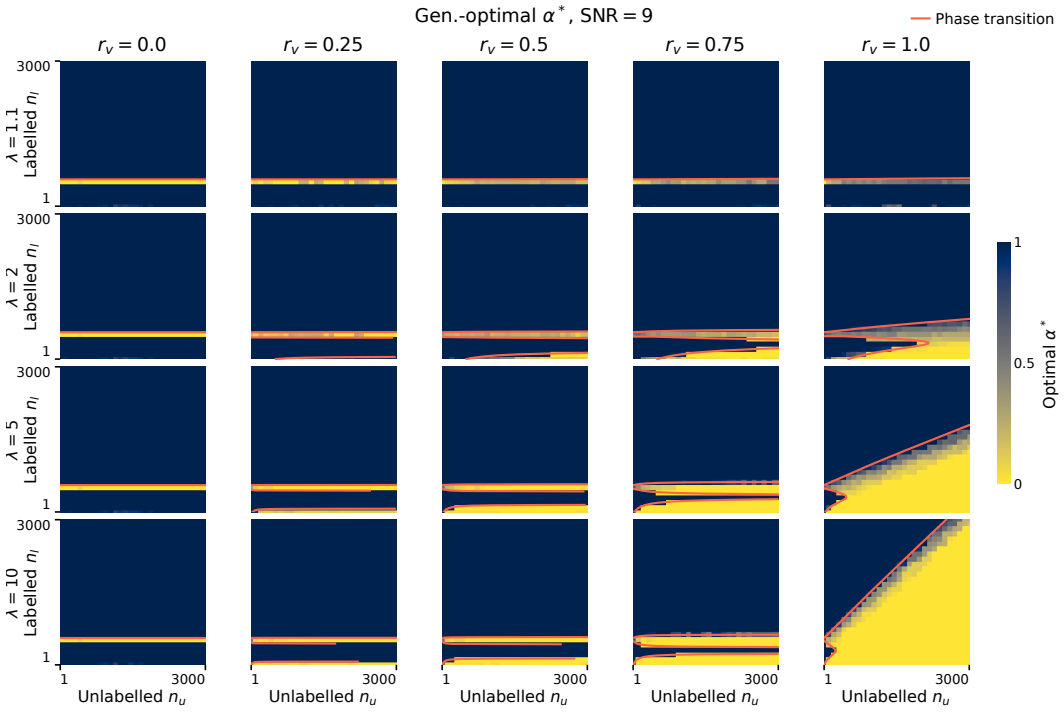


Figure 10: Optimal value of α that minimizes the generalisation error, shown as a function of spike strength λ and task alignment r_v with $\mathbf{w}^\star$. We fix $\mathrm{SNR} = 9$. Red lines indicate the approximate phase transitions.

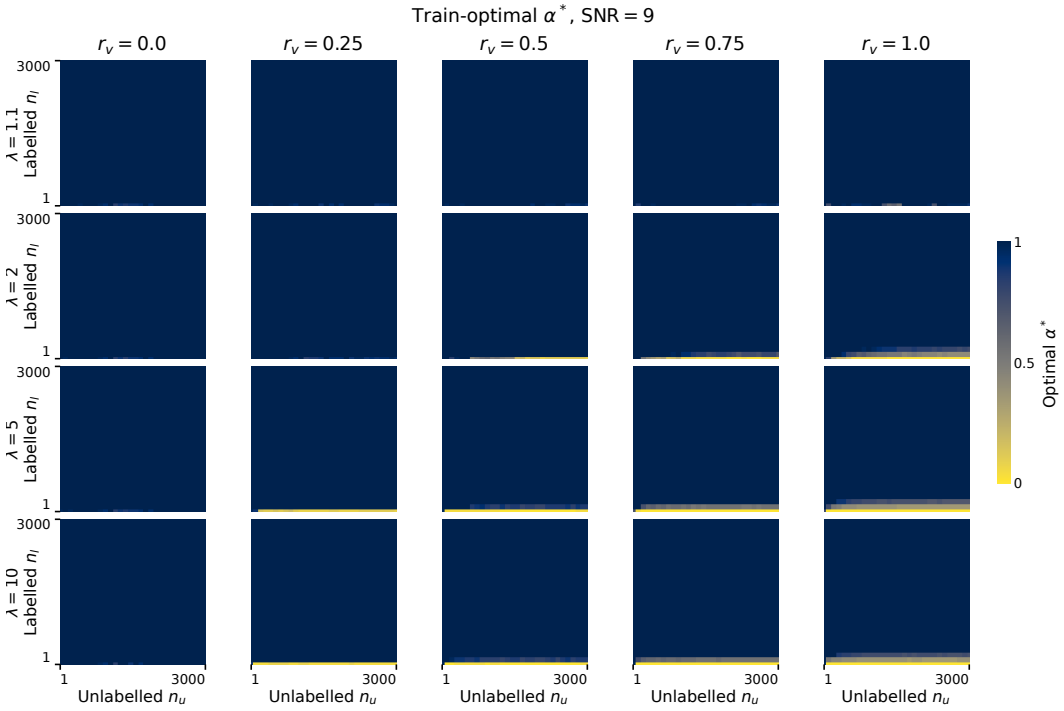


Figure 11: Optimal value of α that minimizes the training error, shown as a function of spike strength λ and task alignment r_v with $\mathbf{w}^\star$. We fix $\mathrm{SNR} = 9$.

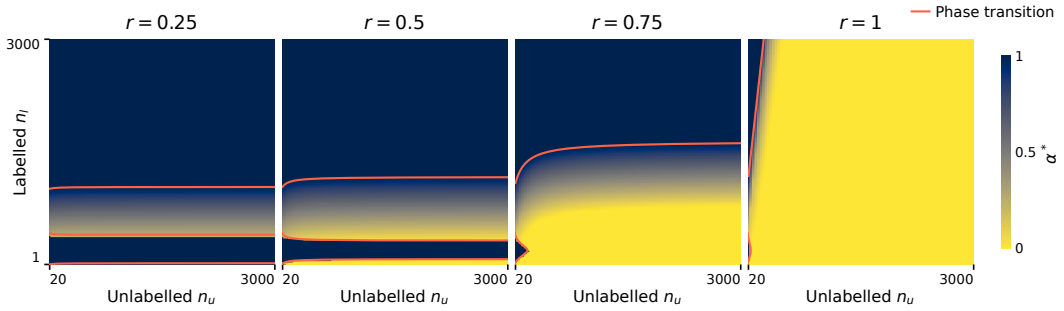


Figure 12: Optimal value of α that minimizes the generalisation error, shown as a function of prior and downstream sample sizes n_p and n_d, for varying spike alignment r (with $\lambda_u = 10$, $\lambda_\ell = 5$ fixed). We fix $\mathrm{SNR} = 5$. Red lines indicate the approximate phase transition.

