THE HIDDEN ADVANTAGE OF MASK RESAMPLING: A THEORY OF MASKED AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Why can masked prediction learn useful representations that unmasked reconstruction misses? We study this question in a high-dimensional model of a masked autoencoder (MAE) trained on data with shared latent structure and heterogeneous noise. We prove that masked linear reconstruction can recover the latent feature at linear sample complexity in regimes where unmasked linear reconstruction, equivalent to PCA, fails. The analysis also quantifies the statistical advantage of mask resampling, an established ingredient of masked pretraining. By introducing a fixed collection of K masks per sample, we characterize its effect on feature recovery and downstream performance, identifying regimes where greater mask diversity lowers sample complexity. Guided by this prediction, we find that random cropping and flipping in standard image-training pipelines can obscure the advantage of mask resampling by renewing the prediction task even when the patch mask is fixed. Removing these transformations reveals a downstream advantage for dynamic over static masking in CNN autoencoders and vision transformers. A complementary BERT pilot finds benefits from greater mask diversity on downstream language tasks. Our results separate the benefit of the masked prediction objective from that of mask diversity, and show how a tractable theory can guide experiments that uncover advantages hidden by standard training practices.

Why can predicting missing parts of an input produce more useful representations than reconstructing the entire input? Building on denoising autoencoders (Vincent et al., 2010), Pathak et al. (2016) predicted missing image regions from visible context and obtained more transferable representations than a standard autoencoder trained without input masking, a paradigm that now underlies masked language and image modeling (Devlin et al., 2019; He et al., 2022; Bao et al., 2022; Xie et al., 2022). Masked learning therefore couples two ingredients: a *prediction objective* that specifies what structure is rewarded, and an *augmentation procedure* that controls how many distinct prediction tasks a given sample ever presents. Despite its empirical success, our theoretical understanding of how these ingredients shape learned representations remains incomplete. We develop a high-dimensional theory that separates their effects and connects them to latent-feature recovery and downstream performance.

Motivated by the observation that unmasked, i.e. the standard autoencoder, reconstruction can favor high-variance nuisance directions over downstream-relevant structure (Balestriero & LeCun, 2024), we consider a data model with a shared latent direction and heterogeneous coordinate-wise noise. The latent direction carries the downstream signal, while heterogeneity allows nuisance variance to dominate the leading principal component. Masking changes this competition by requiring hidden coordinates to be predicted from the visible ones, rather than rewarded for their own variance. We prove that a linear masked autoencoder recovers the latent direction with $n = O(d)$ samples where unmasked linear reconstruction, equivalent to PCA, fails (Sec. 2).

Analyzing this model clearly reveals a second notable effect: the number of distinct masks available per sample controls how many underlying training samples are needed for latent-feature recovery. We distinguish static masking, where each sample is assigned a single random mask that is reused throughout training, from dynamic masking, commonly used in practice (He et al., 2022; Liu et al., 2019), in which the mask is independently resampled each time a sample is used during pretraining. More generally, assigning each sample K independent masks and reusing them thereafter interpolates between these two regimes at fixed number of underlying samples. This allows us to isolate mask diversity from data diversity. Our analysis predicts that increasing K lowers the recovery threshold, with corresponding improvements in downstream performance in the settings studied.

This prediction prompted us to revisit how the contribution of mask diversity is assessed empirically. In language modeling, dynamic masking has been reported to provide only modest gains over reusing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

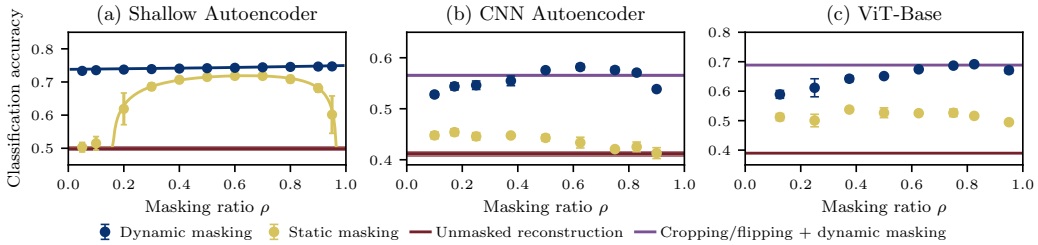


Figure 1: **Isolating mask resampling reveals its contribution to representation learning.** Downstream accuracy versus masking ratio ρ for dynamic masking (blue), static masking (yellow), and unmasked reconstruction (red). **(a)** Shallow autoencoder on spiked model (Sec. 1); curves are theory and markers are simulations. **(b)** CNN autoencoder on CIFAR-10 (test top-1). **(c)** ViT-Base on ImageNet-100 (validation top-5). In (b,c), blue and yellow curves use no stochastic cropping or flipping. The purple reference includes cropping and flipping with *dynamic mask resampling* at $\rho = 0.625$ **(b)** and $\rho = 0.75$ **(c)**. At suitable masking ratios, resampling alone reaches comparable accuracy. Bars and shading show ± 1 standard deviation (SD). See App. H for details.

a fixed collection of masks (Liu et al., 2019), but their static baseline already uses $K = 10$ pre-generated masks per sequence rather than a single fixed mask. Data2Vec 2.0 shows substantial improvements in image classification accuracy by increasing the number of masked views per sample within each training update (Baevski et al., 2023). Furthermore, in vision, self-supervised image learning commonly uses random cropping, flipping, and color transformations to generate changing views of the same sample (Chen et al., 2020b; Tian et al., 2020). In MAE specifically, standard pretraining combines random cropping and flipping with mask resampling (He et al., 2022). Under a newly drawn crop, for example, the same patch mask can hide different content from the original image. We therefore hypothesize that stochastic image augmentation adds another source of prediction-task diversity, obscuring the effect of resampling the mask itself.

Guided by this hypothesis, we compare static and dynamic masking with and without stochastic cropping and flipping. Without these transformations, dynamic masking improves downstream accuracy in CNN autoencoders on CIFAR-10 and ViT-Base on ImageNet-100 (Fig. 1 (b,c)). In the ViT experiments, reintroducing cropping and flipping substantially narrows the observed static–dynamic gap, mainly because static masking improves (Fig. 5). Finite-K experiments further show progressive improvements as each image is exposed to more masks (Fig. 4). A complementary BERT-Medium pilot finds improvements with increasing mask diversity on GLUE tasks, extending the qualitative evidence beyond vision (Sec. 4).

Together, these results distinguish two benefits of masked representation learning: the prediction objective can favor useful correlations over nuisance variance, while renewing the masks can lower the sample complexity required to learn those correlations. By identifying this distinction, the theory motivates a controlled test of its relevance beyond the solvable model.

The **main contributions** of this paper are the following:

- Under suitable heterogeneous noise, we prove that a linear masked autoencoder recovers the latent direction for all sufficiently large finite sample complexity $\alpha = n/d$, whereas unmasked linear reconstruction (PCA) fails for every fixed $\alpha > 0$. Replica calculations complement this guarantee by predicting the recovery threshold (Sec. 2).
- We formulate a tractable high-dimensional model of a two-stage learning protocol: self-supervised learning (SSL) through masked reconstruction, followed by supervised linear probing with a frozen encoder (Sec. 1). For linear and nonlinear autoencoders, we derive replica predictions for feature recovery and downstream accuracy as functions of the masking ratio and mask diversity, showing the benefit of resampling the masks (Sec. 3).
- Guided by our theoretical predictions, we test the role of mask diversity in vision and language. Dynamic masking improves downstream performance without stochastic cropping and flipping; adding these transformations substantially narrows the observed static–dynamic gap in the ViT setting, consistent with overlapping sources of prediction-task diversity. A BERT-Medium pilot provides complementary evidence for the benefit of mask diversity in language modeling (Sec. 4).

FURTHER RELATED WORK

We outline the broad lines of related work here; App. A provides an expanded comparison with the closest references and clarifies the differences in setting, analytical scope, and contribution.

Statistical-physics theory of neural networks. Our replica analysis builds on high-dimensional empirical risk minimization (Aubin et al., 2020; Loureiro et al., 2021; Cui, 2025), shallow-autoencoder dynamics (Refinetti & Goldt, 2022), and denoising autoencoders (Cui & Zdeborová, 2023). Conde Mendes et al. (2026) prove an advantage of nonlinear autoencoders over PCA. Heterogeneous noise has also been studied in supervised learning (Adomaityte et al., 2023; 2024).

Theory of masked learning. Existing work connects masked autoencoding to contrastive learning and downstream guarantees (Zhang et al., 2022b), and establishes latent identifiability in hierarchical models (Kong et al., 2023). Bisulco et al. (2025) characterize linear MAEs with full reconstruction from masked inputs, including visible coordinates rather than only hidden targets. Their analysis relates patch size and masking ratio to the spatial correlations learned. Dong et al. (2025) analyze decoupled coordinate predictors through rigorous ridgeless-regression risk formulas, accounting for both corrupted covariates and the samples supplying a selected prediction target. Single-coordinate prediction also connects to factored attention (Rende et al., 2024). Wortsman Zurich et al. (2026) characterize a spiked-model recovery transition in the joint matrix of single-coordinate ridge predictors and establish prediction-risk advantages over PCA in structured covariance settings. Cabannes et al. (2023) distinguish sample variability from augmentation variability in SSL generalization bounds; additional views can reduce the latter without removing the former.

Spiked models and heteroskedastic noise. Our linear baseline connects to spiked covariance models (Johnstone, 2001) and PCA recovery transitions (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012). Hong et al. (2018) characterize PCA under heteroskedastic noise. Spectral methods addressing heteroskedasticity include weighted PCA (Hong et al., 2023), diagonal correction through HeteroPCA (Zhang et al., 2022a), and optimal estimators for doubly heteroskedastic noise (Zhang & Mondelli, 2024). We study how masking changes the reconstruction objective to promote recovery, without supplying the coordinate variances or their distribution.

1 A HIGH-DIMENSIONAL MODEL OF MASKED LEARNING

Data and noise heterogeneity. Our goal is to compare prediction objectives and masking protocols for a data distribution that enables sharp high-dimensional analysis. Inspired by insights into the advantage of masking from Balestriero & LeCun (2024) and by work on unsupervised learning in high dimensions (Johnstone, 2001; Deshpande & Montanari, 2016; Lesieur et al., 2017), we consider a data model that separates a shared informative feature in the form of a spike from coordinate-specific nuisance fluctuations. Concretely, we consider n samples with d coordinates generated as

$$x_{\mu i} = \sqrt{\frac{\beta}{d}}\,\lambda_\mu u_i^\star + \sqrt{\gamma_i^\star}\,z_{\mu i}, \qquad \mu \in [n], \quad i \in [d]. \tag{1}$$

Here $\boldsymbol{u}^\star$ is the shared latent direction (spike), λ_μ is the latent variable of sample μ, and $\beta > 0$ controls signal strength. The variance $\gamma_i^\star$ sets the strength of nuisance fluctuations in coordinate i: a large value increases input variance independently of the spiked direction. These variances are drawn once and shared by all samples, whereas the noise realizations $z_{\mu i}$ change across samples and coordinates.

The variables λ_μ and $u_i^\star$ are independently drawn from centered, unit-second-moment distributions P_λ and P_u; $z_{\mu i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and the variances $\gamma_i^\star \overset{\text{iid}}{\sim} P_\gamma$ are positive. We normalize $\mathbb{E}_\gamma[\gamma] = 1$ and write $\kappa_2 = \mathbb{E}_\gamma[\gamma^2] < \infty$; hence $\kappa_2 \geq 1$, with equality only for homogeneous noise. All random families are mutually independent. The case $P_\gamma = \delta_1$ is the standard spiked covariance model (Johnstone, 2001). Full assumptions and notation are in Apps. B.1 and B.2.

Architecture and masked prediction. The learner is a single-hidden-unit autoencoder with tied encoder and decoder weights $\boldsymbol{w} \in \mathbb{R}^d$. For each sample μ, we draw K mask vectors $\boldsymbol{\chi}_\mu^{(r)} = (\chi_{\mu i}^{(r)})_{i\in[d]} \in \{0,1\}^d$, with entries drawn independently from $\mathrm{Bernoulli}(\rho)$ and independently of the data. The mask collection $\Xi = (\boldsymbol{\chi}_\mu^{(r)})_{\mu\in[n],\, r\in[K]}$ is held fixed throughout pretraining. An entry of

one marks a hidden coordinate; so ρ is the *expected masking ratio*. Mask r of sample μ encodes only visible coordinates through

$$h_{\mu r,\mathrm{vis}}(\boldsymbol{w}) = \frac{1}{\sqrt{d}}\,\boldsymbol{w}^\top\left[\left(\mathbf{1}_d - \boldsymbol{\chi}_\mu^{(r)}\right)\odot \mathbf{x}_\mu\right], \tag{2}$$

where $\odot$ denotes the elementwise (Hadamard) product and $\mathbf{1}_d$ is the all-ones vector. For the fixed mask collection Ξ, pretraining minimizes the regularized empirical masked risk

$$\mathcal{R}_{K,d}(\boldsymbol{w}) = \frac{\lambda_R}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{K}\sum_{\mu=1}^{n}\sum_{r=1}^{K}\left\|\boldsymbol{\chi}_\mu^{(r)}\odot\left[\mathbf{x}_\mu - \frac{\boldsymbol{w}}{\sqrt{d}}\,\sigma(h_{\mu r,\mathrm{vis}}(\boldsymbol{w}))\right]\right\|_2^2, \tag{3}$$

where σ is the activation function and λ_R controls ridge regularization. The loss is evaluated only on hidden coordinates, so each target is predicted from the visible input. Unmasked reconstruction instead observes and predicts every coordinate; we define this separate objective in Sec. 2.

Static, finite-K, and dynamic masking. The choice $K = 1$ gives static masking. Increasing K adds masked views without adding underlying samples. At a fixed dataset, the average over masks converges to the mask expectation as $K \to \infty$. We refer to this loss as the dynamic masking objective; drawing a fresh mask at each encounter provides stochastic evaluations of it. The high-dimensional analysis takes $n, d \to \infty$ at fixed K and obtains dynamic masking as $K \to \infty$ (App. B.2.4).

Our *two-stage protocol* first learns a representation from unlabeled data and then fits a supervised probe with the encoder frozen. We assess both whether pretraining recovers the shared latent direction and whether the resulting representation supports prediction from a small, independent labeled dataset.

Pretraining and feature recovery. Given the pretraining dataset $\mathcal{D}_{\mathrm{pre}} = \{\boldsymbol{x}_\mu\}_{\mu=1}^n$ with $n = \alpha d$ and the fixed mask collection Ξ, we study a global minimizer $\widehat{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^d}\mathcal{R}_{K,d}(\boldsymbol{w})$ of the risk in (3). In the proportional limit $n, d \to \infty$ at fixed sample complexity $\alpha = n/d$, recovery is measured by the cosine similarity (alignment):

$$\theta_{\boldsymbol{u}^\star} = \frac{|\langle\widehat{\boldsymbol{w}}, \boldsymbol{u}^\star\rangle|}{\|\widehat{\boldsymbol{w}}\|_2\,\|\boldsymbol{u}^\star\|_2} \tag{4}$$

Weak recovery means that there exists $\eta > 0$, independent of n and d such that $\theta_{\boldsymbol{u}^\star} > \eta$ has probability that tends to one as $n, d \to \infty$. A recovery threshold marks the onset of this alignment.