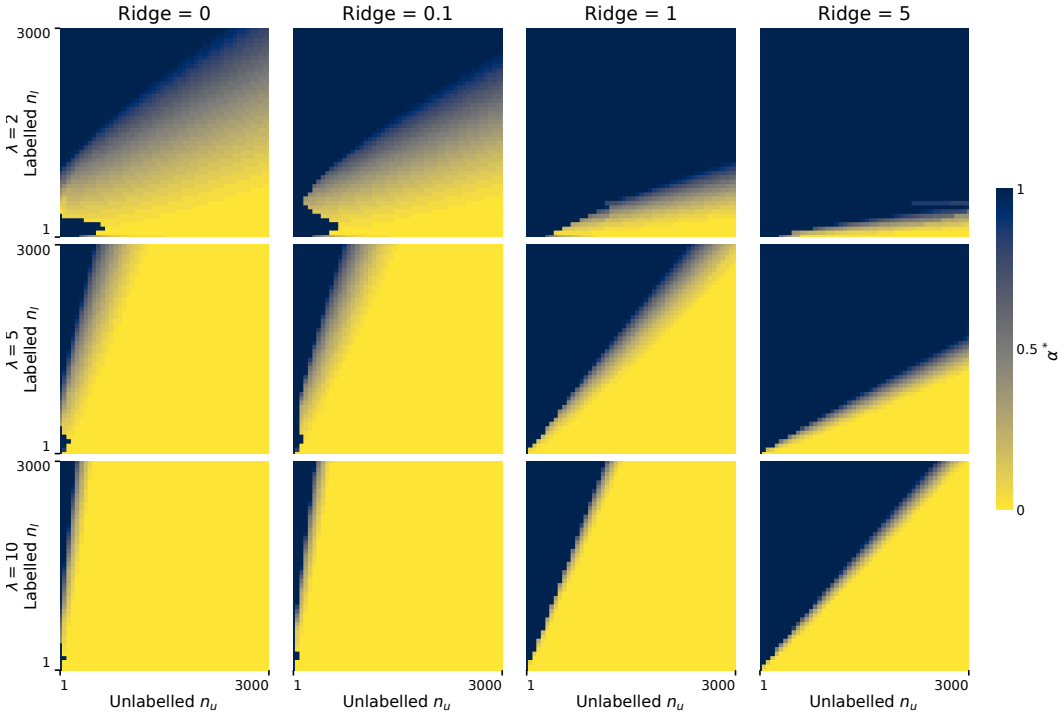


Figure 13: Optimal value of α that minimizes the generalization error in ridge regression, with ridge and spike parameters varied.

and 15.8% for the nonlinear autoencoder (Table 1). Together, these results show that the theory's main conclusions extend to both linear and nonlinear networks: compression is beneficial when downstream data are limited, and the extent of this benefit depends on SNR.

F.2 Additional transformer experiments

We first consider the setting where PCA is applied to representations extracted from the downstream task (Figure 17), rather than from pretraining data as in Figure 4**(e)**. In this regime, compression remains beneficial for generalisation when the downstream dataset is small, with an optimal intermediate dimension m. The dependence on the pretraining dataset size n_u is weaker than in the pretraining-PCA setting, but still observable.

We next return to applying PCA to pretraining data and report validation accuracy across a broader range of downstream dataset sizes (Figure 18). For small downstream datasets, an intermediate compression level is optimal, particularly for models trained for longer durations. In contrast, for

Table 1: Average absolute difference in optimal dimensionality, as a percentage of the number of features $p = 200$, between theory and PCA with regression, a linear autoencoder (AE), and a nonlinear autoencoder, averaged over the (n_u, n_l) grid. Train-opt and Gen-opt compare the training- and generalisation-optimal dimensionalities; Combined is their average.

Model	**Train-opt**	**Gen-opt**	**Combined**
Linear AE	5.59%	8.87%	7.23%
Nonlinear AE	12.95%	18.66%	15.81%
PCA + Regression	6.11%	8.08%	7.10%

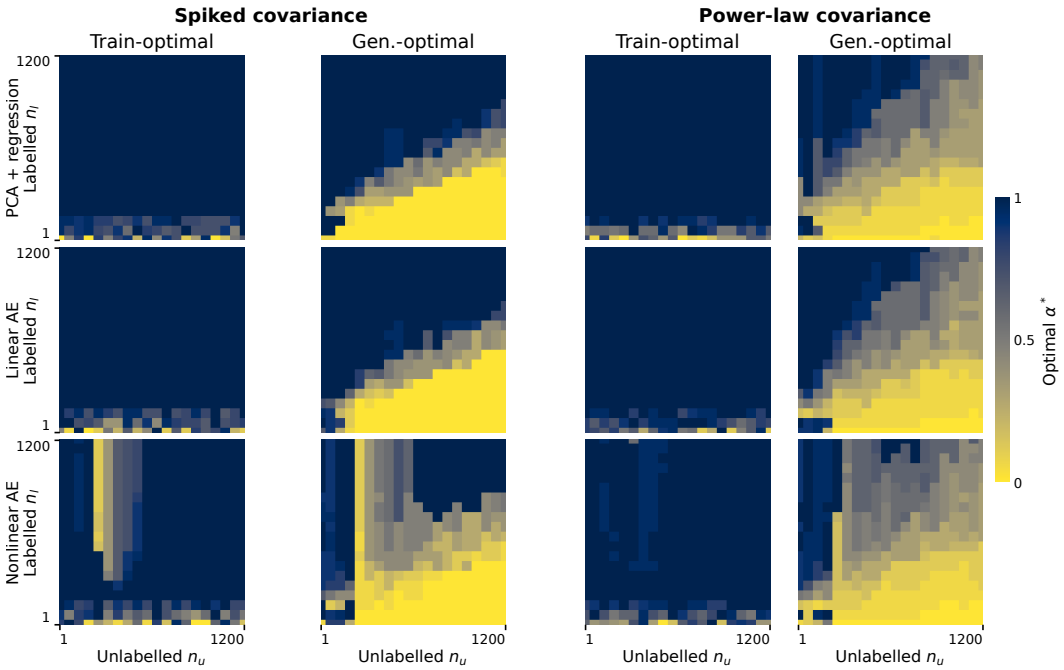


Figure 14: Extended version of panels **(a-d)** in Figure 4: optimal bottleneck size in autoencoders on inputs with different covariance structure. We compare PCA on pretraining data (top row) corresponding to our theoretical setup, a linear autoencoder (middle row) and a non-linear autoencoder (bottom row) trained on Gaussian data, with varying bottleneck size m ($\alpha = m/p$). The input covariance on the first two columns is $\mathbf{\Sigma} = \mathbf{I} + \lambda \mathbf{v}\mathbf{v}^\top$, while in the last two columns the covariance is a Toeplitz matrix $\mathbf{H}$ with a spike, i.e., $\mathbf{\Sigma} = \mathbf{H} + \lambda \mathbf{v}\mathbf{v}^\top$, where $H_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$. The autoencoder results match those of PCA, and the identity vs general covariance exhibit a similar phenomenology. We fix $\mathrm{SNR} = 9.0, \lambda = 5.0$.

large downstream datasets, performance is maximized without compression ($m \approx p$). At the smallest dataset sizes, the optimal m is less stable across runs.

Neither experimental setup provides a direct match to the theoretical model. In particular, the pretraining and downstream datasets differ not only in semantic structure but also in sequence length, which affects the learned representations and limits the transferability of PCA computed on pretraining data. Despite this mismatch, both PCA-based approaches match or outperform the no-compression baseline (Figure 19).

F.3 More datasets

We extended the transformer experiments to two model scales in the Pythia-deduped family—70M and 410M parameters—and to nine GLUE benchmark tasks: the eight classification tasks CoLA, SST-2, MRPC, QQP, QNLI, RTE, WNLI, and MNLI, together with the regression task STS-B.