Downstream task and linear probing. After pretraining, we freeze $\widehat{\boldsymbol{w}}$ and draw a labeled dataset $\mathcal{D}_{\mathrm{ds}} = \{(\boldsymbol{x}_\nu^{\mathrm{ds}}, y_\nu^{\mathrm{ds}})\}_{\nu=1}^{n_{\mathrm{ds}}}$, where $y_\nu^{\mathrm{ds}} = \mathrm{sign}(\lambda_\nu^{\mathrm{ds}})$. These samples follow (1) with the same latent direction and coordinate variances as $\mathcal{D}_{\mathrm{pre}}$, but independently drawn latent coefficients and noise. We fit a one-dimensional logistic classifier on the unmasked scalar features $d^{-1/2}\widehat{\boldsymbol{w}}^\top\boldsymbol{x}_\nu^{\mathrm{ds}}$ and evaluate it on independent test samples from the same model. Thus pretraining uses $n = \alpha d$ unlabeled samples, while the labeled probe size n_{ds} remains fixed as d grows (Apps. G.2 and H.7).

Analytical predictions and optimality. The replica method predicts feature recovery after pretraining and the accuracy of the subsequent frozen-representation probe (Apps. C.1 and C.2). Replica calculations predict weak recovery thresholds by locating where an uninformative solution becomes unstable to a small alignment. These predictions concern continuous recovery transitions and are validated with simulations; an earlier discontinuous transition requires a separate analysis (App. D.2.1). Below, "weak recovery threshold" has this predictive meaning when referring to replica results. Independently, Sec. 2 proves a separation between masked linear autoencoder and PCA.

We also compare feature recovery with a matched Bayes-optimal reference that uses the same unlabeled samples but is supplied with the latent priors and realized coordinate variances. Unmasked reconstruction and finite-K comparisons isolate the effects of the objective and mask reuse, respectively; the Bayes reference instead assesses the gap to matched inference with additional prior information. Following Lesieur et al. (2017), App. F gives the state evolution and heteroskedastic approximate message passing (AMP) updates. Downstream evaluation uses the same scalar logistic probe for AMP and the autoencoders (App. G.2).

2 A PROVABLE ADVANTAGE OF MASKED RECONSTRUCTION

We first isolate the effect of the prediction objective. Unmasked autoencoders learn a compressed representation by reconstructing the full input through a low-dimensional bottleneck. The bottleneck forces the learner to select which variations in the data to preserve, with reconstruction error providing the criterion for this selection. Unmasked reconstruction is therefore a natural baseline for asking whether the features most useful for reconstructing the input also capture its shared latent structure.

Unmasked reconstruction. The autoencoder observes and predicts every coordinate, minimizing

$$\mathcal{R}_{\mathrm{full,d}}(\boldsymbol{w}) = \sum_{\mu=1}^{n} \left\| \mathbf{x}_\mu - \frac{\boldsymbol{w}}{\sqrt{d}}\, \sigma\big(h_\mu^{\mathrm{full}}(\boldsymbol{w})\big) \right\|_2^2, \qquad h_\mu^{\mathrm{full}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{x}_\mu}{\sqrt{d}}. \tag{5}$$

For $\sigma(h) = h$, the global minimizers of this tied single-unit autoencoder point along a direction of largest empirical variance, making it equivalent to ordinary PCA (Bourlard & Kamp, 1988); see App. E.1.1 for the reduction.

What changes when the target is hidden? For the masked objective (3), we likewise consider a linear activation, set $\lambda_R = 0$, and average over masks. Write $\boldsymbol{w} = \sqrt{dQ}\,\boldsymbol{r}$, with $Q \geq 0$ and $\|\boldsymbol{r}\| = 1$, and let $\Sigma = \mathrm{diag}(\gamma_1^\star, \ldots, \gamma_d^\star) + \beta \boldsymbol{u}^\star (\boldsymbol{u}^\star)^\top / d$ be the population covariance conditional on the planted direction and coordinate variances. Both population losses per sample are quadratic in Q. Their terms linear in Q are $-2Q$ times the following rewards:

$$\text{Unmasked:} \quad \boldsymbol{r}^\top \Sigma \boldsymbol{r} \qquad \text{Masked:} \qquad \rho(1-\rho)\, \boldsymbol{r}^\top \big(\Sigma - \mathrm{diag}\,\Sigma\big) \boldsymbol{r} \tag{6}$$

Unmasked reconstruction rewards total variance, whereas masked prediction rewards cross-correlations between different coordinates. The diagonal is absent from the masked reward because masking prevents a coordinate from predicting itself, removing the reward for its own variance: only correlations with visible coordinates can help reconstruct it.

To illustrate this effect, consider a coordinate with large noise variance, preserving its fluctuations in the compressed representation can substantially reduce unmasked reconstruction error, even though those fluctuations carry no information about the shared structure (the spike $\boldsymbol{u}^\star$ in our model). But when that coordinate is hidden, its independent noise cannot be predicted from the visible coordinates. The shared feature $\boldsymbol{u}^\star$, by contrast, creates correlations between visible and hidden coordinates and can therefore improve masked prediction. Noise no longer earns a direct reward through its variance, and the quadratic prediction penalty discourages reliance on noisy inputs. Together, these effects can allow masked reconstruction to favor cleaner coordinates. The full loss decomposition is given in Apps. E.1.4 and E.1.5.

Homogeneous noise provides a reference case. When $P_\gamma = \delta_1$, noise contributes equally to the unmasked reward in every unit direction. For Gaussian latent priors, PCA recovers above $\alpha_c = \beta^{-2}$ (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012), while the replica-symmetric prediction for linear dynamic masking gives the same threshold and asymptotic alignment (Apps. D.3.3 and E.1.1). Fig. 2 (a) illustrates this agreement.

A separation at linear sample complexity. We specialize the model of Sec. 1 to $P_u = P_\lambda = \mathcal{N}(0,1)$ and a fixed, mean-one variance law supported in $[a, \infty)$ for some $a > 0$, with finite second moment. We fix $\beta > 0$ and $\rho \in (0,1)$ as $n, d \to \infty$ with $n/d \to \alpha$. The precise hypotheses, including additional edge conditions for bounded-noise PCA failure, are collected in Assumption E.1.

Theorem 2.1 (Recovery by masked reconstruction and separation from PCA)**.** *Under (A1) and (A2), there exist finite $A_{\mathrm{suff}}(a,\beta,\rho)$ and $\eta(a,\beta,\rho) > 0$ such that, for every fixed $\alpha > A_{\mathrm{suff}}$, with probability tending to one the linear dynamic masking objective, with exact averaging over masks, admits global minimizers and every such minimizer satisfies $\theta^2_{\boldsymbol{u}^\star} \geq \eta$. If, additionally, the variance support is unbounded or the bounded-noise conditions (A3) hold, then unmasked linear reconstruction (PCA) has $\theta^2_{\boldsymbol{u}^\star} \to 0$ in probability for every fixed $\alpha > 0$.*

For bounded noise with upper endpoint γ_+, (A3) requires $\beta\, \mathbb{E}_\gamma[(\gamma_+ - \gamma)^{-1}] < 1$ and $\mathbb{E}_\gamma[\gamma^2(\gamma_+ - \gamma)^{-2}] = \infty$. The first condition excludes a separated population covariance outlier; the second ensures a regular upper edge of the noise-only sample-covariance spectrum at every fixed finite α. P_γ from Fig. 2(b) satisfies these conditions.

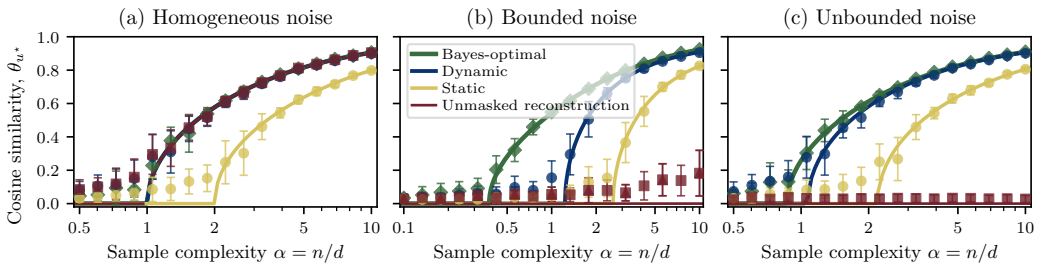


Figure 2: **Masking recovers a feature that PCA misses.** Cosine similarity versus sample complexity $\alpha = n/d$, for **(a)** homogeneous noise, $\gamma = 1$; **(b)** bounded heterogeneous noise, $\gamma = \frac{1}{4} + \frac{9}{4}B$, $B \sim \mathrm{Beta}(1, 2)$; and **(c)** unbounded heterogeneous noise, $\gamma = \frac{3}{4} + \frac{1}{4}e^{Z-1/2}$, $Z \sim \mathcal{N}(0, 1)$. Curves are theoretical predictions (linear masked autoencoders), markers are finite-dimensional simulations, and error bars show ± 1 SD across seeds. Green curves give the matched Bayes-optimal reference; green diamonds show AMP results (Apps. F.5 and H.14). Unmasked reconstruction is ordinary PCA. The homogeneous panel provides a control; the heterogeneous panels illustrate recovery by masking where the asymptotic PCA alignment vanishes. See App. H for details.

Theorem 2.1 therefore shows that the same dataset contains a recoverable feature that its leading principal component misses: changing the prediction objective is enough to obtain positive alignment with $n = O(d)$ samples, without supplying coordinate variances to either learner. The proof first shows that the population masked objective favors an informative direction over all directions with sufficiently small overlap. It then controls empirical fluctuations uniformly over candidate directions, relative to the noise penalty, so that this preference survives at sufficiently large finite α. The argument applies to every global minimizer of the objective (Apps. E.1 and E.2).

Interpreting the separation. Fig. 2 (b,c) illustrates how heterogeneity changes the comparison with the homogeneous reference: masking recovers while PCA can remain uninformative[1]. The bounded example (panel b) shows that this effect does not require arbitrarily large coordinate variances; the signal can remain below the noise spectral edge even when the variance support is compact. With unbounded support (panel c), extreme coordinate variances instead drive the leading sample component away from the shared feature. The respective PCA arguments are given in Apps. E.1.2 and E.1.3. The masked reward in (6) involves the same off-diagonal covariance structure used by diagonal-deletion PCA (Florescu & Perkins, 2016), but the full masked objective also contains a noise-sensitive, direction-dependent prediction penalty. In our simulations, dynamic masking achieves better alignment than both diagonal-deletion PCA and HeteroPCA (App. H.4).

The theorem guarantees recovery for all sufficiently large finite α for dynamic masking, but does not locate the sharp transition, for which the replica analysis provides a prediction. Having established why the masked objective can favor useful structure, we next ask how its performance changes when each sample supplies only a finite collection of masks.

3 Mask Diversity and Sample Complexity

We now keep the original dataset fixed and vary the number K of masks per sample. Reusing a fixed collection of masks repeatedly presents the same visible-to-hidden prediction tasks; increasing K adds new tasks without adding independent observations. The question is how this change affects the weak recovery thresholds and downstream performance.

Predicting learning across mask diversity. Our replica analysis yields high-dimensional learning curves for static, finite-K, and dynamic masking, resolving their dependence on sample complexity α, masking ratio ρ, and activation function (App. C.1). It predicts alignment with the planted feature and the effective noise along the learned direction; the downstream analysis then translates these quantities into classification accuracy under the linear-probe protocol introduced in Sec. 1.

[1] App. H.2 shows that the positive cosine similarity of PCA in Fig. 2(b) is due to finite-size effects.

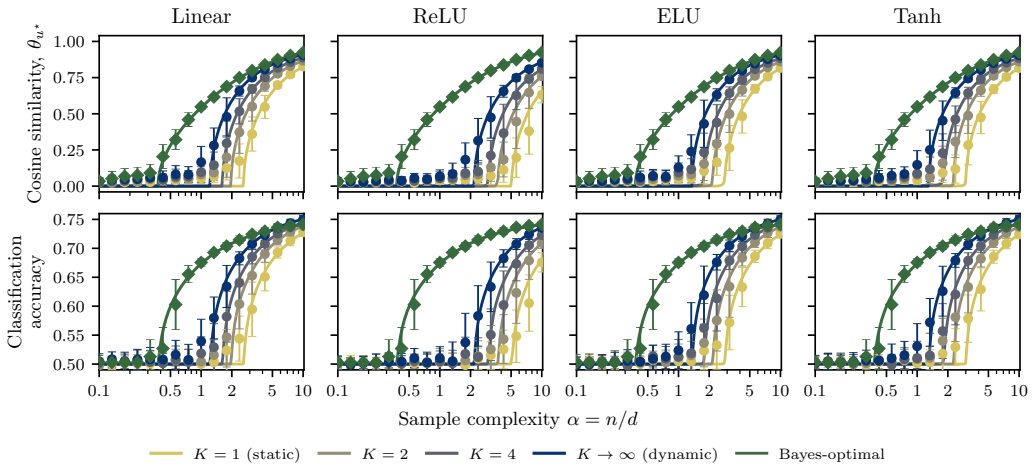


Figure 3: **Replica predictions track learning across mask diversity.** Cosine similarity (top) and downstream classification accuracy (bottom) versus sample complexity $\alpha = n/d$ under bounded heterogeneous noise (variance distribution of Fig. 2(b)), for linear, ReLU, ELU, and tanh activations. Colors distinguish $K = 1$(static), $2, 4$ and the dynamic masking limit. Solid curves are replica predictions and markers are finite-dimensional simulations. Green curves show the Bayes-optimal pretraining reference, evaluated with the same logistic probe for the lower row; green diamonds show AMP results (Apps. F.5 and H.14). Error bars show ± 1 SD. See App. H for experimental details.

Fig. 3 compares these predictions with numerical training at finite dimension. Across the tested activations, increasing mask diversity K lowers the weak recovery threshold and improves alignment above the transition in the settings shown. The replica predictions capture not only these threshold shifts, but also the magnitude and α-dependence of the cosine similarity over the displayed range. The theory also captures training and test errors (Fig. 10). Thus the analysis quantifies how mask resampling changes the representation learned from a given number of examples.

From feature recovery to downstream performance. The lower row of Fig. 3 shows how these representation gains translate into classification accuracy. The linear-probe predictions (App. G.2) track the numerical results as sample complexity and mask diversity vary across the four activation functions. Together, the two rows show that the theory captures both latent-feature recovery during pretraining and the downstream usefulness of the frozen representation.

The Bayes-optimal pretraining curve provides the reference using the matched priors and the realized coordinate variances (App. F). Increasing mask diversity brings the representations closer to this reference in the displayed comparisons, while a gap can remain even with dynamic masking (Fig. 3).