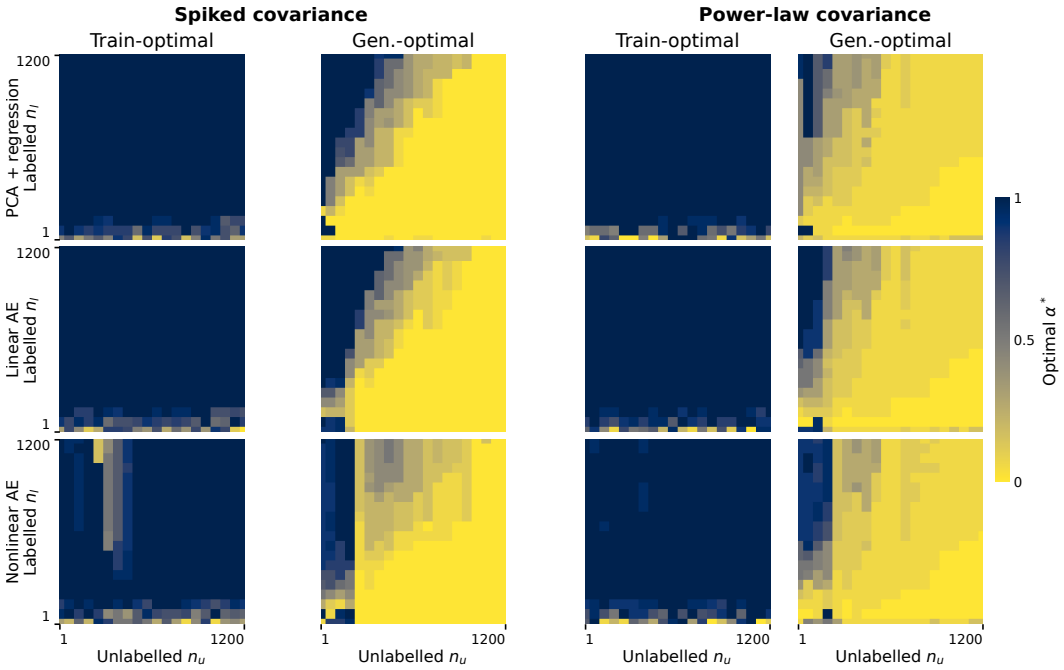


Figure 15: Effect of SNR on optimal bottleneck size in autoencoders on inputs with different covariance structure. We compare PCA on pretraining data (top row) corresponding to our theoretical setup, a linear autoencoder (middle row) and a non-linear autoencoder (bottom row) trained on Gaussian data, with varying bottleneck size m ($\alpha = m/p$). The input covariance on the first two columns is $\mathbf{\Sigma} = \mathbf{I} + \lambda \mathbf{v}\mathbf{v}^\top$, while in the last two columns the covariance is a Toeplitz matrix $\mathbf{H}$ with a spike, i.e., $\mathbf{\Sigma} = \mathbf{H} + \lambda \mathbf{v}\mathbf{v}^\top$, where $H_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$. The autoencoder results match those of PCA, and the identity vs general covariance exhibit a similar phenomenology. We fix $\mathrm{SNR} = 1.8, \lambda = 5.0$.

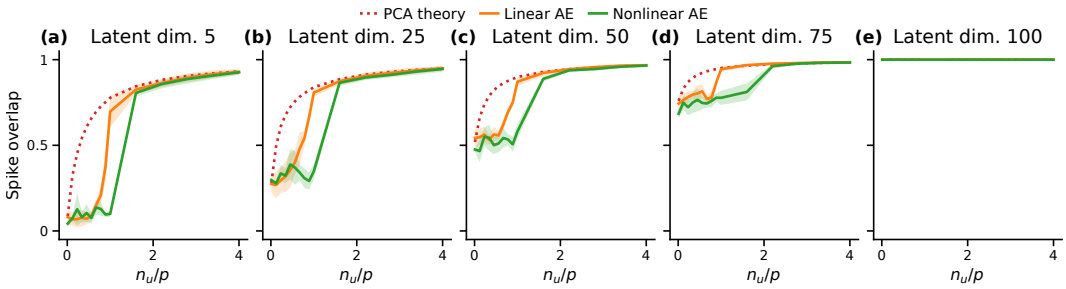


Figure 16: Spike direction recovery as a function of sample ratio n_u/p for linear and nonlinear autoencoders. Each panel shows the overlap between the singular vectors of the encoder weights and the population spike $\mathbf{v}$ as a function of the sample ratio n_u/p, for a fixed latent dimension $m \in \{5, 25, 50, 75, 100\}$. Data are generated from a spiked covariance model in $p = 100$ dimensions with spike strength $\lambda = 5$. Solid lines show empirical means over 5 random seeds; shaded bands show ± 1 standard deviation. The dotted line shows the theoretical PCA overlap $\bar{p}_{\mathbf{v},\mathbf{v}}$ derived from Lemma 3. Both linear and nonlinear autoencoders show a mismatch in low sample regimes, with a more evident mismatch in nonlinear autoencoders and in autoencoders with larger bottlenecks. For large sample regimes, linear and nonlinear autoencoders match PCA well, and show the phase transitions predicted by our theory (see panels **(a-d)** in Figure 4 discussed in Section 5.1, and Figures 14-15 discussed in Appendix F.1).

Across both model sizes and all nine datasets, we observe qualitatively similar behaviour in the validation-optimal PCR dimension, with no systematic dependence on model parameter count.

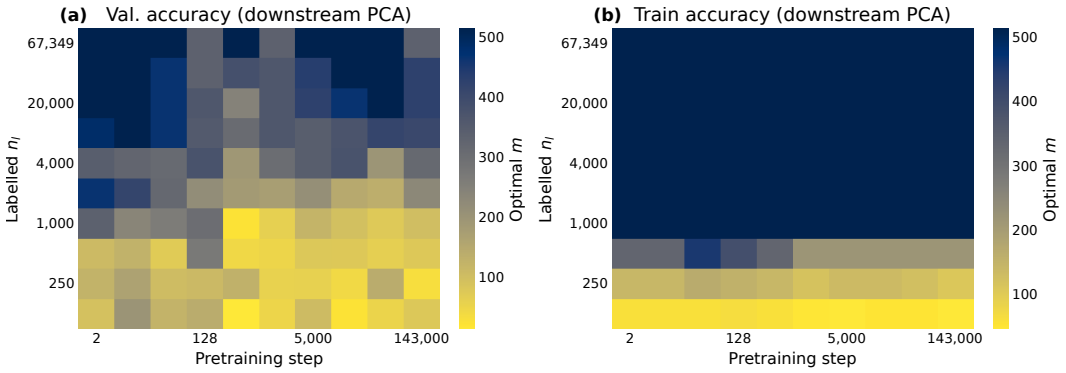


Figure 17: Optimal m using PCA on representations of the downstream task. We use the same layout as Figure 4**(e)**, but the PCA basis is fit on the downstream labelled set rather than on the pretraining corpus. **(a)** Optimal m for validation accuracy; **(b)** optimal m for training accuracy. Compared to doing PCA on the pretraining corpus (Figure 4**(e)**), downstream PCA tends to favor larger m and the dependence on the checkpoint is weaker.

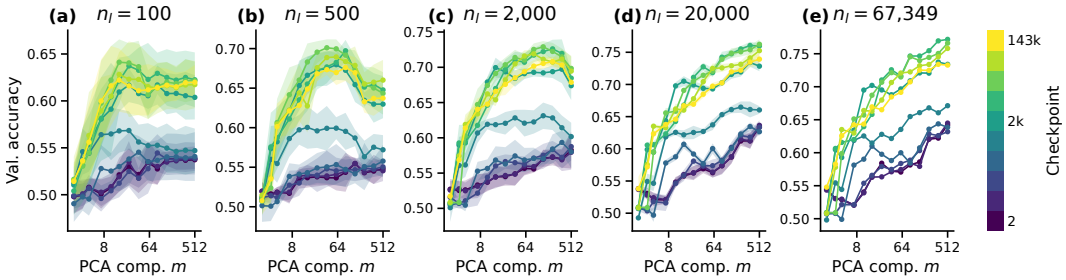


Figure 18: Validation accuracy vs number of PCA components m across five downstream task sizes. Note that the values of m are on a log scale. Each panel shows a different value of n_l spanning the full range from $n_l = 100$ to $n_l = 67{,}349$; curves are coloured by pretraining checkpoint. Shaded bands show the standard deviation over six random seeds. For smaller tasks, the curves have a clear peak and become progressively more monotone as n_l grows, reflecting that small tasks benefit from aggressive compression while large tasks can exploit finer-grained directions in representation space.

On both model sizes, we also tested the theoretical prediction that reducing the effective signal-to-noise ratio favours stronger dimensionality reduction. We independently corrupted 0%, 20%, 50%, or 80% of the downstream training targets while leaving the validation targets unchanged. For the classification tasks, each selected label was redrawn uniformly from the task's class labels; for STS-B, selected continuous similarity scores were redrawn uniformly from the observed range of training scores. The downstream- and pretrain-PCA optimal dimensionality grids for Pythia-70M are shown in Figures 20 and 21 for three of the four corruption levels (0, 0.5, 0.8; the full sweep, including the intermediate 0.2 level omitted from the figures for space, is reported in Tables 2–5).

Consistent with the theoretical prediction, increasing target noise generally shifts the validation-optimal dimension downward across datasets and both model sizes, though the per-dataset trend is not perfectly monotonic (Tables 2–5). Averaging the validation-optimal dimension across datasets, downstream PCA becomes 46.8% more compressed under the highest noise level relative to the noise-free setting for Pythia-70M (mean dimension: $185.2 \rightarrow 98.5$) and 54.2% for Pythia-410M (mean dimension: $320.8 \rightarrow 146.8$); pretraining PCA shows the same direction of effect, at 39.9% and 50.9% respectively.

Note that dataset WNLI does not follow this otherwise consistent trend: its estimated optimal dimensionality increases at higher noise levels instead of decreasing. We believe this is primarily a finite-sample instability rather than evidence of a systematic reversal of the predicted relationship. WNLI is by far the smallest GLUE task in our experiments, containing only 634 training examples, while the other datasets range up to hundreds of thousands of samples. Moreover, its development

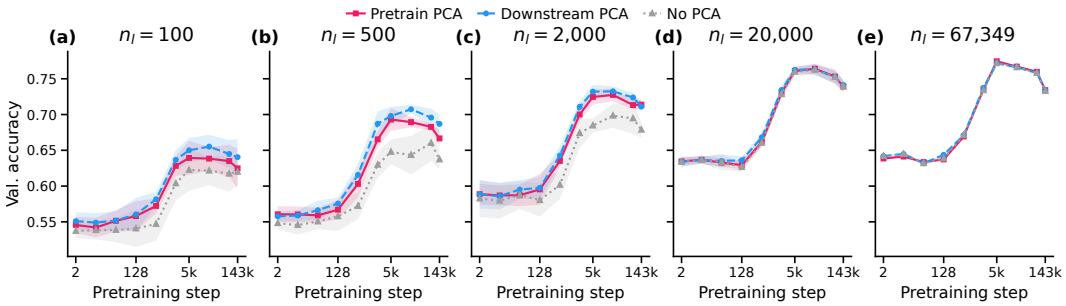


Figure 19: Validation accuracy vs pretraining checkpoint for three probing conditions across five downstream task sizes. Each panel shows a different downstream task size n_l; the x-axis is the pretraining step on a log scale. Three curves are compared: Pretrain PCA (pink, solid) fits the PCA basis on `Pile` representations from the same checkpoint and selects the best m; Downstream PCA (blue, dashed) fits PCA on the downstream task representations (`sst2`) and selects the best m; No PCA (grey, dotted) uses the full-dimensional representations (largest available m). For each curve, the displayed value is the mean over six seeds at the per-seed optimal m, with the shaded area corresponding to 1 standard deviation. The gap between PCA and the no-PCA baseline is largest at small and intermediate n_l, where dimensionality reduction acts as regularization. Pretrain PCA matches downstream PCA for intermediate and large n_l, with the mismatch coming from the difference in the type of data available in the pretraining and downstream datasets.

Dataset	Noise = 0	Noise = 0.2	Noise = 0.5	Noise = 0.8
MNLI	364.1	351.7	305.9	227.3
QQP	320.5	288.4	243.5	145.1
QNLI	307.7	274.4	232.5	159.1
SST-2	273.1	243.7	198.5	105.5
CoLA	17.0	13.3	6.8	10.4
STS-B	221.1	192.4	132.2	81.9
MRPC	41.5	24.1	14.1	18.5
RTE	113.0	85.5	89.6	103.9
WNLI	8.1	11.1	16.4	34.8

Table 2: Mean validation-optimal PCA dimension m for downstream PCA, `pythia-70m-deduped`, averaged over checkpoints, downstream training-set sizes, and seeds, for each GLUE dataset and label-noise level. Datasets are ordered by training-set size, descending. Hidden-layer size is $p = 512$.

set has a known adversarial construction artifact, as sentences showing up in training with one label often show up in validation with a flipped label [101]. Consequently, both the measured validation performance and the selected minimizing value of m are substantially more variable than for the other datasets, and we interpret WNLI as a noisy finite-sample exception rather than a systematic effect of increasing noise.

G Further discussion

G.1 Relation to classical principal component regression

Our two-stage estimator is closely related to classical principal component regression (PCR), but differs in a key structural aspect: the PCA subspace and the regression estimator are learned from independent datasets. In classical PCR [45], the principal components are computed from the same labelled design matrix used for regression. There is therefore a single sample size, whereas our setting separates the amount of unlabelled pretraining data n_u from the amount of labelled downstream data n_l.

This difference is already visible at the level of the estimator. In our setting,

$$\hat{\mathbf{w}} = (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{y}_l = (\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star + (\mathbf{X}_l \mathcal{P}_m)^\dagger \xi. \tag{71}$$