How mask diversity lowers sample complexity. The replica equations describe learning at finite signal strength β. A weak-signal expansion ($\beta \ll 1$) isolates a simple expression for the sample complexity as a function of the number of masks. For activations satisfying $\sigma(0) = 0$ and $\sigma(x) = s_\sigma x + O(x^2)$ as $x \to 0$, with $0 < s_\sigma < \infty$, fixed ρ, $\lambda_R = 0$, we have that the weak recovery threshold follows

$$\alpha_{c,\mathrm{sm}}^{(K)} = \frac{\kappa_2}{\beta^2}\left[1 + \frac{1}{K}\left(\frac{1}{2\rho(1-\rho)} - 1\right)\right][1 + o_\beta(1)], \qquad \kappa_2 = \mathbb{E}_\gamma[\gamma^2]. \tag{7}$$

The derivation and conditions underlying this expansion are given in Apps. D.2 and D.2.1. The prefactor κ_2/β^2 sets the weak-signal scale, including the cost of variance heterogeneity. At leading order, mask diversity lowers the excess threshold by a term proportional to $1/K$: doubling K halves this excess sample requirement, with smaller absolute gains at each doubling.

This leading law is shared by linear, tanh, and ELU activations; ReLU has a different coefficient but retains the same ordering in K (App. D.2). Beyond weak signal, Fig. 18 compares linear threshold ratios computed from the full replica equations with the asymptotic prediction. The ratios continue to decrease with K at the finite β, even when they depart from the weak-signal expression. Finite-dimensional training in Fig. 3 supports this predicted ordering across all four activations.

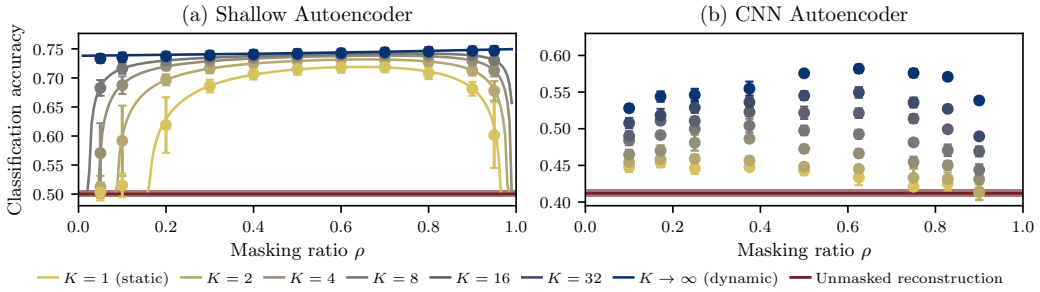


Figure 4: **Finite mask diversity interpolates between static and dynamic training.** Downstream accuracy versus masking ratio for **(a)** the solvable model, with theoretical curves and simulation markers, and **(b)** a CNN autoencoder on CIFAR-10. Increasing K improves accuracy in the settings shown, and the curves approach the dynamic masking reference. The number of training samples is held fixed. Bars and shading show ± 1 SD across seeds; see App. H for details.

Mask diversity can change the learning mechanism. Beyond the weak-signal expansion, replicas reveal a qualitative difference between static and dynamic masking. In the linear case, at the weak recovery transition, static masking does not preferentially use cleaner coordinates to identify the feature: its threshold reflects noise heterogeneity only through the aggregate moment κ_2. Dynamic masking can instead find and favor cleaner coordinates as recovery begins, implicitly learning which coordinates provide more reliable information. At suitable masking ratios, this adaptation provides an additional benefit not captured by the leading weak-signal law (7). App. D.4 derives this distinction and compares noise distributions with identical static but different dynamic thresholds.

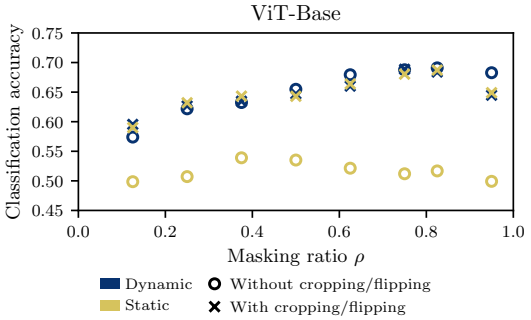


Figure 5: **Mask renewal and image augmentation.** Linear-probe validation top-5 accuracy for ViT-Base on ImageNet-100. Blue/yellow indicate dynamic/static masking; circles/crosses denote pretraining without/with stochastic cropping and flipping, respectively. All four conditions use a common seed, data split, initialization, optimizer settings, and linear-probing protocol. Each point represents one run. The blue cross at $\rho = 0.75$ is the purple horizontal line in Fig. 1(c). See App. H for training details.

Interaction with the masking ratio. The finite-K penalty in (7) is smallest at balanced masking and grows towards either endpoint. Near these extremes, each view offers relatively few visible or hidden coordinates, and a small collection of masks provides few distinct prediction tasks. Increasing K reduces this penalty. The shallow-model sweep in Fig. 4(a) illustrates the associated improvement in accuracy and the broader range of useful masking ratios as K increases. At finite signal strength, the best downstream masking ratio also depends on the activation and the learned representation's noise, so it need not minimize the leading threshold alone.

4 Testing the mask-diversity prediction in vision and language

Fig. 4 links the theoretical prediction to the CNN experiments: increasing the number K of fixed masks per sample improves downstream accuracy towards the dynamic-masking reference in both the solvable model (a) and CNN autoencoder (b), with the underlying dataset fixed. We test mask diversity further in ViT and BERT, and examine whether random cropping and flipping narrow the static–dynamic gap in vision.

Isolating mask resampling. We test the qualitative mask-diversity prediction in CNN autoencoders on CIFAR-10 and ViT-Base on ImageNet-100. Without stochastic cropping or flipping during pretraining, dynamic masking improves downstream accuracy over static masking in both architectures (Fig. 1(b,c)); see Fig. 7 for ViT top-1 accuracy. At suitable masking ratios, resampling alone reaches accuracy comparable to our reference combining dynamic masking with random cropping

and flipping (He et al., 2022). With extended pretraining and probing schedules, dynamic masking (without stochastic cropping or flipping) reaches accuracy comparable to the published MAE baseline (Tables 3 and 4).

Interaction with cropping and flipping. Fig. 5 compares static and dynamic ViT masking with and without cropping and flipping. Under static masking, patch positions remain fixed while transformations change the visible content and reconstruction targets. Cropping and flipping substantially improve static masking, narrowing the gap to dynamic masking. Fig. 14 extends the comparison to our CNN experiments with CIFAR-10 leading to the same qualitative conclusions.

Finite mask diversity. In the CNN sweep of Fig. 4(b), each image is assigned K independently drawn masks that are reused throughout pretraining, whereas dynamic masking draws a fresh mask at every presentation. For validation-selected encoders, mean downstream accuracy generally increases with K, approaching the dynamic-masking reference while using the same pretraining images. To assess sensitivity to the training protocol, we repeat the comparison at $\rho = 0.625$ with five paired seeds and a fixed learning rate under a schedule defined in optimizer updates. The mask-diversity trend persists both with validation-selected checkpoints and when all encoders are evaluated at the same 125k-update checkpoint (App. H.10).

For unmasked reconstruction in Fig. 1, we found that lower reconstruction error need not imply better representations for the downstream task: continued pretraining can reduce reconstruction MSE while linear probing degrades (Figs. 12 and 13).

Mask diversity in language modeling. A BERT-Medium pilot compares $K = 1$, $K = 10$, and dynamic masking, using only `[MASK]` replacements (Table 1). The corpus, architecture, and pretraining budget are shared across masking conditions. The ordering dynamic masking $> K = 10 > K = 1$ holds in mean development score on seven of the eight GLUE tasks. Additional results and experimental details are provided in App. H.13.

Table 1: Development scores ($\times 100$, $\uparrow$) for BERT-Medium under different masking procedures, on a subset of GLUE tasks. AVG is the unweighted mean of the task-specific scores over all eight evaluated tasks (accuracy; Spearman correlation for STS-B; MCC for MRPC, CoLA, and RTE), using MNLI-m for MNLI. Entries are means over three pretraining seeds, each averaged over two fine-tuning runs; standard deviations are in Table 5. See App. H.13 for details.

Masking	MNLI-m	QNLI	QQP	SST-2	AVG
$K = 1$	72.9	82.6	85.9	87.6	62.7
$K = 10$	76.3	84.4	87.0	88.1	65.8
Dynamic	76.8	85.5	87.2	89.4	66.4

5 Discussion

Our analysis separates two benefits of masked learning: masked prediction can favor shared structure over nuisance variance, while mask resampling can reduce the sample complexity required to recover that structure. We establish the first benefit rigorously in the linear model and quantify the second through replica calculations that agree with simulations. The CNN and ViT experiments, together with the BERT pilot, support the relevance of mask diversity across architectures and data modalities.

This perspective connects masked reconstruction to theories of data augmentation. Analyses based on transformation groups explain statistical gains through invariance and averaging (Chen et al., 2020a; Mei et al., 2021); in joint-embedding self-supervised learning, multiple views can reduce augmentation variability while retaining the variability associated with the original observations (Cabannes et al., 2023). Our analysis gives a complementary account for masked reconstruction: changing K preserves both the underlying dataset and the distribution of individual masks, yet changes the predicted recovery threshold. The weak-signal law quantifies the additional sample requirement at finite K and the diminishing gains from increasing mask diversity.

The ViT comparison suggests that geometric augmentation can partly compensate for mask reuse. A predictive theory of augmentation should therefore explain when geometric transformations and mask resampling offer complementary benefits, and when one reduces the benefit of the other. Our replica analysis also identifies a qualitative change in learning: dynamic masking can exploit cleaner coordinates at the recovery transition, while static masking lacks this adaptation. Whether resampling similarly changes which information is learned in larger models remains an open question.

Better representations need not require more data; they can come from a greater diversity of prediction tasks on the same observations.

AI USE STATEMENT

We used generative AI tools in an iterative, author-led research process to assist with literature retrieval, drafting, and polishing. Additional uses included formulating and refining mathematical claims, developing and revising derivations and proof arguments, including input of proof ideas from AI, discussing experimental design and interpretation, implementing and debugging numerical routines, and preparing scientific plots. Throughout this collaboration, the authors actively directed the work, critically examined the suggestions, and checked and revised AI-assisted outputs at every stage, including the mathematical arguments, computational methods, references, and written presentation. All scientific judgments and decisions about the manuscript's content remained with the authors. The authors take full responsibility for the manuscript.

ETHICS STATEMENT

This work studies the statistical properties of masked representation learning through mathematical analysis, synthetic simulations, and experiments on existing vision and language benchmarks. It involves no new collection of data from human participants. Learned representations may inherit biases and sensitive information from their training data; improvements in downstream accuracy do not establish fairness or privacy guarantees. These issues require separate assessment for any downstream application.

REPRODUCIBILITY STATEMENT

The data model, training objectives, and evaluation protocol are defined in Sec. 1 and App. B.2. The replica calculations, threshold derivations, and linear separation are documented in Apps. C.1, D.2 and E; the Bayes and downstream analyses are in Apps. F and G.2. App. H records the experimental settings, evaluation protocols, and uncertainty conventions. Source code and run configurations are supplied to reviewers.

REFERENCES

Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.

Urte Adomaityte, Gabriele Sicuro, and Pierpaolo Vivo. Classification of heavy-tailed features in high dimensions: a superstatistical approach. *Advances in Neural Information Processing Systems*, 36: 43880–43893, 2023.

Urte Adomaityte, Leonardo Defilippis, Bruno Loureiro, and Gabriele Sicuro. High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(11):114002, 2024.

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International conference on machine learning*, pp. 1416–1429. PMLR, 2023.

Z. D. Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1): 316–345, 1998. doi: 10.1214/aop/1022855421.

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993. doi: 10.1214/aop/1176989118.

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(5):1643–1697, 2005. doi: 10.1214/009117905000000233.

Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/balestriero24b.html.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*, 2022. arXiv preprint arXiv:2106.08254.

Peter N. Belhumeur, Joao P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Anthony Bisulco, Rahul Ramesh, Randall Balestriero, and Pratik Chaudhari. From linearity to non-linearity: How masked autoencoders capture spatial correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16441–16450, 2025.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11): 1093–1095, 2013.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The SSL interplay: Augmentations, inductive bias, and generalization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3252–3298. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cabannes23a.html.

Shuxiao Chen, Edgar Dobriban, and Jane Lee. A group-theoretic framework for data augmentation. In *Advances in Neural Information Processing Systems*, volume 33, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/f4573fc71c731d5c362f0d7860945b88-Abstract.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Vicente Conde Mendes, Lorenzo Bardone, Cédric Koller, Jorge Medina Moreira, Vittorio Erba, Emanuele Troiani, and Lenka Zdeborová. A solvable high-dimensional model where nonlinear autoencoders learn structure invisible to PCA while test loss misaligns with generalization. *International Conference on Machine Learning*, 2026. arXiv:2602.10680.

Hugo Cui. High-dimensional learning of narrow neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(2):023402, 2025.

Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems*, 36:11850–11890, 2023.

Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. *Journal of Machine Learning Research*, 17(141):1–41, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Mingze Dong, Leda Wang, and Yuval Kluger. Understanding and enhancing mask-based pretraining towards universal representations. *Advances in Neural Information Processing Systems*, 38: 142696–142739, 2025.

Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pp. 943–959. PMLR, 2016.

Stanislav Fort, Andrew Brock, Razvan Pascanu, Soham De, and Samuel L Smith. Drawing multiple augmentation samples per image during training efficiently decreases test error. *arXiv preprint arXiv:2105.13343*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

David Hong, Laura Balzano, and Jeffrey A Fessler. Asymptotic performance of PCA for high-dimensional heteroscedastic data. *Journal of multivariate analysis*, 167:435–452, 2018.

David Hong, Fan Yang, Jeffrey A Fessler, and Laura Balzano. Optimally weighted PCA for high-dimensional heteroscedastic data. *SIAM Journal on Mathematics of Data Science*, 5(1):222–250, 2023.

Peter Izsak, Moshe Berchansky, and Omer Levy. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10644–10652. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.831. URL https://aclanthology.org/2021.emnlp-main.831/.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. doi: 10.3150/15-BEJ730. URL https://arxiv.org/abs/1405.2468.

Lingjing Kong, Martin Q Ma, Guangyi Chen, Eric P Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7918–7928. IEEE, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

William Leeb and Elad Romanov. Optimal spectral shrinkage and PCA with heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(5):3009–3037, 2021.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 680–687. IEEE, 2015.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10144–10157. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3351–3418. PMLR, 2021. URL https://proceedings.mlr.press/v134/mei21a.html.

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Valentina Njaradi, Clémentine Dominé, Rachel Swanson, Marco Mondelli, and Andrew Saxe. Optimal representation size: High-dimensional analysis of pretraining and linear probing. *arXiv preprint arXiv:2605.20105*, 2026.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-4009. URL https://aclanthology.org/N19-4009/.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pp. 18499–18519. PMLR, 2022.

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. Mapping of attention mechanisms to a generalized potts model. *Physical Review Research*, 6(2):023057, 2024.

Jack W. Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995. doi: 10.1006/jmva.1995.1058.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019. URL https://arxiv.org/abs/1908.08962.

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(110):3371–3408, 2010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. 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*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2985–3000, 2023.

Arie Wortsman Zurich, Federica Gerace, Bruno Loureiro, and Yue M. Lu. A random matrix theory of masked self-supervised learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UuH4mmAGPJ.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9643–9653. IEEE, 2022.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Anru R Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80, 2022a.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022b.

Yihan Zhang and Marco Mondelli. Matrix denoising with doubly heteroscedastic noise: Fundamental limits and optimal spectral methods. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2955. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a96368eb38bce0956a1132154d70d72d-Abstract-Conference.html.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html.

APPENDIX CONTENTS

A Detailed comparison with the closest prior work **17**

B Model, notation, and effective output channel **20**

B.1 Notation 20

B.2 Detailed model and effective output channel 21

C Replica derivation and output-channel specializations **25**

C.1 Generic replica derivation 25

C.2 Output-channel specializations 33

D Replica recovery predictions and the role of mask diversity **36**

D.1 Recovery criterion and fixed-signal factorization 36

D.2 Mask diversity in the weak-signal regime 40

D.3 Dynamic linear masking at fixed signal strength 47

D.4 Mask diversity changes how the learner uses heterogeneous noise 49

E Proofs of the linear reconstruction separation **52**

E.1 PCA obstructions and bounded-noise masking recovery 52

E.2 Rigorous dynamic linear recovery with unbounded coordinate noise 58

F Bayes-optimal pretraining and AMP **62**

F.1 Mapping to bipartite low-rank estimation 62

F.2 Asymptotic identifiability of the coordinate variances 63

F.3 Bayes-optimal output channel and Fisher information 64

F.4 Denoisers 65

F.5 Explicit heteroskedastic bipartite AMP 65

F.6 State evolution 67

F.7 Fixed-point equations in the present notation 68

F.8 Continuous weak-recovery threshold 68

F.9 Gaussian-prior specialization 69

G Additional observables and finite-shot downstream analysis **69**

G.1 Training and test errors 69

G.2 Downstream task 70

H Experimental settings and reporting **72**

H.1 Masking and augmentation protocols 72

H.2 Finite-size effects of PCA 75

H.3 Top-1 accuracy for ViT 75

H.4 Noise-aware spectral methods 75
H.5 Dynamic masking and nonlinear autoencoders 77
H.6 Reconstruction errors across mask diversity 77
H.7 Number of downstream labels 78
H.8 Early-stopping without masking 78
H.9 Mask renewal and image augmentation in CNNs 80
H.10 CNN training budgets and validation selection 80
H.11 Numerical assessment of finite-mask threshold ratios 81
H.12 Synthetic experiments and numerical predictions 82
H.13 Mask diversity in BERT 84
H.14 Bayes curves and AMP estimates 85

A Detailed comparison with the closest prior work

The main text outlines the broad lines of related work. Here we expand on the closest references to clarify both their settings and what their analyses establish. We compare data models, prediction objectives, and masking protocols, as well as the observables and guarantees studied. These distinctions locate our contributions: recovery under heterogeneous noise for the exact mask-averaged objective, and replica predictions for mask diversity and downstream performance.

Reconstruction and useful features. The reconstruction–perception mismatch identified by Balestriero & LeCun (2024) motivates our separation: a direction can explain substantial input variance without carrying useful latent information. Their closed-form linear denoising analysis shows how corruption, including masking, can improve this alignment. In our spiked model, coordinate-dependent noise makes the obstruction explicit, and hidden-target prediction admits a recovery guarantee at linear sample complexity: every global minimizer of the exact mask-averaged linear objective weakly recovers at linear sample complexity while ordinary PCA fails (Theorem 2.1).

Full reconstruction from masked inputs. The linear MAE analyzed by Bisulco et al. (2025) has untied encoder and decoder matrices, whereas our bottleneck uses tied weights. More crucially, their mask-averaged analytical loss reconstructs all coordinates from masked inputs, including visible coordinates; ours scores only hidden targets. The visible-coordinate rewards retain a contribution from reproducing input variance, changing both the loss landscape and its masking-ratio dependence. Their characterization of critical points and of the spatial correlations selected by patch size and masking ratio is therefore closely relevant, but its spectral solution does not directly solve our objective. Our statistical analysis follows how the structure favored by hidden-target prediction translates into recovery at finite sample size and under finite mask reuse. These are different objectives and constraints, rather than a general performance ordering.

Single-token prediction and factored attention. Rende et al. (2024) connect factored attention to generalized Potts conditionals and use replicas to predict generalization in a Gaussian relaxation. There, predicting one coordinate from all the others probes a full-rank random precision matrix. This differs from our low-rank signal with diagonal heterogeneous noise both in correlation structure and in the number of targets per mask: a positive limiting fraction is hidden in our model. The shared tied bottleneck couples these prediction tasks, so its alignment with the planted feature, and the downstream utility of that alignment, become explicit observables as the number of reused masks varies.

The joint matrix of single-coordinate predictors. Single-coordinate prediction can itself reveal latent structure: Wortsman Zurich et al. (2026) analyze the joint matrix of self-supervised ridge predictors, rigorously characterizing its spectrum, prediction risk, and spiked-model BBP recovery transition. They also establish prediction-risk advantages over PCA in structured covariance settings. Coupling an extensive set of hidden targets through a shared tied bottleneck changes the estimator in our work; mask diversity is an additional statistical control. Our analysis follows its effects on recovery and downstream performance via replicas, alongside a rigorous recovery separation under heteroskedastic noise for the exact mask-averaged objective. Nevertheless, it is important to stress that recovering latent structure through the data correlations exploited by masked prediction is already established in their work.

Masking ratio and a selected prediction target. Dong et al. (2025) start from a loss over all hidden coordinates, but decouple their linear predictors and solve a stylized regression for one selected target, $y = X\boldsymbol{\beta} + \epsilon$ in their notation. Their isotropic and spiked covariance assumptions describe the remaining covariates X; the target coefficient and residual noise are specified separately. Consequently, this target need not be statistically representative of the other coordinates of the augmented sample (X, y). The resulting risk formulas rigorously quantify how masking both corrupts covariates and selects the samples supplying this target. In our common-factor model, every coordinate follows the same generative construction without a designated response, and all hidden targets train the same tied bottleneck; the analysis then follows collective feature recovery. Their fixed-ratio theory motivates randomized masking ratios, with downstream benefits assessed across

modalities; our finite-reuse calculation keeps the ratio fixed and links representation alignment to a finite-label probe, with replica rather than rigorous guarantees for the general learning curves.

Contrastive connections and downstream guarantees. The connection between masked reconstruction and useful representations is also explicit in Zhang et al. (2022b): reconstruction induces alignment on a mask-based augmentation graph, supporting downstream guarantees and an analysis of masking ratio and dimensional collapse. Our recovery thresholds address a more specific statistical question: how reusing a finite collection of masks changes the learned alignment at a fixed sample-to-dimension ratio. The downstream calculation then tracks the effect of that alignment on a probe trained with finitely many labels.

Identifiability in hierarchical models. Under the hierarchical generative and invertibility assumptions of Kong et al. (2023), masking ratio and patch size determine which latent variables are shared between visible and hidden regions and can be identified. In our additive-noise model, the relevant observable is partial alignment with a planted direction at finite sample-to-dimension ratio. The recovery transition and its dependence on mask diversity concern statistical estimation under these different data and encoder assumptions, rather than an alternative identifiability statement for their hierarchy.

Nonlinear recovery beyond PCA. Conde Mendes et al. (2026) show that nonlinear autoencoders recover higher-order structure invisible to PCA, and that reconstruction loss can misalign with representation quality. Our signal is covariance-visible, but heterogeneous variances obstruct ordinary PCA. Hidden-target prediction changes which covariance structure the objective rewards, yielding a separation that does not require nonlinearity. Complementary high-dimensional work studies how pretraining sample size and representation dimensionality jointly control downstream linear-probe generalization in spiked models (Njaradi et al., 2026).

Noise-aware spectral alternatives. HeteroPCA (Zhang et al., 2022a) and whitening-based methods (Leeb & Romanov, 2021) address the same nuisance of heterogeneous noise through spectral correction. Fig. 8 compares such estimators on the same observations: our separation from ordinary PCA does not imply a separation from all spectral methods. The comparison locates masking among several routes to recovery. It changes which structure the reconstruction objective rewards, without supplying the autoencoder with the coordinate variances or their distribution.

Bayes-optimal inference and the role of prior information. Our Bayes analysis builds on the low-rank matrix-estimation framework of Lesieur et al. (2015; 2017), which connects matched priors and output channels to approximate message passing and state evolution. We specialize this framework to coordinate-dependent Gaussian noise to obtain the reference learning curves and AMP updates in App. F. For Gaussian latent priors, the continuous recovery threshold is $\alpha_c^{\mathrm{BO}} = [\beta^2 \mathbb{E}_\gamma(\gamma^{-2})]^{-1}$ when the inverse second moment is finite (App. F.8). Its dependence on inverse variances highlights the information available in low-noise coordinates. Zhang & Mondelli (2024) provide an independent rigorous foundation: their analysis of doubly heteroskedastic matrix denoising yields this weak-estimation threshold under their Gaussian-prior and compact, positive covariance-support assumptions, and a covariance-aware spectral estimator attains it under their additional technical condition. Thus the Bayes reference connects our reconstruction-based learners to an established optimal-inference benchmark, rather than introducing a new information-theoretic limit.

The distinction in available information matters for interpreting the comparison. Our Bayes reference uses the matched latent priors and the realized coordinate variances, whereas the autoencoders receive neither the variances nor their distribution. A remaining gap to Bayes therefore reflects both the reconstruction objective and the additional information used by matched inference. Increasing mask diversity improves how that objective uses the observations; it does not supply the learner with the noise model. The variance-consistency calculation in App. F discusses learning the noise levels from the samples, separately from the known-variance benchmark used here.

Augmentation multiplicity and finite reuse. Prior work has studied the benefit of drawing multiple augmented views of the same example. Fort et al. (2021) show empirically that averaging several freshly sampled augmentations per image can improve supervised generalization, while Cabannes

et al. (2023) analyze the corresponding multi-view setting theoretically and separate variability due to sampling original examples from variability due to sampling their augmentations. These works vary the number of views used for an example while preserving the underlying augmentation distribution. We study a different axis: *finite reuse*. Each training sample is assigned K masks that are held fixed throughout learning, so changing K changes the finite collection of prediction tasks available to the learner rather than the number of views averaged within an update. Replica predictions quantify how this finite mask diversity changes feature recovery and downstream accuracy, including the weak-signal excess sample requirement in (7). The masks contain no additional information about the planted feature; we show that their benefit comes from how a richer collection of prediction tasks allows the objective to extract information already present in the observed samples.

Mask resampling and other augmentations. Dynamic masking in RoBERTa (Liu et al., 2019) and multiple masked views per update in data2vec 2.0 (Baevski et al., 2023) already exploit variation in prediction tasks. The interaction with image augmentation matters when interpreting its benefit: even fixed patch positions expose changing content under stochastic cropping and flipping. Our vision comparisons isolate this interaction, while the BERT pilot provides complementary evidence for finite mask diversity. Group-based accounts of augmentation through invariance and averaging (Chen et al., 2020a; Mei et al., 2021) offer another perspective; here the controlled variable is the reuse of prediction tasks from each original observation. These experiments test the qualitative role of mask diversity beyond the solvable model, without assuming that its quantitative recovery law transfers to these architectures and datasets.

B MODEL, NOTATION, AND EFFECTIVE OUTPUT CHANNEL

B.1 NOTATION

Symbol	Meaning
Norms	
$\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2$	Euclidean vector norm, $(\sum_i \|v_i\|^2)^{1/2}$
$\|A\|_{\mathrm{op}}$	matrix operator norm, $\sup_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$
$\|A\|_{\mathrm{F}}$	Frobenius norm, $(\sum_{i,j} \|A_{ij}\|^2)^{1/2}$
Data and model	
$d,\ n,\ \alpha = n/d$	dimension, number of samples, sample complexity
$\boldsymbol{u}^\star$	planted latent direction
λ_μ	latent variable of sample μ; $\mathbb{E}[\lambda^2] = 1$
β	signal strength/signal-to-noise ratio
$\gamma_i^\star \sim P_\gamma$	noise variance of coordinate $i \in [1, d]$; $\mathbb{E}_\gamma[\gamma] = 1$
$\kappa_2 = \mathbb{E}_\gamma[\gamma^2]$	variance heterogeneity; $\kappa_2 = 1$ iff homoskedastic
$\Gamma^\star$	$\mathrm{diag}(\gamma_1^\star, \dots, \gamma_d^\star)$
$\boldsymbol{w},\ \widehat{\boldsymbol{w}}$	tied encoder/decoder weights; empirical-risk minimizer
σ	activation function
λ_R	ridge parameter ($\lambda_R = 0$ is ridgeless)
Masking	
ρ	masking probability per coordinate
K	number of fixed masks per sample (mask diversity)
$\chi_{\mu i}^{(r)}$	mask indicator; 1 if coordinate i hidden in view r
h	full preactivation, shared by all K views of a sample
δ_r	mask-induced fluctuation of view r
Order parameters	
$M = d^{-1}(\boldsymbol{u}^\star)^\top \boldsymbol{w}$	teacher–student overlap
$q_I = d^{-1}\|\boldsymbol{w}\|_2^2$	Euclidean self-overlap
$q_\Gamma = d^{-1}\boldsymbol{w}^\top \Gamma^\star \boldsymbol{w}$	noise-weighted self-overlap
$\theta_{\boldsymbol{u}^\star} = \|M\|/[\sqrt{q_I}(\|\boldsymbol{u}^\star\|/\sqrt{d})]$	cosine similarity with $\boldsymbol{u}^\star$
$\Psi^{(s)}$	replicated free entropy with s replicas
η_{th}	Gibbs inverse temperature; $\eta_{\mathrm{th}} \to \infty$ for ERM
Thresholds and response functions	
$\alpha_c^{(K)}$	weak-recovery threshold at mask diversity K
α_c^{BO}	Bayes-optimal recovery threshold
$\widetilde{\boldsymbol{F}}_K$	whitened output force
$\mathcal{X}_K$	weak-signal limit of the averaged informative response
T_K	limiting, q_Γ-normalized force second moment
$\eta_{K,c}^2 = \mathcal{E}_{K,c}/(q_{\Gamma,c}\mathcal{X}_{K,c}^2) \geq 1$	exact output-efficiency factor, (D.28)
$v_K(\gamma) \propto (1 + \tau_K\gamma)^{-1}$	input susceptibility of the uninformative saddle
$\tau_{K,c}$	susceptibility shape; sign decides up- or down-weighting
$\mathcal{R}_\gamma(\tau)$	heteroskedastic reweighting factor; $\mathcal{R}_\gamma(0) = \kappa_2$
ζ_c	spectral parameter of the full-reconstruction branch
γ_+	essential upper edge of P_γ

Bold symbols denote vectors, and a star marks ground-truth quantities. Unless conditioning is explicitly stated or specified in the surrounding text, an unsubscripted $\mathbb{E}$ denotes expectation over all model randomness. Subscripts specify the variables being averaged, with other quantities held fixed: in particular, $\mathbb{E}_\gamma$ averages over the noise-variance distribution P_γ, and $\mathbb{E}_\chi$ averages over masks at a fixed dataset and candidate weight vector. Conditional expectations are written as $\mathbb{E}[\,\cdot \mid \cdot\,]$. Population quantities defined by averaging over fresh samples at fixed quenched variables retain their dependence on the realized teacher $\boldsymbol{u}^\star$ and variance matrix $\Gamma^\star$; for example, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mid \boldsymbol{u}^\star, \Gamma^\star] = \Gamma^\star + \beta\boldsymbol{u}^\star(\boldsymbol{u}^\star)^\top/d$. We also write $[m] = \{1, \dots, m\}$ for any positive integer m.