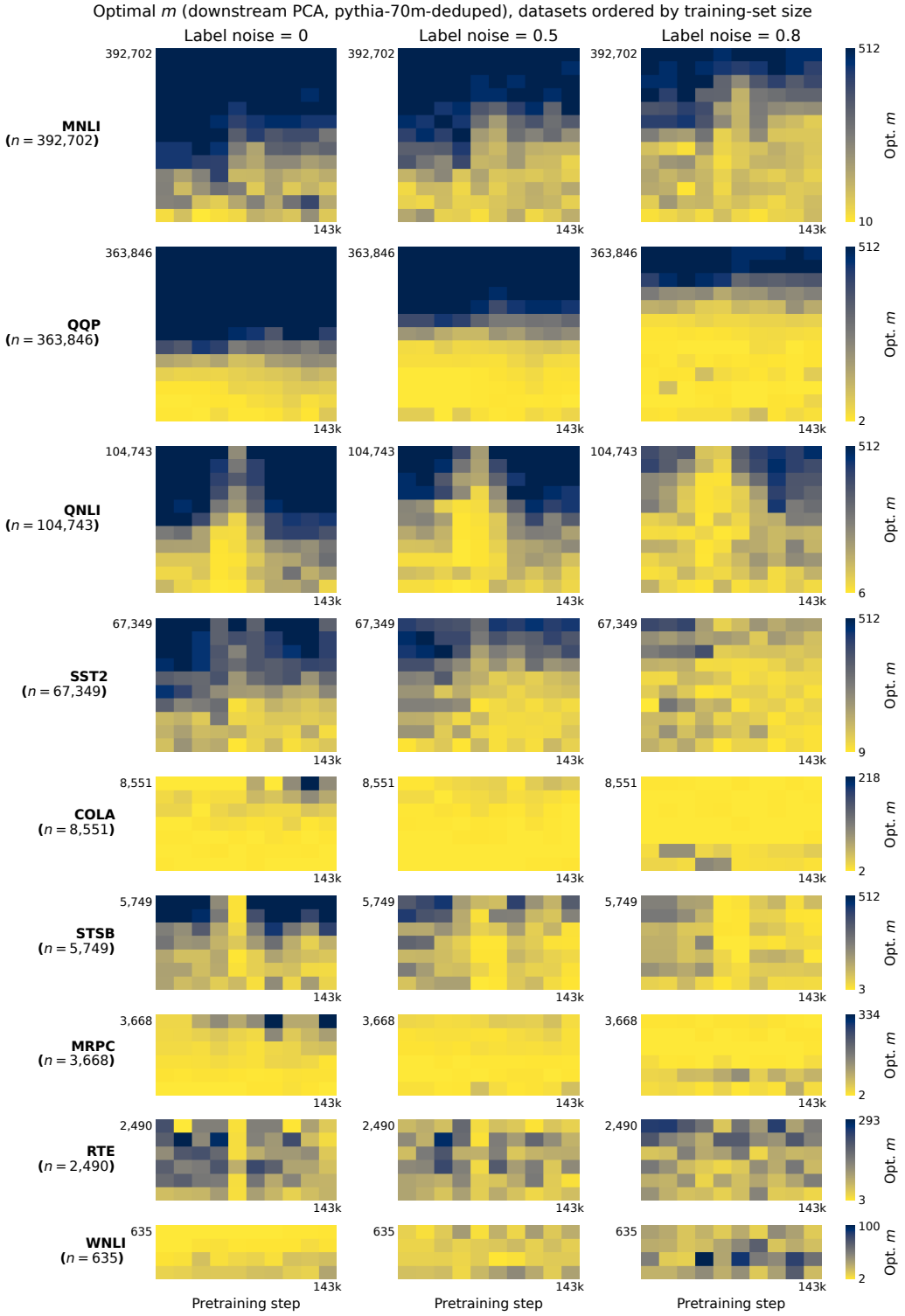


Figure 20: Validation-optimal number of PCA components m for linear probes on `pythia-70m-deduped` representations ($p = 512$), with PCA fit on the labelled downstream data, across nine GLUE datasets and three label-noise levels (columns: fraction of corrupted training targets 0, 0.5, 0.8). Each heatmap shows the optimal m as a function of pretraining checkpoint (horizontal axis, up to step 143k) and labelled training-set size n_l (vertical axis, largest at the top). The optimal m is the argmax of validation accuracy for each seed (smaller m on ties), averaged over 6 seeds. Datasets are ordered by training-set size, descending, and each dataset has its own colour scale. Increasing label noise generally lowers the optimal m, with WNLI the main exception (Table 2).

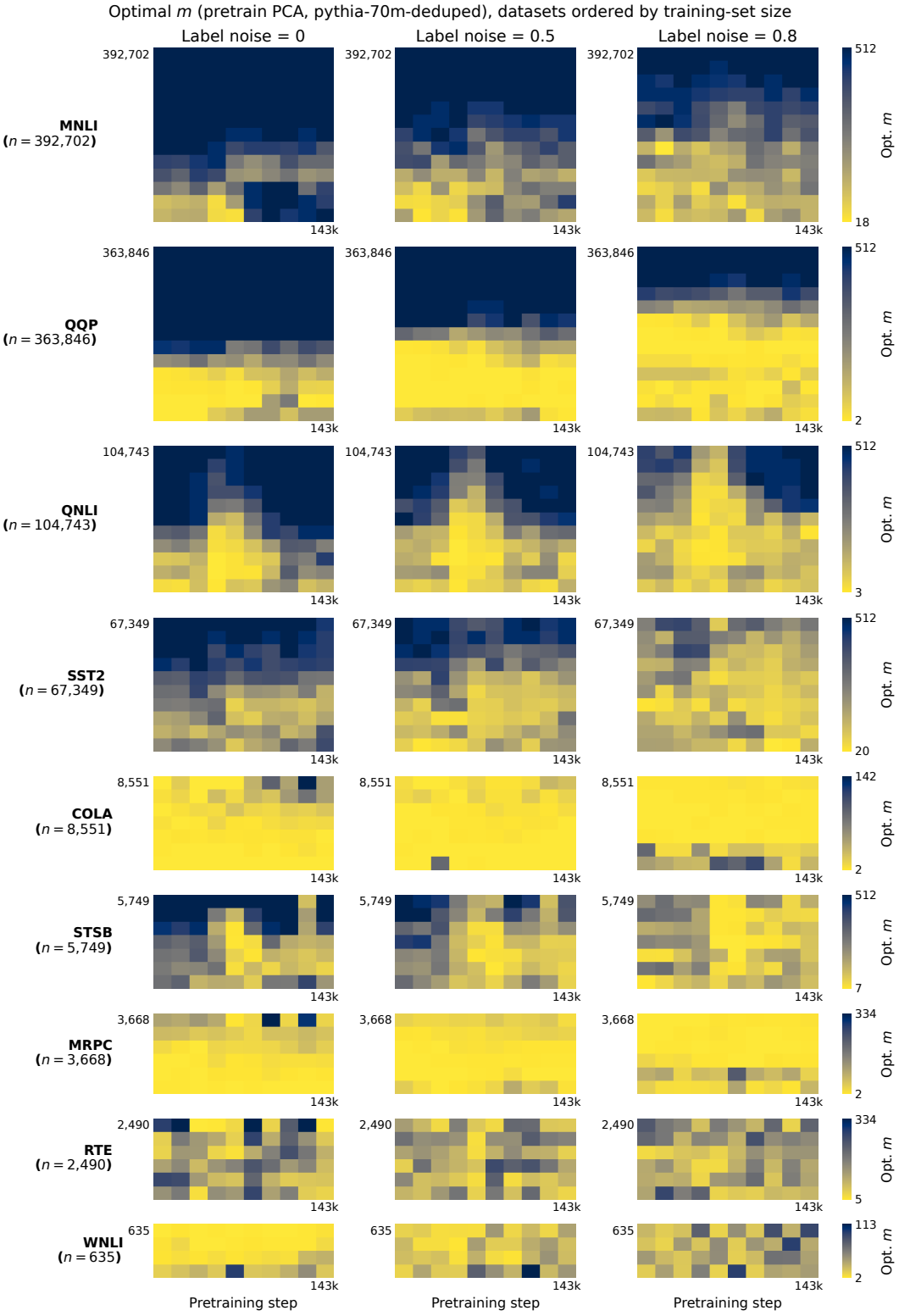


Figure 21: As Figure 20, but with PCA fit on pretraining data (the Pile) rather than the labelled downstream data. The validation-optimal m shows the same downward shift with increasing label noise (Table 3).

Dataset	Noise = 0	Noise = 0.2	Noise = 0.5	Noise = 0.8
MNLI	431.9	419.9	367.5	306.1
QQP	340.6	315.4	267.3	183.6
QNLI	334.3	315.6	271.8	191.6
SST-2	326.5	294.6	244.5	154.0
CoLA	13.7	12.2	6.5	13.6
STS-B	275.9	240.6	199.5	116.7
MRPC	34.9	27.6	15.9	22.6
RTE	127.6	116.8	107.9	113.2
WNLI	11.9	14.4	24.3	37.1

Table 3: Mean validation-optimal PCA dimension m for pretraining PCA, `pythia-70m-deduped`, averaged over checkpoints, downstream training-set sizes, and seeds, for each GLUE dataset and label-noise level. Datasets are ordered by training-set size, descending. Hidden-layer size is $p = 512$.