Samples index rows and coordinates index columns throughout:

$$X = (x_{\mu i}), \quad Z = (z_{\mu i}) \in \mathbb{R}^{n\times d}, \qquad X = \sqrt{\beta/d}\,\boldsymbol{\lambda}(\boldsymbol{u}^\star)^\top + Z(\Gamma^\star)^{1/2}. \tag{B.1}$$

Here $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^\top$.

Here $\mathcal{E}_{K,c} = \mathbb{E}\|\widetilde{\boldsymbol{F}}_{K,c}\|_2^2$ and $\mathcal{X}_{K,c}$ are the force second moment and averaged informative response at criticality. Along the regular weak-signal branch at fixed finite K, $\eta_{K,c}^2 \to T_K/\mathcal{X}_K^2$ as $\beta \downarrow 0$, with the limiting quantities defined in (D.65)–(D.66).

The cosine identity in the table is exact for nonzero teacher and weights. Since $\|\boldsymbol{u}^\star\|^2/d \to 1$, its high-dimensional expression is $|M|/\sqrt{q_I}$. For a fixed candidate weight vector, independent masks make the fluctuations δ_r independent across views conditional on the sample; averaging over the shared sample need not preserve independence at finite dimension. Independence in the limiting Gaussian channel is understood under the assumptions in App. B.2.

B.2 Detailed model and effective output channel

As mentioned in the main document, we study a single hidden-neuron autoencoder with tied weights and $K = O_d(1)$ fixed masks per training sample. The high-dimensional limit is the linear one where

$$n, d \to \infty, \qquad \frac{n}{d} \to \alpha, \tag{B.2}$$

with K fixed. Dynamic masking is obtained afterward through the sequential limit $K \to \infty$ at the level of the effective output channel.

B.2.1 Data model

Why allow unequal noise variances? The distinction between variance and task-relevant information motivates the model. Reconstruction objectives can emphasize components that are weakly useful for perception (Balestriero & LeCun, 2024). Examples of nuisance variation dominating leading components include illumination in face recognition (Belhumeur et al., 1997), natural-image backgrounds superimposed on handwritten digits (Abid et al., 2018), and heterogeneous noise in biological data (Brennecke et al., 2013). These examples motivate separating useful structure from variance; they do not identify their covariance structure with the coordinate-noise model studied here. Our model isolates this distinction using a single shared feature and independently varying coordinate noise levels.

The training samples $\boldsymbol{x}_\mu \in \mathbb{R}^d$ are generated from the heteroskedastic rank-one model

$$x_{\mu i} = \sqrt{\frac{\beta}{d}}\,\lambda_\mu u_i^\star + \sqrt{\gamma_i^\star}\,z_{\mu i}, \qquad \mu = 1, \dots, n, \quad i = 1, \dots, d. \tag{B.3}$$

The sample latent variables satisfy

$$\lambda_\mu \overset{\text{iid}}{\sim} P_\lambda, \qquad \mathbb{E}[\lambda] = 0, \qquad \mathbb{E}[\lambda^2] = 1. \tag{B.4}$$

The teacher direction is quenched and has coordinates

$$u_i^\star \overset{\text{iid}}{\sim} P_u, \qquad \mathbb{E}[u^\star] = 0, \qquad \mathbb{E}[(u^\star)^2] = 1. \tag{B.5}$$

The noise variables are

$$z_{\mu i} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1). \tag{B.6}$$

The coordinate variances are also quenched,

$$\gamma_i^\star \overset{\text{iid}}{\sim} P_\gamma, \qquad \gamma_i^\star > 0, \qquad \mathbb{E}_\gamma[\gamma] = 1, \qquad \mathbb{E}_\gamma[\gamma^2] < \infty. \tag{B.7}$$

It is convenient to collect the coordinate variances into the quenched diagonal matrix

$$\Gamma^\star := \operatorname{diag}(\gamma_1^\star, \dots, \gamma_d^\star)\,. \tag{B.8}$$

Thus $\boldsymbol{u}^\star$ and $\boldsymbol{\gamma}^\star = (\gamma_1^\star, \dots, \gamma_d^\star)$, equivalently $\Gamma^\star$, are coordinate-wise quenched disorder, whereas λ_μ and $z_{\mu i}$ are sample-dependent random variables. Unless stated otherwise, all random families introduced in the data model are mutually independent. In particular,

$$\boldsymbol{u}^\star, \qquad \boldsymbol{\gamma}^\star, \qquad \{\lambda_\mu\}_{\mu=1}^n, \qquad \{z_{\mu i}\}_{\mu,i} \tag{B.9}$$

are mutually independent.

B.2.2 FINITE-K MASKED AUTOENCODER

Fix a masking probability $0 < \rho < 1$. For each unmasked sample $\boldsymbol{x}_\mu$, we draw K independent masks

$$\chi_{\mu i}^{(r)} \overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \qquad r = 1, \dots, K, \tag{B.10}$$

with the convention

$$\chi_{\mu i}^{(r)} = 1 \quad \Longleftrightarrow \quad \text{coordinate } i \text{ is masked in view } r \text{ of sample } \mu. \tag{B.11}$$

The corresponding visible indicator is simply $1 - \chi_{\mu i}^{(r)}$; meaning that the input and output of the autoencoder will be complementary.

The collection of masks $\{\chi_{\mu i}^{(r)}\}_{\mu,i,r}$ is independent of all random variables entering the data-generating model. For fixed μ, however, the K masked views share the same underlying sample $\boldsymbol{x}_\mu$. The finite-K dataset is therefore

$$\mathcal{D}_K = \left\{ \left(\boldsymbol{x}_\mu, \boldsymbol{\chi}_\mu^{(1)}, \dots, \boldsymbol{\chi}_\mu^{(K)} \right) \right\}_{\mu=1}^{n}. \tag{B.12}$$

The masks are part of the training disorder and remain fixed during optimization.

Architecture and masked reconstruction loss. The model has one hidden unit, no bias, a scalar activation function σ, and tied encoder/decoder weights $\boldsymbol{w} \in \mathbb{R}^d$. For view r of sample μ, the encoder receives only the visible coordinates. Its preactivation is

$$h_{\mu r,\text{vis}}(\boldsymbol{w}) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} (1 - \chi_{\mu i}^{(r)}) w_i x_{\mu i}. \tag{B.13}$$

The tied decoder reconstructs coordinate i as

$$\widehat{x}_{\mu i}^{(r)} = \frac{w_i}{\sqrt{d}} \, \sigma(h_{\mu r,\text{vis}}(\boldsymbol{w})) \, . \tag{B.14}$$

Only the masked coordinates contribute to the reconstruction loss. The loss of view r is therefore

$$\ell_d^{(r)} \left(\boldsymbol{w}; \boldsymbol{x}_\mu, \boldsymbol{\chi}_\mu^{(r)} \right) = \sum_{i=1}^{d} \chi_{\mu i}^{(r)} \left[x_{\mu i} - \frac{w_i}{\sqrt{d}} \sigma(h_{\mu r,\text{vis}}(\boldsymbol{w})) \right]^2 . \tag{B.15}$$

The finite-K loss of the original sample is the average over its K fixed masks,

$$\ell_{K,d}(\boldsymbol{w}; \boldsymbol{x}_\mu, \boldsymbol{\chi}_\mu^{(1:K)}) = \frac{1}{K} \sum_{r=1}^{K} \ell_d^{(r)} \left(\boldsymbol{w}; \boldsymbol{x}_\mu, \boldsymbol{\chi}_\mu^{(r)} \right) . \tag{B.16}$$

and we study the standard risk minimization of

$$\mathcal{R}_{K,d}(\boldsymbol{w}) = \frac{\lambda_R}{2} \|\boldsymbol{w}\|_2^2 + \sum_{\mu=1}^{n} \ell_{K,d} \left(\boldsymbol{w}; \boldsymbol{x}_\mu, \boldsymbol{\chi}_\mu^{(1:K)} \right) . \tag{B.17}$$

Effective sample loss. For each view, define the contribution of the masked coordinates to the full preactivation,

$$h_{\mu r,\text{mask}}(\boldsymbol{w}) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \chi_{\mu i}^{(r)} w_i x_{\mu i}. \tag{B.18}$$

Expanding (B.15) and dropping the term independent of $\boldsymbol{w}$ gives

$$\ell_d^{(r),\text{inf}} = \frac{\sigma(h_{\mu r,\text{vis}})^2}{d} \sum_{i=1}^{d} \chi_{\mu i}^{(r)} w_i^2 - 2\sigma(h_{\mu r,\text{vis}}) h_{\mu r,\text{mask}}. \tag{B.19}$$

Introduce the Euclidean self-overlap

$$Q_I(\boldsymbol{w}) = \frac{1}{d} \sum_{i=1}^{d} w_i^2 = \frac{1}{d} \boldsymbol{w}^\top \boldsymbol{w}. \tag{B.20}$$

Self-averaging of the masked norm. In the one-sample disorder average, the candidate weight vector $\boldsymbol{w}$ is held fixed while the masks are averaged. Hence, we need to assume the following

$$Q_I(w) = O(1), \qquad \frac{1}{d^2}\sum_{i=1}^{d} w_i^4 \longrightarrow 0.$$

to keep the replica derivation consistent. The second condition excludes a macroscopic contribution from any single coordinate to the Euclidean norm. For independent Bernoulli mask entries, with the candidate weights held fixed,

$$\mathbb{E}_\chi\left[\frac{1}{d}\sum_{i=1}^{d}(\chi_{\mu i}^{(r)} - \rho)w_i^2\right] = 0, \qquad \mathbb{E}_\chi\left[\left(\frac{1}{d}\sum_{i=1}^{d}(\chi_{\mu i}^{(r)} - \rho)w_i^2\right)^2\right] = \frac{\rho(1-\rho)}{d^2}\sum_{i=1}^{d} w_i^4.$$

Consequently, for each fixed sample and fixed K,

$$\max_{1\le r\le K}\left|\frac{1}{d}\sum_{i=1}^{d}\chi_{\mu i}^{(r)} w_i^2 - \rho Q_I(w)\right| \xrightarrow[d\to\infty]{\mathbb{P}_\chi} 0. \tag{B.21}$$

The same conclusion holds jointly for any fixed finite collection of candidate replicas. For fixed K, under the high-dimensional self-averaging used throughout the analysis,

$$\frac{1}{d}\sum_{i=1}^{d}\chi_{\mu i}^{(r)} w_i^2 \xrightarrow[d\to\infty]{} \rho Q_I(\boldsymbol{w}), \qquad r = 1,\dots,K. \tag{B.22}$$

Hence the effective high-dimensional sample loss is

$$\ell_K\left(\{h_{r,\mathrm{vis}}, h_{r,\mathrm{mask}}\}_{r=1}^{K}; Q_I\right) = \frac{1}{K}\sum_{r=1}^{K}\left[\rho Q_I \sigma(h_{r,\mathrm{vis}})^2 - 2\sigma(h_{r,\mathrm{vis}})h_{r,\mathrm{mask}}\right] \tag{B.23}$$

From this point onward, the sample index μ is suppressed when discussing the one-sample channel. We also work with the centered, $\boldsymbol{w}$-dependent objective obtained after dropping the reconstruction term independent of $\boldsymbol{w}$; this leaves the Gibbs measure and the ERM minimizers unchanged.

B.2.3 Low-dimensional field representation

Define the ordinary full preactivation

$$h(\boldsymbol{w}) = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} w_i x_i. \tag{B.24}$$

Since every coordinate is either visible or masked in view r,

$$h_{r,\mathrm{vis}} + h_{r,\mathrm{mask}} = h, \qquad r = 1,\dots,K. \tag{B.25}$$

This identity is exact at finite d. In particular, the $2K$ quantities

$$(h_{1,\mathrm{vis}},\dots,h_{K,\mathrm{vis}}, h_{1,\mathrm{mask}},\dots,h_{K,\mathrm{mask}}) \tag{B.26}$$

contain only $K+1$ independent directions.

Define the centered visible-field fluctuation of mask r by

$$\delta_r := h_{r,\mathrm{vis}} - (1-\rho)h. \tag{B.27}$$

Using (B.13) and (B.24),

$$\delta_r = \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\big(\rho - \chi_i^{(r)}\big)w_i x_i. \tag{B.28}$$

Therefore

$$h_{r,\mathrm{vis}} = (1-\rho)h + \delta_r, \qquad h_{r,\mathrm{mask}} = \rho h - \delta_r. \tag{B.29}$$

The natural field vector for the finite-K problem is thus

$$\boldsymbol{f}_K = (h, \delta_1, \dots, \delta_K) \in \mathbb{R}^{K+1}. \tag{B.30}$$

The centered mask coefficients satisfy

$$\mathbb{E}[\rho - \chi_i^{(r)}] = 0, \tag{B.31}$$

and, by independence of different masks,

$$\mathbb{E}\Big[(\rho - \chi_i^{(r)})(\rho - \chi_i^{(s)})\Big] = \rho(1-\rho)\delta_{rs}. \tag{B.32}$$

This is the basic reason why the mask fluctuations become diagonal in the reduced representation.

Substituting (B.29) into (B.23) gives the finite-K effective loss directly in terms of the nonredundant fields,

$$\ell_K(h, \delta_1, \dots, \delta_K; Q_I) = \frac{1}{K}\sum_{r=1}^{K}\Big[\rho Q_I \sigma((1-\rho)h + \delta_r)^2 - 2\sigma((1-\rho)h + \delta_r)\,(\rho h - \delta_r)\Big]. \tag{B.33}$$

Finite-K Gaussian channel. We use the replica method (Mézard et al., 1987; Mézard & Montanari, 2009; Zdeborová & Krzakala, 2016), a heuristic tool from statistical physics.

For replicas a, b, define the teacher–student overlap

$$M^{\mathsf{a}} := \frac{1}{d}\sum_{i=1}^{d} u_i^\star w_i^{\mathsf{a}} = \frac{1}{d}(\boldsymbol{u}^\star)^\top \boldsymbol{w}^{\mathsf{a}}. \tag{B.34}$$

More generally, for a deterministic or quenched metric $A \in \mathbb{R}^{d\times d}$, define the metric-dependent replica overlap

$$Q_A^{\mathsf{ab}} := \frac{1}{d}(\boldsymbol{w}^{\mathsf{a}})^\top A \boldsymbol{w}^{\mathsf{b}}. \tag{B.35}$$

The two metrics needed in the present heteroskedastic problem are

$$Q_I^{\mathsf{ab}} = \frac{1}{d}\sum_{i=1}^{d} w_i^{\mathsf{a}} w_i^{\mathsf{b}}, \qquad Q_\Gamma^{\mathsf{ab}} = \frac{1}{d}\sum_{i=1}^{d} \gamma_i^\star w_i^{\mathsf{a}} w_i^{\mathsf{b}} = \frac{1}{d}(\boldsymbol{w}^{\mathsf{a}})^\top \Gamma^\star \boldsymbol{w}^{\mathsf{b}}. \tag{B.36}$$

Thus Q_I^{ab} is the ordinary Euclidean replica overlap, whereas Q_Γ^{ab} is the noise-weighted overlap controlling the covariance of the projected Gaussian data noise. In the homoskedastic limit $\Gamma^\star = I$, the two coincide: $Q_\Gamma^{\mathsf{ab}} = Q_I^{\mathsf{ab}}$. For a single replica we use the shorthand $Q_I := Q_I^{\mathsf{aa}}$ and $Q_\Gamma := Q_\Gamma^{\mathsf{aa}}$.

In the following sample and mask averages, the teacher $\boldsymbol{u}^\star$, variance realization $\Gamma^\star$, and candidate replicas are held fixed; their conditioning is suppressed. Conditionally also on the sample latent variable λ, the full preactivation satisfies

$$\mathbb{E}[h^{\mathsf{a}} \mid \lambda] \longrightarrow \sqrt{\beta}\lambda M^{\mathsf{a}}, \qquad \mathrm{Cov}(h^{\mathsf{a}}, h^{\mathsf{b}} \mid \lambda) \longrightarrow Q_\Gamma^{\mathsf{ab}}. \tag{B.37}$$

For the centered mask fluctuations,

$$\mathbb{E}[\delta_r^{\mathsf{a}} \mid \lambda] \longrightarrow 0, \tag{B.38}$$

$$\mathrm{Cov}(h^{\mathsf{a}}, \delta_r^{\mathsf{b}} \mid \lambda) \longrightarrow 0, \tag{B.39}$$

$$\mathrm{Cov}(\delta_r^{\mathsf{a}}, \delta_s^{\mathsf{b}} \mid \lambda) \longrightarrow \rho(1-\rho)\delta_{rs} Q_\Gamma^{\mathsf{ab}}. \tag{B.40}$$

The rank-one signal contribution to the centered fluctuations is subleading: because $\rho - \chi_i^{(r)}$ is centered, the corresponding empirical sum is $o_d(1)$ under the same self-averaging assumptions.

Let

$$\boldsymbol{M}_s = (M^1, \dots, M^s)^\top, \qquad \mathsf{Q}_{\Gamma,s} = (Q_\Gamma^{\mathsf{ab}})_{\mathsf{a},\mathsf{b}\le s}, \qquad \boldsymbol{e}_0 = (1, 0, \dots, 0)^\top \in \mathbb{R}^{K+1}. \tag{B.41}$$

Define

$$\Omega_K := \mathrm{diag}\,(1, \rho(1-\rho)\mathbf{I}_K)\,. \tag{B.42}$$

Then the replicated finite-K field vector has the asymptotic conditional law

$$(\boldsymbol{f}_K^1, \dots, \boldsymbol{f}_K^s) \mid \lambda \Longrightarrow \mathcal{N}_{(K+1)s}\Big(\sqrt{\beta}\lambda(\boldsymbol{M}_s \otimes \boldsymbol{e}_0), \mathsf{Q}_{\Gamma,s} \otimes \Omega_K\Big)\,. \tag{B.43}$$

For $0 < \rho < 1$, Ω_K is positive definite. This is why the reduced $(K+1)$-dimensional representation is preferable to the raw $2K$-dimensional visible/masked representation.

B.2.4 DYNAMIC-MASK OUTPUT CHANNEL

Dynamic masking is represented by the sequential limit

$$n, d \to \infty \quad \text{with } K \text{ fixed}, \qquad \text{followed by} \qquad K \to \infty. \tag{B.44}$$

For one replica, the reduced Gaussian channel of App. B.2.3 can be represented as

$$\delta_r = \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi_r, \qquad \xi_r \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad r = 1, \dots, K, \tag{B.45}$$

independently of the full preactivation h. Substituting this representation into (B.33) gives

$$\begin{aligned}\ell_K = \frac{1}{K}\sum_{r=1}^{K}\Big[&\rho Q_I \sigma\Big((1-\rho)h + \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi_r\Big)^2 \\ &- 2\sigma\Big((1-\rho)h + \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi_r\Big)\Big(\rho h - \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi_r\Big)\Big].\end{aligned} \tag{B.46}$$

By the law of large numbers,

$$\begin{aligned}\ell_K \xrightarrow[K\to\infty]{} \ell_{\text{dyn}}(h; Q_I, Q_\Gamma) := \mathbb{E}_\xi\Bigg[&\rho Q_I\,\sigma\Big((1-\rho)h + \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi\Big)^2 \\ &- 2\,\sigma\Big((1-\rho)h + \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi\Big)\Big(\rho h - \sqrt{\rho(1-\rho)Q_\Gamma}\,\xi\Big)\Bigg].\end{aligned} \tag{B.47}$$

Thus dynamic masking does not introduce a new high-dimensional field. Once the mask average is performed, the only explicit sample-dependent field is the ordinary preactivation h; the fluctuations generated by the masks are retained inside the averaged loss (B.47).

The reduced $(K+1)$-dimensional representation is sufficient for the replica calculation. For completeness, the covariance in the redundant $2K$ visible/masked representation is recorded here as a consistency check. The singular raw covariance can be recovered from the reduced variables, but it is not needed for the replica calculation. Let

$$\boldsymbol{h}_{\text{vis}} = (h_{1,\text{vis}}, \dots, h_{K,\text{vis}})^\top, \qquad \boldsymbol{h}_{\text{mask}} = (h_{1,\text{mask}}, \dots, h_{K,\text{mask}})^\top. \tag{B.48}$$

Then

$$\begin{pmatrix}\boldsymbol{h}_{\text{vis}} \\ \boldsymbol{h}_{\text{mask}}\end{pmatrix} = \underbrace{\begin{pmatrix}(1-\rho)\mathbf{1}_K & \mathbf{I}_K \\ \rho\mathbf{1}_K & -\mathbf{I}_K\end{pmatrix}}_{L_K}\begin{pmatrix}h \\ \boldsymbol{\delta}\end{pmatrix}. \tag{B.49}$$

Suppressing the common replica factor Q_Γ^{ab},

$$\Omega_K^{\text{raw}} = L_K\begin{pmatrix}1 & 0 \\ 0 & \rho(1-\rho)\mathbf{I}_K\end{pmatrix}L_K^\top, \tag{B.50}$$

which gives

$$\Omega_K^{\text{raw}} = \begin{pmatrix}(1-\rho)^2\mathbf{1}_K\mathbf{1}_K^\top + \rho(1-\rho)\mathbf{I}_K & \rho(1-\rho)(\mathbf{1}_K\mathbf{1}_K^\top - \mathbf{I}_K) \\ \rho(1-\rho)(\mathbf{1}_K\mathbf{1}_K^\top - \mathbf{I}_K) & \rho^2\mathbf{1}_K\mathbf{1}_K^\top + \rho(1-\rho)\mathbf{I}_K\end{pmatrix}. \tag{B.51}$$

For $K \geq 2$ this matrix is singular because

$$h_{1,\text{vis}} + h_{1,\text{mask}} = \cdots = h_{K,\text{vis}} + h_{K,\text{mask}} = h. \tag{B.52}$$

For $0 < \rho < 1$,

$$\operatorname{rank}(\Omega_K^{\text{raw}}) = K+1, \qquad \dim\ker(\Omega_K^{\text{raw}}) = K-1. \tag{B.53}$$

The singularity is therefore only a consequence of using $2K$ coordinates to represent a $(K+1)$-dimensional random object.

C REPLICA DERIVATION AND OUTPUT-CHANNEL SPECIALIZATIONS

C.1 GENERIC REPLICA DERIVATION

The computation is performed once for a generic finite-dimensional output channel. The model-specific information enters only through the field dimension L, the signal loading $\boldsymbol{\kappa}$, the covariance matrix Ω, and the output factor P_{out}. Finite-K and dynamic masking are specified after the generic saddle-point calculation.

C.1.1 HIGH-LEVEL DESCRIPTION

Let

$$\boldsymbol{f}(\boldsymbol{w}) = (f_1(\boldsymbol{w}), \dots, f_L(\boldsymbol{w})) \in \mathbb{R}^L, \qquad L = O_d(1), \tag{C.1}$$

be a finite collection of sample-dependent projected fields of the form

$$f_\ell(\boldsymbol{w}) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} r_{\ell i} w_i x_i, \qquad \ell = 1, \dots, L, \tag{C.2}$$

where $r_{\ell i}$ denotes the sample-disorder coefficient associated with component ℓ. We assume that, for every fixed number of replicas, the empirical mean and covariance coefficients self-average as

$$\frac{1}{d} \sum_{i=1}^{d} r_{\ell i} u_i^\star w_i^{\mathsf{a}} \longrightarrow \kappa_\ell M^{\mathsf{a}}, \tag{C.3}$$

$$\frac{1}{d} \sum_{i=1}^{d} r_{\ell i} r_{ki} \gamma_i^\star w_i^{\mathsf{a}} w_i^{\mathsf{b}} \longrightarrow \Omega_{\ell k} Q_\Gamma^{\mathsf{ab}}, \tag{C.4}$$

for a deterministic loading vector $\boldsymbol{\kappa} \in \mathbb{R}^L$ and a positive-semidefinite matrix $\Omega \in \mathbb{R}^{L \times L}$.

The generic sample contribution is encoded through

$$P_{\text{out}}\left(\boldsymbol{f}, Q_I, Q_\Gamma\right), \tag{C.5}$$

where the diagonal overlaps Q_I and Q_Γ are allowed to enter explicitly. We keep the weight factor

$$P_w(\boldsymbol{w}) = \exp\left\{-\eta_{\text{th}} \frac{\lambda_R}{2} \|\boldsymbol{w}\|_2^2\right\}. \tag{C.6}$$

Here $\eta_{\text{th}} > 0$ is the Gibbs inverse temperature. The corresponding partition function is

$$\mathcal{Z}(\mathcal{D}) = \int \mathrm{d}\boldsymbol{w}\, P_w(\boldsymbol{w}) \prod_{\mu=1}^{n} P_{\text{out}}(\boldsymbol{f}_\mu(\boldsymbol{w}), Q_I(\boldsymbol{w}), Q_\Gamma(\boldsymbol{w}))\,, \tag{C.7}$$

with

$$Q_I(\boldsymbol{w}) = \frac{1}{d} \boldsymbol{w}^\top \boldsymbol{w} = \frac{1}{d} \sum_i w_i^2, \qquad Q_\Gamma(\boldsymbol{w}) = \frac{1}{d} \boldsymbol{w}^\top \Gamma^\star \boldsymbol{w} = \frac{1}{d} \sum_i \gamma_i^\star w_i^2. \tag{C.8}$$

C.1.2 REPLICATED MOMENT AND OVERLAP PARAMETERS

The free entropy per dimension is

$$\Phi = \lim_{d \to \infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}(\mathcal{D}), \tag{C.9}$$

and the replica method formally uses

$$\mathbb{E}[\log \mathcal{Z}] = \lim_{s \to 0} \frac{\mathbb{E}[\mathcal{Z}^s] - 1}{s}. \tag{C.10}$$

For an integer $s \geq 1$, introduce replicas $\boldsymbol{w}^1, \dots, \boldsymbol{w}^s$ and the same overlap family as above,

$$M^{\mathsf{a}} := \frac{1}{d} (\boldsymbol{u}^\star)^\top \boldsymbol{w}^{\mathsf{a}}, \qquad Q_A^{\mathsf{ab}} := \frac{1}{d} (\boldsymbol{w}^{\mathsf{a}})^\top A \boldsymbol{w}^{\mathsf{b}}, \tag{C.11}$$

with the two specializations

$$Q_I^{\mathsf{ab}} = \frac{1}{d} (\boldsymbol{w}^{\mathsf{a}})^\top \boldsymbol{w}^{\mathsf{b}}, \qquad Q_\Gamma^{\mathsf{ab}} = \frac{1}{d} (\boldsymbol{w}^{\mathsf{a}})^\top \Gamma^\star \boldsymbol{w}^{\mathsf{b}}. \tag{C.12}$$

Then

$$\mathcal{Z}^s = \int \prod_{\mathsf{a}=1}^{s} \mathrm{d}\boldsymbol{w}^{\mathsf{a}}\, P_w(\boldsymbol{w}^{\mathsf{a}}) \prod_{\mu=1}^{n} \prod_{\mathsf{a}=1}^{s} P_{\text{out}}\left(\boldsymbol{f}_\mu^{\mathsf{a}}, Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}}\right). \tag{C.13}$$

The output factor of replica a depends explicitly only on the diagonal quantities Q_I^{aa} and Q_Γ^{aa}. The off-diagonal overlaps Q_Γ^{ab} enter through the joint Gaussian law because all replicas are evaluated on the same training sample and their projected data-noise fields are correlated. The off-diagonal Euclidean overlaps Q_I^{ab} do not enter the output covariance directly; they become relevant in the weight-side saddle-point constraints. Because $\gamma_i^\star$ is quenched, the primitive overlap variables are resolved by the variance type $\gamma \sim P_\gamma$. For each replica a, define the teacher overlap restricted to coordinates with variance γ

$$m^a(\gamma) = \frac{1}{d\,P_\gamma(\mathrm{d}\gamma)} \sum_{i:\gamma_i^\star \simeq \gamma} u_i^\star w_i^a, \tag{C.14}$$

and for each pair a, b define the corresponding type-resolved Euclidean overlap

$$Q^{ab}(\gamma) = \frac{1}{d\,P_\gamma(\mathrm{d}\gamma)} \sum_{i:\gamma_i^\star \simeq \gamma} w_i^a w_i^b. \tag{C.15}$$

This notation is understood by first discretizing the support of P_γ and then taking the continuum limit. The global overlaps are the two metric moments of this type-resolved overlap:

$$M^a = \mathbb{E}_\gamma[m^a(\gamma)], \qquad Q_I^{ab} = \mathbb{E}_\gamma[Q^{ab}(\gamma)], \qquad Q_\Gamma^{ab} = \mathbb{E}_\gamma[\gamma Q^{ab}(\gamma)]. \tag{C.16}$$

Thus $Q_I^{ab} = Q_{A=I}^{ab}$ is the Euclidean overlap, whereas $Q_\Gamma^{ab} = Q_{A=\Gamma^\star}^{ab}$ is the overlap seen by the heteroskedastic Gaussian noise. In the homoskedastic limit $\Gamma^\star = I$, the two matrices coincide.