Dataset	Noise = 0	Noise = 0.2	Noise = 0.5	Noise = 0.8
MNLI	672.1	633.2	532.8	355.0
QQP	576.6	522.1	419.9	211.6
QNLI	562.4	515.4	427.7	260.8
SST-2	428.2	351.1	285.9	167.3
CoLA	65.3	33.6	12.1	16.3
STS-B	365.9	258.0	160.3	106.2
MRPC	47.1	31.3	14.5	21.6
RTE	161.8	160.0	130.7	140.4
WNLI	8.0	12.7	19.6	42.0

Table 4: Mean validation-optimal PCA dimension m for downstream PCA, `pythia-410m-deduped`, averaged over checkpoints, downstream training-set sizes, and seeds, for each GLUE dataset and label-noise level. Datasets are ordered by training-set size, descending. Hidden-layer size is $p = 1024$.

In classical PCR, $\mathcal{P}_m$ is constructed from the right singular vectors of the same matrix $\mathbf{X}_l$. Consequently,

$$(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l = \mathcal{P}_m, \tag{72}$$

and the noiseless component of the estimator reduces to $\mathcal{P}_m \mathbf{w}^\star$.

In the two-sample setting, $\mathcal{P}_m$ is instead estimated from the independent pretraining matrix $\mathbf{X}_u$. Hence $\mathcal{P}_m$ and $\mathbf{X}_l^\top \mathbf{X}_l$ do not generally share an eigenbasis, and

$$(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \neq \mathcal{P}_m. \tag{73}$$

The noiseless estimator therefore cannot be reduced to a direct projection of $\mathbf{w}^\star$ onto the retained empirical PCs. Its analysis requires additional quantities describing the interaction between the independently learned representation and the downstream design, including the spectra of $\mathbf{X}_l \mathbf{U}_m$ and $\mathbf{X}_l \mathbf{U}_\perp$ and deterministic equivalents for bilinear forms involving the downstream row-space projector Π. By contrast, the deterministic limits involving $\mathcal{P}_m$ alone are the same as, or closely related to, the corresponding quantities appearing in classical PCR.

The two-sample formulation also changes the relevant interpolation threshold. The downstream regression problem is controlled by

$$\gamma_{\text{eff}} = \frac{m}{n_l} = \alpha \gamma_l. \tag{74}$$

Because the representation is learned independently of the labelled dataset, m is not constrained by the rank of $\mathbf{X}_l$, and both $\gamma_{\text{eff}} < 1$ and $\gamma_{\text{eff}} > 1$ can arise as the representation dimension is varied. This allows the underparameterized and overparameterized linear-probe regimes to be studied within the same model.

Finally, the failure of (72) generates additional finite-sample contributions to the error. In particular,

$$(\mathbf{X}_l \mathcal{P}_m)^\dagger \mathbf{X}_l \mathbf{w}^\star$$

Dataset	Noise = 0	Noise = 0.2	Noise = 0.5	Noise = 0.8
MNLI	839.3	789.2	671.3	507.0
QQP	733.8	628.3	484.9	297.2
QNLI	665.0	584.8	505.4	315.0
SST-2	601.7	504.4	405.5	252.7
CoLA	60.9	24.1	10.2	21.8
STS-B	584.5	437.2	301.5	191.1
MRPC	61.5	33.4	22.7	55.0
RTE	261.9	241.1	203.8	186.6
WNLI	9.3	15.1	27.4	49.2

Table 5: Mean validation-optimal PCA dimension m for pretraining PCA, `pythia-410m-deduped`, averaged over checkpoints, downstream training-set sizes, and seeds, for each GLUE dataset and label-noise level. Datasets are ordered by training-set size, descending. Hidden-layer size is $p = 1024$.

contains contributions from signal components discarded by $\mathcal{P}_m$ but statistically coupled to the retained downstream features. These produce the additional missing-signal and leaked-signal bias terms, together with an associated variance contribution, appearing in our asymptotic expressions. Their magnitude depends jointly on the pretraining and downstream sample sizes, covariance structure, task alignment, and signal-to-noise ratio. In classical PCR, the same term reduces to $\mathcal{P}_m \mathbf{w}^\star$, so the corresponding bias is determined directly by the signal excluded from the retained empirical-PC subspace.

Thus, although both settings use PCA followed by linear regression, separating representation learning from downstream fitting changes both the estimator algebra and the resulting high-dimensional asymptotics.

Numerical comparison with regression and PCR We compare three approaches: standard regression, pretrained regression (PR) where PCA is learned on prior data and transferred to the downstream task, and principal component regression (PCR), where PCA is learned directly on the downstream data.

In the fully aligned setting ($\mathbf{w}^\star \parallel \mathbf{v}$, $\cos\theta = 1$), Figure 22 (left) shows that incorporating a PCA step consistently improves performance over standard regression. Moreover, when the pretraining dataset is larger than the downstream dataset ($n_u > n_l$), PR outperforms PCR, reflecting the more accurate estimation of the signal subspace from additional data.

In the misaligned setting ($\cos\theta = 1/\sqrt{2}$, right panel in Figure 22), the advantage of PR is restricted to regimes with large n_u and small n_l. As n_l increases, PCR becomes comparable to PR, and both methods approach the performance of standard regression (within a small error gap). This reflects the diminishing benefit of transferring a pretrained subspace when sufficient downstream data is available to estimate the relevant structure directly.

G.2 Fine-tuning

In practice, fine-tuning often involves attaching a re-initialized output head with small weights. This regime is closely related to the setting considered in our analysis. As demonstrated by [102–104], a small initialization can lead to dynamics in which the encoder weights change relatively little, while most of the adaptation occurs in the output weights. We therefore expect our theory to provide a useful approximation to full fine-tuning in regimes where feature learning remains limited. More generally, however, full fine-tuning may transform the projection matrix, turning the leakage term from a fixed inherited artefact into an adaptive component. Gradient descent can rotate the encoder and change the learned subspace, so the fixed decomposition into the projection matrix and its orthogonal complement need not remain valid throughout training. In the extreme case,the leakage may be reduced substantially or even eliminated. The leakage term depends on the missing signal , corresponding to task-relevant directions outside the initially inherited subspace. By updating the encoder, full fine-tuning may incorporate some of these directions into the learned representation, thereby reducing the contribution of the leakage term.

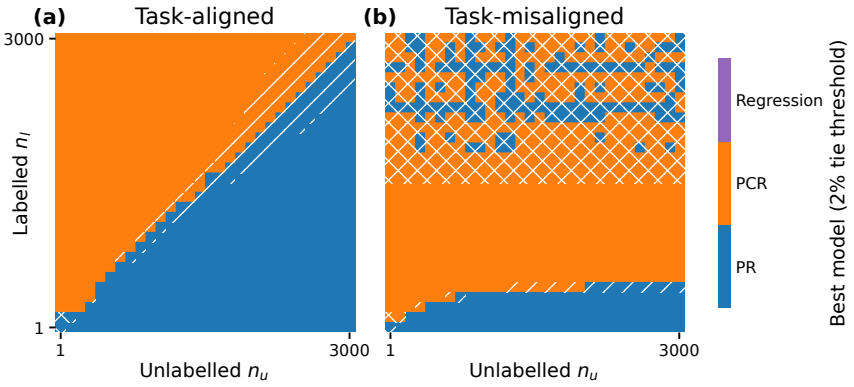


Figure 22: Heatmaps over the (n_u, n_l) grid ($p = 500$, SNR $= 1.8$, $\lambda = 5$) showing which method – standard regression (violet), pretrained regression (PR, blue) or principal component regression (PCR, orange) – achieves lower generalisation error. Forward hatching (//) marks regions where PR and PCR are tied (within 2% error); cross hatching (×) marks regions where all three methods (PR, PCR, and standard regression) are tied. When the task is aligned with the pretraining signal, PR wins for $n_u > n_l$ and PCR wins otherwise, with the boundary tracking the diagonal. Under misalignment, PCR dominates across most of the grid; PR retains an advantage when $n_u \gg n_l$.

H Implementation details

Code to reproduce all figures and experiments is available on GitHub.

H.1 Figure 1

Panel (a) Diagram of the model made with `figma`.

Model. We define the signal-to-noise ratio as $\text{SNR} = \frac{\|\mathbf{w}^\star\|^2}{\sigma^2}$ and the alignment between the regression target and the spike as $r_v^2 = \cos\theta^2$. Throughout, we fix $p = 500$ and $\sigma^2 = 1$ aligned spikes $\mathbf{v} = \mathbf{q}$, no regularizer.

Panels (b-c) (generalisation and training error vs α). We fix $n_u \approx n_l \approx 300$, SNR $= 9.0$, and $r_v^2 = 1.0$. The parameter α is swept over 30 values in $[0.01, 1.0]$ for each of four spike strengths $\lambda \in \{1.1, 2.0, 5.0, 10.0\}$. Theoretical predictions (solid curves) are compared against empirical averages over multiple realisations (dashed curves, shading corresponds to 1 standard deviation).

H.2 Figure 2

Model. We define the signal-to-noise ratio as $\text{SNR} = \frac{\|\mathbf{w}^\star\|^2}{\sigma^2}$ and the alignment between the regression target and the spike as $r_v^2 = \cos\theta^2$. Throughout, we fix $p = 500$ and $\sigma^2 = 1$ aligned spikes $\mathbf{v} = \mathbf{q}$, no regularizer.

Panel (a) (minimum generalisation error). For SNR $= 9.0$, $\lambda^\star = 5.0$, and $r_v^2 = 1.0$, we plot $\min_\alpha E_{\text{gen}}(\alpha)$ over a 100×100 grid with $n_u, n_l \in [1, 3000]$. Dashed white lines mark $n_u = p$ and $n_l = p$ ($p = 500$). The colour scale is logarithmic.

Panel (b) (gain from tuning α relative to $\alpha = 1$). Under the same setting, we compute $E_{\text{gen}}(\alpha = 1) - E_{\text{gen}}(\alpha^\star)$ over the same grid. Positive values indicate regions where using the full prior representation is suboptimal and tuning α reduces generalisation error. The colour scale is logarithmic.

Panel (c) (marginal substitution rate). Under the same setting, we compute the ratio $(\partial E_{\text{gen}}/\partial n_u)/(\partial E_{\text{gen}}/\partial n_l)$ evaluated at the optimal $\alpha^\star$, with gradients approximated numerically via central differences on the 100×100 grid. Values above 1 (red) indicate that collecting an

additional unlabelled sample n_u reduces generalisation error more than collecting a labelled sample n_l; values below 1 (blue) indicate the converse. The colour scale is centred at 1.

Panel (d) (generalisation-optimal $\alpha^\star$). Under the same setting, we plot the value of $\alpha^\star = \arg\min_\alpha E_{\text{gen}}(\alpha)$ over the same grid, normalised so that $\alpha \in [0, 1]$.

H.3 Figure 3

Model. We define the signal-to-noise ratio as $\text{SNR} = \frac{\|\mathbf{w}^\star\|^2}{\sigma^2}$ and the alignment between the regression target and the spike as $r_v^2 = \cos\theta^2$. Throughout, we fix $p = 500$ and $\sigma^2 = 1$.

Panel (a) (phase portrait SNR). At fixed $\lambda = 5.0$, $r_v^2 = 1.0$, aligned spikes $\mathbf{v} = \mathbf{q}$ and no regulariser, we plot the smooth and sharp transition boundaries in the (n_u, n_l) plane for each of the six SNR values $\{0.1, 0.5, 1.0, 3.0, 5.0, 9.0\}$. Increasing SNR progressively reduces the region in which full-rank representations ($\alpha \approx 1$) are suboptimal.

Panel (b) (phase portrait spike strength). At fixed $\text{SNR} = 9.0$, $r_v^2 = 1.0$, aligned spikes $\mathbf{v} = \mathbf{q}$ and no regulariser, we show the corresponding transition curves for $\lambda \in \{2, 3, 5, 7, 10\}$. Stronger spikes shift the phase boundaries.

Panel (c) (phase portrait alignment). At fixed $\text{SNR} = 9.0$,$\lambda = 5.0$, aligned spikes $\mathbf{v} = \mathbf{q}$ and no regulariser, we plot the transition curves for $r_v^2 \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. All panels are evaluated over $n_u, n_l \in [1, 3000]$.

Panel (d) (phase portrait alignment). At fixed $\text{SNR} = 5.0$, pretraining spike strength $\lambda_u = 10.0$, downstream spike strength $\lambda_\ell = 5.0$, and task-spike alignment $\rho_{wq} = 1.0$ (the task vector $\mathbf{w}^\star$ fully aligned with the downstream spike $\mathbf{q}$), we plot the smooth and sharp transition boundaries in the (n_u, n_l) plane for the two-spike setting, sweeping the spike-spike alignment $r = \mathbf{v}^\top \mathbf{q} \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ between the pretraining spike $\mathbf{v}$ and the downstream spike $\mathbf{q}$. Increasing r enlarges the region in which compression ($\alpha < 1$) is optimal.

H.4 Figure 4

Autoencoders

Data generation. Both covariances use $p = 200$ features, noise standard deviation $\sigma = 1$, $\text{SNR} = 9$, $\|\mathbf{w}^\star\| = 3$.

- *Spiked identity ($\lambda = 5.0$)*: $\mathbf{\Sigma} = \mathbf{I}_p + \lambda \mathbf{v}\mathbf{v}^\top$, i.e. the top eigenvalue is $1 + \lambda = 6.0$; true weight vector aligned with the spike $\mathbf{w}^\star = \|\mathbf{w}^\star\|\mathbf{v}$, where $\mathbf{v}$ is the top eigenvector of $\mathbf{\Sigma}$.
- *Power-law ($\beta = 0.3$)*: $\mathbf{\Sigma} = \text{diag}\big(1^{-\beta}, 2^{-\beta}, \ldots, p^{-\beta}\big)$; no additional spike is added, and $\mathbf{w}^\star$ is aligned with the top eigenvector (the first standard basis vector, since eigenvalues are already sorted in decreasing order).

Each grid point (n_u, n_l) averages five independent replications, with $n_u, n_l \in [1, 1200]$ each swept over 20 uniform values.

Models. For each (n_u, n_l) pair, $\alpha \in \{0.01, 0.05, \ldots, 1.0\}$ (20 values) sets the latent dimension $m = \lfloor \alpha p \rfloor$. The three models below are evaluated.

1. *PCA + Regression (PR)*: Eigenvectors of $\mathbf{S}_u = \mathbf{X}_u^\top \mathbf{X}_u / n_u$ are computed; $\mathbf{X}_l$ is projected onto the top-m eigenvectors and OLS is applied in the latent space.
2. *Linear AE*: Encoder ($\mathbb{R}^p \to \mathbb{R}^m$) and linear decoder ($\mathbb{R}^m \to \mathbb{R}^p$) trained by minimizing MSE on $\mathbf{X}_u$; the frozen encoder-decoder pair reconstructs $\mathbf{X}_l$ back into $\mathbb{R}^p$, and OLS is applied to this reconstructed p-dimensional representation (not the m-dimensional latent code).
3. *Nonlinear AE*: Same as the linear AE but with a sigmoid nonlinearity applied at the encoder, $\mathbf{z} = \sigma(\mathbf{W}_{\text{enc}}\, \mathbf{x})$.