Here $Q^{ab}(\gamma)$ is a type-resolved overlap density. It is not a third global metric overlap: Q_I^{ab} and Q_Γ^{ab} are the two moments of this same object that are required by the Euclidean decoder geometry and the heteroskedastic field covariance, respectively.

Because $u_i^\star$ and $\gamma_i^\star$ are independent and $\mathbb{E}[(u^\star)^2] = 1$, the teacher self-overlap at fixed variance type self-averages to

$$\frac{1}{d\,P_\gamma(\mathrm{d}\gamma)} \sum_{i:\gamma_i^\star \simeq \gamma} (u_i^\star)^2 \longrightarrow 1. \tag{C.17}$$

Hence the teacher self-overlap is fixed by the prior and does not introduce an additional variational order parameter.

Conditioning on the quenched coordinate variances $\boldsymbol{\gamma}^\star$, independence across the sample index μ gives

$$\mathbb{E}_{\mathcal{D}|\boldsymbol{\gamma}^\star} \mathcal{Z}^s = \mathbb{E}_{\boldsymbol{u}^\star} \int \prod_{\mathsf{a}=1}^{s} \mathrm{d}\boldsymbol{w}^{\mathsf{a}}\, P_w(\boldsymbol{w}^{\mathsf{a}}) \times \left[\mathbb{E}_{\lambda,\boldsymbol{z},\boldsymbol{r}} \prod_{\mathsf{a}=1}^{s} P_{\mathrm{out}}(\boldsymbol{f}^{\mathsf{a}}, Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}}) \right]^n \tag{C.18}$$

The expectation in square brackets is a one-sample quantity. Although it originates from a d-dimensional sample, the output factor depends on that sample only through the finite set of Ls replicated fields.

C.1.3 Gaussian reduction and replicated output factor

Using the data model (B.3), the field of replica a decomposes as

$$f_\ell^{\mathsf{a}} = \frac{\sqrt{\beta}\lambda}{d} \sum_{i=1}^{d} r_{\ell i} u_i^\star w_i^{\mathsf{a}} + \frac{1}{\sqrt{d}} \sum_{i=1}^{d} r_{\ell i} \sqrt{\gamma_i^\star}\, w_i^{\mathsf{a}} z_i. \tag{C.19}$$

The first term converges, by (C.3), to

$$\sqrt{\beta}\lambda\, \kappa_\ell M^{\mathsf{a}}. \tag{C.20}$$

The second term is Gaussian conditional on the coefficients $r_{\ell i}$ because the z_i are Gaussian. Its empirical covariance between components (ℓ, a) and (k, b) is

$$\frac{1}{d} \sum_{i=1}^{d} r_{\ell i} r_{ki} \gamma_i^\star w_i^{\mathsf{a}} w_i^{\mathsf{b}}, \tag{C.21}$$

which converges by (C.4) to

$$\Omega_{\ell k} Q_\Gamma^{\mathsf{ab}}. \tag{C.22}$$

Thus no additional central-limit approximation is needed for the Gaussian data-noise term; only its empirical covariance is replaced by its high-dimensional limit.

Let

$$\boldsymbol{M}_s = (M^1, \dots, M^s)^\top, \qquad \mathsf{Q}_{\Gamma,s} = (Q_\Gamma^{\mathsf{ab}})_{\mathsf{a},\mathsf{b}\le s}. \tag{C.23}$$

Then

$$(\boldsymbol{f}^1, \dots, \boldsymbol{f}^s) \mid \lambda \Longrightarrow \mathcal{N}_{Ls}\Big(\sqrt{\beta}\lambda(\boldsymbol{M}_s \otimes \boldsymbol{\kappa}), \mathsf{Q}_{\Gamma,s} \otimes \Omega\Big)\,. \tag{C.24}$$

If Ω is singular, this notation denotes the corresponding Gaussian probability measure on the support of its covariance matrix.

Define

$$\Psi_{\text{out}}^{(s)}\left(\boldsymbol{M}_s, \mathsf{Q}_{\Gamma,s}, \{Q_I^{\mathsf{aa}}\}, \{Q_\Gamma^{\mathsf{aa}}\}\right) := \mathbb{E}_\lambda \mathbb{E}_{(\boldsymbol{f}^1,\dots,\boldsymbol{f}^s)|\lambda}\left[\prod_{\mathsf{a}=1}^{s} P_{\text{out}}(\boldsymbol{f}^{\mathsf{a}}, Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}})\right]. \tag{C.25}$$

The inner expectation is with respect to (C.24). Equivalently,

$$\begin{aligned}\Psi_{\text{out}}^{(s)} = \mathbb{E}_\lambda \int \mathcal{N}_{Ls}\Big(\mathrm{d}\boldsymbol{f}^1 \cdots \mathrm{d}\boldsymbol{f}^s; \sqrt{\beta}\lambda(\boldsymbol{M}_s \otimes \boldsymbol{\kappa}), \mathsf{Q}_{\Gamma,s} \otimes \Omega\Big) \\ \times \prod_{\mathsf{a}=1}^{s} P_{\text{out}}(\boldsymbol{f}^{\mathsf{a}}, Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}})\,.\end{aligned} \tag{C.26}$$

Hence the replicated moment takes the generic form

$$\mathbb{E}_{\mathcal{D}|\boldsymbol{\gamma}^\star} \mathcal{Z}^s = \mathbb{E}_{\boldsymbol{u}^\star} \int \prod_{\mathsf{a}=1}^{s} \mathrm{d}\boldsymbol{w}^{\mathsf{a}}\, P_w(\boldsymbol{w}^{\mathsf{a}}) \left[\Psi_{\text{out}}^{(s)}\right]^n. \tag{C.27}$$

This is the point at which the model-specific sample disorder has been compressed into a finite-dimensional output contribution. The remainder of the replica saddle-point calculation depends on the learning problem only through $\Psi_{\text{out}}^{(s)}$.

C.1.4 Overlap constraints and replica action

The low-dimensional output factor depends on the microscopic replicas only through the type-resolved overlaps. We enforce their definitions with conjugate fields, which decouples the coordinate dependence in the usual way. We enforce the type-resolved student overlaps with their conjugate fields

$$\begin{aligned}1 = \int \mathcal{D}m\, \mathcal{D}Q\, \mathcal{D}\widehat{m}\, \mathcal{D}\widehat{Q} \exp\Bigg\{ d\mathbb{E}_\gamma \left[\sum_{a=1}^{s} \widehat{m}^a(\gamma) m^a(\gamma) + \sum_{a\le b} \widehat{Q}^{ab}(\gamma) Q^{ab}(\gamma)\right] \\ - \sum_{i=1}^{d}\left[\sum_{a=1}^{s} \widehat{m}^a(\gamma_i^\star) u_i^\star w_i^a + \sum_{a\le b} \widehat{Q}^{ab}(\gamma_i^\star) w_i^a w_i^b\right]\Bigg\}.\end{aligned} \tag{C.28}$$

Substituting this into Eq. (C.27) and factorizing, one can show that

$$\mathbb{E}_{\mathcal{D}|\boldsymbol{\gamma}^\star} \mathcal{Z}^s = \int \mathcal{D}m\, \mathcal{D}Q\, \mathcal{D}\widehat{m}\, \mathcal{D}\widehat{Q} \exp\left\{ d\left[\Phi_{\text{trace}}^{(s)} + \Phi_{\text{prior}}^{(s)} + \alpha \Phi_{\text{out}}^{(s)}\right]\right\}, \tag{C.29}$$

where

$$\Phi_{\text{trace}}^{(s)} = \mathbb{E}_\gamma \left[\sum_{a=1}^{s} \widehat{m}^a(\gamma) m^a(\gamma) + \sum_{a\le b} \widehat{Q}^{ab}(\gamma) Q^{ab}(\gamma)\right], \tag{C.30}$$

$$\Phi_{\text{prior}}^{(s)} = \mathbb{E}_\gamma \log \mathbb{E}_{u^\star \sim P_u} \int \prod_{a=1}^{s} \mathrm{d}w^a\, P_w(w^a) \exp\left\{ -\sum_{a=1}^{s} \widehat{m}^a(\gamma) u^\star w^a - \sum_{a\le b} \widehat{Q}^{ab}(\gamma) w^a w^b \right\}, \tag{C.31}$$

$$\Phi_{\text{out}}^{(s)} = \log \Psi_{\text{out}}^{(s)}\left(\boldsymbol{M}_s, \mathsf{Q}_{\Gamma,s}, \{Q_I^{aa}\}, \{Q_\Gamma^{aa}\}\right). \tag{C.32}$$

C.1.5 REPLICA-SYMMETRIC ANSATZ AND $s \to 0$ REDUCTION

We now impose the replica-symmetric (RS) ansatz

$$m^a(\gamma) = m(\gamma), \tag{C.33}$$

$$Q^{ab}(\gamma) = q(\gamma) + V(\gamma)\delta_{ab}, \qquad a, b = 1, \dots, s. \tag{C.34}$$

For the conjugated overlaps we use

$$\widehat{m}^a(\gamma) = -\widehat{m}(\gamma), \tag{C.35}$$

$$\widehat{Q}^{ab}(\gamma) = \left(\frac{\widehat{V}(\gamma) - \widehat{q}(\gamma)}{2} + \widehat{q}(\gamma) \right) \delta_{ab} - \widehat{q}(\gamma), \tag{C.36}$$

Under the RS assumption, we have that

$$\begin{aligned} &- \sum_{a=1}^{s} \widehat{m}^a(\gamma) u^\star w^a - \sum_{a \le b} \widehat{Q}^{ab}(\gamma) w^a w^b \\ &= \widehat{m}(\gamma) u^\star \sum_{a=1}^{s} w^a - \frac{\widehat{V}(\gamma)}{2} \sum_{a=1}^{s} (w^a)^2 + \frac{\widehat{q}(\gamma)}{2} \left(\sum_{a=1}^{s} w^a \right)^2 . \end{aligned} \tag{C.37}$$

and a few additional order parameters

$$\begin{aligned} M &:= \mathbb{E}_\gamma[m(\gamma)], \\ q_I &:= \mathbb{E}_\gamma[q(\gamma)], \qquad & q_\Gamma &:= \mathbb{E}_\gamma[\gamma q(\gamma)], \\ V_I &:= \mathbb{E}_\gamma[V(\gamma)], \qquad & V_\Gamma &:= \mathbb{E}_\gamma[\gamma V(\gamma)]. \end{aligned} \tag{C.38}$$

Consequently, the two global replica-overlap matrices are

$$Q_I^{ab} = q_I + V_I \delta_{ab}, \qquad Q_\Gamma^{ab} = q_\Gamma + V_\Gamma \delta_{ab}. \tag{C.39}$$

Trace term. For the type-resolved quadratic overlap

$$\begin{aligned} \sum_{a \le b} \widehat{Q}^{ab} Q^{ab} &= s \, \frac{\widehat{V} - \widehat{q}}{2} (q + V) - \frac{s(s-1)}{2} \widehat{q}\, q \\ &= \frac{s}{2} \left[\widehat{V} q + (\widehat{V} - \widehat{q}) V \right] - \frac{s^2}{2} \widehat{q}\, q. \end{aligned} \tag{C.40}$$

The trace term can be rewritten as

$$\Phi_{\text{trace}}^{(s)} = s \left\{ -\mathbb{E}_\gamma[\widehat{m}(\gamma) m(\gamma)] + \frac{1}{2} \mathbb{E}_\gamma[\widehat{V}(\gamma) q(\gamma)] + \frac{1}{2} \mathbb{E}_\gamma[(\widehat{V}(\gamma) - \widehat{q}(\gamma)) V(\gamma)] \right\} + O(s^2). \tag{C.41}$$

Prior term. Substituting the RS ansatz into Eq. (C.31), the only term coupling replicas is quadratic in $\sum_a w^a$. For each fixed γ,

$$\exp\left\{ \frac{\widehat{q}(\gamma)}{2} \left(\sum_{a=1}^{s} w^a \right)^2 \right\} = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \exp\left\{ \sqrt{\widehat{q}(\gamma)} \xi \sum_{a=1}^{s} w^a \right\}. \tag{C.42}$$

Define the scalar input partition function

$$\mathcal{Z}_{\text{in}}(\gamma, u^\star, \xi) := \int_{\mathbb{R}} \mathrm{d}w \, P_w(w) \exp\left\{ \sqrt{\widehat{q}(\gamma)}\, \xi w - \frac{\widehat{V}(\gamma)}{2} w^2 + \widehat{m}(\gamma) u^\star w \right\}. \tag{C.43}$$

Using Eq. (C.42), the finite-s site partition becomes exactly

$$\mathcal{Z}_{\text{in}}^{(s)}(\gamma) = \mathbb{E}_{u^\star, \xi} \left[\mathcal{Z}_{\text{in}}(\gamma, u^\star, \xi) \right]^s . \tag{C.44}$$

Therefore

$$\Phi_{\text{prior}}^{(s)} = \mathbb{E}_\gamma \log \mathbb{E}_{u^\star, \xi} \exp\left\{ s \log \mathcal{Z}_{\text{in}}(\gamma, u^\star, \xi) \right\}, \tag{C.45}$$

expanding around $s = 0$ gives

$$\log \mathbb{E} e^{sX} = s \mathbb{E} X + O(s^2). \tag{C.46}$$

Applying this with $X = \log \mathcal{Z}_{\text{in}}$ gives

$$\Phi_{\text{prior}}^{(s)} = s \, \mathbb{E}_{\gamma, u^\star, \xi} \log \int \mathrm{d}w \, P_w(w) \exp\left\{ \sqrt{\widehat{q}(\gamma)} \xi w - \frac{1}{2} \widehat{V}(\gamma) w^2 + \widehat{m}(\gamma) u^\star w \right\} + O(s^2). \tag{C.47}$$

Output term. Under Eq. (C.39), the replicated field covariance is

$$(q_\Gamma \mathbf{1}_s \mathbf{1}_s^\top + V_\Gamma \mathbf{I}_s) \otimes \Omega. \tag{C.48}$$

It can be generated by one common Gaussian vector $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I}_L)$ and independent replica-specific vectors $\boldsymbol{z}^a \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_L)$:

$$\boldsymbol{f}^a = \sqrt{\beta}\lambda \boldsymbol{\kappa} M + \sqrt{q_\Gamma}\, \Omega^{1/2} \boldsymbol{z} + \sqrt{V_\Gamma}\, \Omega^{1/2} \boldsymbol{z}^a. \tag{C.49}$$

Define the common field

$$\boldsymbol{f}_0 := \sqrt{\beta}\lambda \boldsymbol{\kappa} M + \sqrt{q_\Gamma}\, \Omega^{1/2} \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I}_L). \tag{C.50}$$

Conditionally on $(\lambda, \boldsymbol{z})$, the replicas are independent and

$$\boldsymbol{f}^a \mid \lambda, \boldsymbol{z} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{f}_0, V_\Gamma \Omega). \tag{C.51}$$

The diagonal norm arguments are

$$Q_I^{aa} = q_I + V_I, \qquad Q_\Gamma^{aa} = q_\Gamma + V_\Gamma. \tag{C.52}$$

Define the scalar output partition function

$$\mathcal{Z}_{\text{out}}(\boldsymbol{f}_0) := \int \mathrm{d}\boldsymbol{f}\, \mathcal{N}(\boldsymbol{f}; \boldsymbol{f}_0, V_\Gamma \Omega) P_{\text{out}}(\boldsymbol{f}, q_I + V_I, q_\Gamma + V_\Gamma). \tag{C.53}$$

Then the replicated output factor is exactly

$$\Psi_{\text{out}}^{(s)} = \mathbb{E}_{\lambda, \boldsymbol{z}} \left[\mathcal{Z}_{\text{out}}(\boldsymbol{f}_0) \right]^s, \tag{C.54}$$

so that the same cumulant expansion gives

$$\Phi_{\text{out}}^{(s)} = s\, \mathbb{E}_{\lambda, \boldsymbol{z}} \log \int \mathrm{d}\boldsymbol{f}\, \mathcal{N}(\boldsymbol{f}; \boldsymbol{f}_0, V_\Gamma \Omega) P_{\text{out}}(\boldsymbol{f}, q_I + V_I, q_\Gamma + V_\Gamma) + O(s^2). \tag{C.55}$$

C.1.6 Replica-symmetric free entropy

Using the analytic continuation around $s = 0$, the replica trick can equivalently be written as

$$\Phi = \lim_{s \to 0} \frac{1}{s} \lim_{d \to \infty} \frac{1}{d} \log \mathbb{E}_{\mathcal{D}}[\mathcal{Z}^s], \tag{C.56}$$

where the equality is understood through the analytic continuation around $s = 0$.