All autoencoders are implemented in `JAX/Equinox` and trained with Adam ($\mathrm{lr} = 10^{-3}$, 20,000 epochs, batch size 500). Weights are initialized as $\mathbf{W}_{i,j} \sim \mathcal{N}(0, 1/\sqrt{d_{\mathrm{in}}})$.

Panel descriptions. Each panel is a 20×20 heatmap with n_u on the x-axis, n_l on the y-axis, and color encoding $\alpha^\star \in [0, 1]$ (the value minimizing mean loss over the five replications).

The autoencoder component of the figure is organized as 2 rows $\times$ 2 columns:

	Linear AE	Nonlinear AE
Spiked identity	(a)	(b)
Power law	(c)	(d)

Columns compare linear vs nonlinear AEs; rows contrast spiked identity and power-law covariance structures; each panel shows generalisation-optimal $\alpha^\star$. A single shared colorbar applies to all panels.

LLMs

Model and representations. All experiments in this Figure use `Pythia-70M-deduped` [79] (70M parameters, trained on the deduplicated Pile [80, 90]). We evaluate 10 checkpoints at steps 2, 8, 32, 128, 512, 2000, 5000, 20000, 80000, and 143,000. Representations are last-token hidden states from the final layer; downstream (SST-2) inputs are never truncated, while the Pile sequences used to fit the pretrain-PCA basis are truncated to 64 tokens, yielding vectors in $\mathbb{R}^{512}$.

Downstream task. Binary sentiment classification on SST-2 from the GLUE benchmark [3]. Fixed 872-sentence validation set; training sizes $n_l \in \{100, 250, 500, 1000, 2000, 4000, 10000, 20000, 40000, 67349\}$, n_l training examples are randomly chosen out of the maximum 67349 for each seed.

PCA. We sweep $m \in \{2, 3, 4, 7, 11, 16, 25, 39, 60, 92, 142, 218, 334, 512\}$ (approximate $\log_2$ grid), and consider the two conditions below.

1. *Downstream PCA* (Figure 17): basis fit on the n_l labelled training representations. The representation used is the hidden state at the last non-padding token.
2. *Pretrain-PCA* (Figure 4(e-g)): basis fit on up to 102,400 Pile sequences (100 batches $\times$ 1,024 sequences) from the same checkpoint, using `torch.linalg.eigh`. For earlier checkpoints, the basis is fit on $1024 \times n_{\mathrm{checkpoint}}$ sequences. Each Pile sequence is first truncated to its first 64 tokens to match the lengths of sequences in SST2, and the representation used is the final-layer hidden state at the last token of that truncated window.

Linear probe. Projected representations are fed to a 2-output linear layer trained with cross-entropy loss using Adam ($\mathrm{lr} = 10^{-3}$, weight decay 10^{-4}), for up to 10,000 epochs with early stopping (200 steps without improving the best loss by more than 10^{-4}). Linear probes were randomly initialized with small weights for each seed.

Seeds and aggregation. Each (checkpoint, n_l, m) configuration is run with 6 random seeds (affecting probe initialization and downstream data shuffling). Figures show mean $\pm$ std over seeds; optimal-m heatmaps report, for each (checkpoint, n_l), the mean over seeds of the per-seed argmax-m (the m maximizing validation accuracy is first selected independently for each seed, and these per-seed optimal values of m are then averaged).

Panel descriptions.

- **(e)** Heatmap for optimal m under pretrain-PCA, selected by validation accuracy.
- **(f)** Eigenvalue histogram at step 143,000, computed from the 872-sentence SST-2 validation covariance; log-density on the y-axis.
- **(g)** Validation accuracy vs m (x-axis in $\log_2$ scale) under pretrain-PCA for $n_l = 2000$, with curves colored by pretraining progress.

H.5 Compute

Theory curves. Theory curves were computed on a single CPU (Intel Xeon E5-2680 v4 @ 2.40 GHz, no GPU required). Each configuration sweeps a 100×100 grid of (n_u, n_l) values over 30 values of α, amounting to approximately 300,000 evaluations of closed-form expressions. The full sweep covers 24 configurations, $\lambda^* \in \{1.1, 2.0, 5.0, 10.0\} \quad \times \quad \mathrm{SNR} \in \{0.1, 0.5, 1.0, 3.0, 5.0, 9.0\}$, with alignment $r_v^2 = 1.0$, totaling approximately 7.2 million evaluations. When executed in parallel, the full sweep completes in approximately 2h. Peak RAM usage was approximately 2 GB per configuration.

Simulations. Simulations were distributed over a SLURM cluster. Each job was allocated 4 CPUs and 30 GB of RAM. The full grid 6 SNR $\times$ 4 $\lambda_1 \times 4$ $\lambda_2 \times 30$ n_b produced 2,880 SLURM jobs, each processing 30 values of n_l with 100 independent replications at $p = 500$. Each replication involves an eigendecomposition and pseudoinverse of a $p \times p$ matrix for each of the 30 values of α. When executed in parallel, the full sweep completes in approximately 10h.

Autoencoder experiment. Simulations were distributed over a SLURM cluster. Each job was allocated 1 GPU, 8 CPUs, and 16 GB of RAM. The full grid 20 $n_{\mathrm{base}} \times 20$ n_{task} produced 400 SLURM jobs, each processing 20 values of α with 5 independent replications at $p = 200$. Each replication trains a linear and a nonlinear autoencoder for 20,000 epochs for each value of α, followed by OLS regression in the learned latent space. The individual jobs completed in 10-50min depending on parameters.

LLM probe experiment. Simulations were distributed over a SLURM cluster. Each job was allocated 1 GPU, 16 CPUs, and 32 GB of RAM. Two sets of 10 SLURM jobs were submitted: one per checkpoint of Pythia-70M-deduped for the downstream-PCA condition and one for the pretrain-PCA condition. Each job sweeps over 10 training-set sizes and 6 seeds, fitting a linear probe for up to 10,000 epochs with early stopping across up to 14 PCA dimensionalities m.

H.6 LLM Usage

Large language models (LLMs) were used throughout this research project to assist with writing, editing, and formatting the manuscript, including improving clarity and refining LaTeX expressions. LLMs were also used as an aid for understanding and discussing technical concepts, as well as for debugging and drafting portions of the experimental and simulation code. All scientific results, analyses, and conclusions were verified and validated by the authors.

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: See section 1.

   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: The paragraph is copied here for reference. "Several limitations motivate future work: relaxing the shared-spike assumption to capture data mismatch, extending the analysis to ridge-regularised learning, and considering richer covariance models. Another important direction is to replace the closed-form regression estimator with gradient-based training, enabling the study of optimisation-induced bias and its dependence on initialisation from pretraining. Finally, extending the framework beyond PCA-based projections to learned feature representations would allow for a closer connection to empirical observations in deep networks."

   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: The assumptions are stated in the problem statement (Section 4), and the proofs can be found in Appendix A.

   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: Please refer to the simulation details in Appendix H.

   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: As written in Appendix H, code to reproduce all figures and experiments is available on GitHub. .

   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: Please refer to the simulation details in Appendix H and the caption of the figures.

   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: Please refer to the simulation details in Appendix H and the caption of the figures.

   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: Please refer to the simulation details in Appendix H.

   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: We follow the NeurIPS Code of Ethics.

   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: This work is theoretical and not directly tied to deployed systems or immediate societal applications. Potential positive impacts include improving the efficiency and interpretability of machine learning systems through a better understanding of pretrained representations and downstream generalization.

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: [N/A]

    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: [Yes]

    Justification: For justification please see the reference section.

    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: [N/A]

    Justification: [N/A]

    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: [N/A].

    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: [N/A]

    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: [Yes]

    Justification: Please refer to the LLM usage section in Appendix H.6.

    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.