At fixed s, the large-d saddle gives

$$\lim_{d \to \infty} \frac{1}{d} \log \mathbb{E}_{\mathcal{D}}[\mathcal{Z}^s] = \underset{m, q, V, \widehat{m}, \widehat{q}, \widehat{V}}{\text{extr}} \left\{ \Phi_{\text{trace}}^{(s)} + \Phi_{\text{prior}}^{(s)} + \alpha \Phi_{\text{out}}^{(s)} \right\}. \tag{C.57}$$

Here the extremization is over the γ-resolved functions $m(\gamma), q(\gamma), V(\gamma)$ and their conjugates $\widehat{m}(\gamma), \widehat{q}(\gamma), \widehat{V}(\gamma)$.

Equations (C.41), (C.47), and (C.55) have the common small-s structure

$$\Phi_X^{(s)} = s\, \Phi_X + O(s^2), \qquad X \in \{\text{trace}, \text{prior}, \text{out}\}. \tag{C.58}$$

Therefore, dividing Eq. (C.57) by s and taking $s \to 0$ yields a single replica-symmetric saddle for the quenched free entropy,

$$\Phi = \underset{m(\cdot), q(\cdot), V(\cdot), \widehat{m}(\cdot), \widehat{q}(\cdot), \widehat{V}(\cdot)}{\text{extr}} \left\{ \Phi_{\text{trace}} + \Phi_{\text{prior}} + \alpha \Phi_{\text{out}} \right\} \tag{C.59}$$

The three $s \to 0$ contributions are

$$\Phi_{\text{trace}} = -\mathbb{E}_\gamma[\widehat{m}(\gamma) m(\gamma)] + \frac{1}{2} \mathbb{E}_\gamma[\widehat{V}(\gamma) q(\gamma)] \quad + \frac{1}{2} \mathbb{E}_\gamma[(\widehat{V}(\gamma) - \widehat{q}(\gamma)) V(\gamma)]. \tag{C.60}$$

For the input contribution,

$$\Phi_{\text{prior}} = \mathbb{E}_{\gamma, u^\star, \xi} \log \mathcal{Z}_{\text{in}}(\gamma, u^\star, \xi), \tag{C.61}$$

with $\mathcal{Z}_{\rm in}$ defined in Eq. (C.43). For the output contribution,

$$\Phi_{\rm out} = \mathbb{E}_{\lambda,\boldsymbol{z}} \log \mathcal{Z}_{\rm out}(\boldsymbol{f}_0), \tag{C.62}$$

where

$$\mathcal{Z}_{\rm out}(\boldsymbol{f}_0) = \int \mathrm{d}\boldsymbol{f}\, \mathcal{N}(\boldsymbol{f}; \boldsymbol{f}_0, V_\Gamma \Omega) P_{\rm out}(\boldsymbol{f}, q_I + V_I, q_\Gamma + V_\Gamma), \tag{C.63}$$

$$\boldsymbol{f}_0 = \sqrt{\beta}\lambda \boldsymbol{\kappa} M + \sqrt{q_\Gamma}\, \Omega^{1/2} \boldsymbol{z}. \tag{C.64}$$

The global metric moments entering the output channel remain

$$q_I = \mathbb{E}_\gamma[q(\gamma)], \qquad q_\Gamma = \mathbb{E}_\gamma[\gamma q(\gamma)], \qquad V_I = \mathbb{E}_\gamma[V(\gamma)], \qquad V_\Gamma = \mathbb{E}_\gamma[\gamma V(\gamma)]. \tag{C.65}$$

The replicated problem has therefore reduced to a single RS variational saddle. The zero-temperature ERM limit is taken only after this $s \to 0$ reduction.

C.1.7 Zero-temperature ERM limit

We will specialize our equations for ridge regularization:

$$P_w(w) = \exp\left\{-\eta_{\rm th}\frac{\lambda_R}{2}w^2\right\}, \qquad P_{\rm out}(\boldsymbol{f}, Q_I, Q_\Gamma) = \exp\{-\eta_{\rm th}\ell_{\rm out}(\boldsymbol{f}; Q_I, Q_\Gamma)\}. \tag{C.66}$$

We rescale the replica-diagonal excess overlap as

$$V(\gamma) := \frac{v(\gamma)}{\eta_{\rm th}}, \qquad V_I = \frac{v_I}{\eta_{\rm th}}, \quad v_I := \mathbb{E}_\gamma[v(\gamma)], \qquad V_\Gamma = \frac{v_\Gamma}{\eta_{\rm th}}, \quad v_\Gamma := \mathbb{E}_\gamma[\gamma v(\gamma)]. \tag{C.67}$$

$$\widehat{m}(\gamma) \mapsto \eta_{\rm th}\widehat{m}(\gamma), \qquad \widehat{V}(\gamma) \mapsto \eta_{\rm th}\widehat{v}(\gamma), \qquad \widehat{q}(\gamma) \mapsto \eta_{\rm th}^2\widehat{q}(\gamma). \tag{C.68}$$

Input proximal. Using Laplace's method in Eq. (C.47), the prior term contribution at leading order is

$$\Phi^0_{\rm prior} = -\mathbb{E}_{\gamma,u^\star,\xi} \min_{w\in\mathbb{R}} \left\{ \frac{\lambda_R + \widehat{v}(\gamma)}{2} w^2 - \left[\widehat{m}(\gamma)u^\star + \sqrt{\widehat{q}(\gamma)}\xi\right] w \right\}. \tag{C.69}$$

Defining the coordinate proximal

$$w_\star(\gamma, u^\star, \xi) = \arg\min_{w\in\mathbb{R}} \left\{ \frac{\lambda_R + \widehat{v}(\gamma)}{2} w^2 - \left[\widehat{m}(\gamma)u^\star + \sqrt{\widehat{q}(\gamma)}\xi\right] w \right\}. \tag{C.70}$$

For ridge regularization, this gives

$$w_\star(\gamma, u^\star, \xi) = \frac{\widehat{m}(\gamma)u^\star + \sqrt{\widehat{q}(\gamma)}\xi}{\lambda_R + \widehat{v}(\gamma)} \tag{C.71}$$

Stationarity with respect to the conjugates gives the first saddle point equations

$$m(\gamma) = \mathbb{E}_{u^\star,\xi}[u^\star w_\star(\gamma, u^\star, \xi)], \qquad q(\gamma) = \mathbb{E}_{u^\star,\xi}[w_\star(\gamma, u^\star, \xi)^2], \tag{C.72}$$

$$v(\gamma) = \mathbb{E}_{u^\star,\xi}\left[\frac{\xi w_\star(\gamma, u^\star, \xi)}{\sqrt{\widehat{q}(\gamma)}}\right]. \tag{C.73}$$

For Eq. (C.71), this reduces to

$$m(\gamma) = \frac{\widehat{m}(\gamma)}{\lambda_R + \widehat{v}(\gamma)}, \qquad q(\gamma) = \frac{\widehat{m}(\gamma)^2 + \widehat{q}(\gamma)}{(\lambda_R + \widehat{v}(\gamma))^2}, \qquad v(\gamma) = \frac{1}{\lambda_R + \widehat{v}(\gamma)} \tag{C.74}$$

Output proximal. For the zero-temperature output analysis we assume that

$$\Omega \succ 0. \tag{C.75}$$

This is satisfied by all output channels considered in the present work. In particular, for finite-K masking,

$$\Omega_K = \mathrm{diag}(1, \rho(1-\rho)\mathbf{I}_K) \succ 0, \qquad 0 < \rho < 1, \tag{C.76}$$

while for dynamic masking $\Omega = 1$. Hence Ω^{-1} is well defined in the Moreau-envelope representation below.

The output term Eq. (C.55) becomes, after using $V_\Gamma = v_\Gamma/\eta_{\rm th}$,

$$\int \mathrm{d}\boldsymbol{f}\, \mathcal{N}\left(\boldsymbol{f}; \boldsymbol{f}_0, \frac{v_\Gamma}{\eta_{\rm th}}\Omega\right) \exp\{-\eta_{\rm th}\ell_{\rm out}(\boldsymbol{f}; q_I, q_\Gamma)\}. \tag{C.77}$$

Laplace's method gives the standard Moreau envelope

$$\mathcal{M}_{\rm out}(\boldsymbol{f}_0; q_I, q_\Gamma, v_\Gamma) = \min_{\boldsymbol{f}\in\mathbb{R}^L} \left\{ \ell_{\rm out}(\boldsymbol{f}; q_I, q_\Gamma) + \frac{1}{2v_\Gamma}(\boldsymbol{f}-\boldsymbol{f}_0)^\top \Omega^{-1}(\boldsymbol{f}-\boldsymbol{f}_0) \right\}. \tag{C.78}$$

Let $\boldsymbol{f}_\star$ be any minimizer. In the $\eta_{\rm th} \to \infty$ limit we have that

$$\Phi^0_{\rm out} = -\mathbb{E}_{\lambda,\boldsymbol{z}}\left[\mathcal{M}_{\rm out}(\boldsymbol{f}_0; q_I, q_\Gamma, v_\Gamma)\right], \tag{C.79}$$

with $\boldsymbol{f}_0$ given by Eq. (C.50). Define

$$\boldsymbol{F} = \frac{1}{v_\Gamma}\Omega^{-1}(\boldsymbol{f}_\star - \boldsymbol{f}_0). \tag{C.80}$$

hence

$$\boldsymbol{F} + \nabla_{\boldsymbol{f}}\ell_{\rm out}(\boldsymbol{f}_\star; q_I, q_\Gamma) = 0. \tag{C.81}$$

Since $\partial_M \boldsymbol{f}_0 = \sqrt{\beta}\lambda\boldsymbol{\kappa}$, we have that

$$A_\star = \boldsymbol{\kappa}^\top \boldsymbol{F}. \tag{C.82}$$

Moreover,

$$\frac{\partial \mathcal{M}_{\rm out}}{\partial v_\Gamma} = -\frac{1}{2}\boldsymbol{F}^\top \Omega \boldsymbol{F}. \tag{C.83}$$

For the derivative with respect to the noise-weighted off-diagonal overlap q_Γ, define

$$D_\Gamma(\boldsymbol{f}_\star) = \frac{\partial \mathcal{M}_{\rm out}}{\partial q_\Gamma} = \partial_{q_\Gamma}\ell_{\rm out}(\boldsymbol{f}_\star; q_I, q_\Gamma) - \boldsymbol{F}^\top \partial_{q_\Gamma}\boldsymbol{f}_0, \tag{C.84}$$

where

$$\partial_{q_\Gamma}\boldsymbol{f}_0 = \frac{1}{2\sqrt{q_\Gamma}}\Omega^{1/2}\boldsymbol{z}. \tag{C.85}$$

Saddle-point equations. At order $\eta_{\rm th}$, the RS extremization gives

$$\Phi_0 = \operatorname*{extr}_{m,q,v,\widehat{m},\widehat{q},\widehat{v}} \left\{ \Phi^0_{\rm trace} + \Phi^0_{\rm prior} + \alpha\Phi^0_{\rm out} \right\}, \tag{C.86}$$

where

$$\Phi^0_{\rm trace} = -\mathbb{E}_\gamma[\widehat{m}(\gamma)m(\gamma)] - \frac{1}{2}\mathbb{E}_\gamma[\widehat{q}(\gamma)v(\gamma)] + \frac{1}{2}\mathbb{E}_\gamma[\widehat{v}(\gamma)q(\gamma)]. \tag{C.87}$$

with

$$\widehat{m}(\gamma) = \alpha\sqrt{\beta}\,\mathbb{E}_{\lambda,\boldsymbol{z}}[\lambda A_\star] \tag{C.88}$$

The conjugate overlap associated with the replica-diagonal excess follows

$$\widehat{q}(\gamma) = \alpha\gamma\,\mathbb{E}_{\lambda,\boldsymbol{z}}[\boldsymbol{F}^\top\Omega\boldsymbol{F}] \tag{C.89}$$

Because $q(\gamma)$ enters the output channel through both $q_I = \mathbb{E}_\gamma[q(\gamma)]$ and $q_\Gamma = \mathbb{E}_\gamma[\gamma q(\gamma)]$, we obtain

$$\widehat{v}(\gamma) = 2\alpha\,\mathbb{E}_{\lambda,\boldsymbol{z}}\left[\partial_{q_I}\ell_{\rm out}(\boldsymbol{f}_\star; q_I, q_\Gamma) + \gamma D_\Gamma(\boldsymbol{f}_\star)\right] \tag{C.90}$$

Equations (C.72)–(C.73) and (C.88)–(C.90), together with the metric moments, give

$$\begin{aligned} M &= \mathbb{E}_\gamma[m(\gamma)], \\ q_I &= \mathbb{E}_\gamma[q(\gamma)], \quad q_\Gamma = \mathbb{E}_\gamma[\gamma q(\gamma)], \\ v_I &= \mathbb{E}_\gamma[v(\gamma)], \quad v_\Gamma = \mathbb{E}_\gamma[\gamma v(\gamma)]. \end{aligned} \tag{C.91}$$

C.2 Output-channel specializations

The generic computation of App. C.1 applies to both masking schemes. The weight-side saddle equations are common to the two models; only the output channel and the corresponding conjugate equations need to be specialized.

C.2.1 Finite-K specialization

For the finite-K masked autoencoder,

$$L = K + 1, \qquad \boldsymbol{f} = \boldsymbol{f}_K = (h, \delta_1, \dots, \delta_K), \qquad \boldsymbol{\kappa} = \boldsymbol{e}_0, \tag{C.92}$$

$$\Omega = \Omega_K = \mathrm{diag}(1, \rho(1-\rho)\mathbf{I}_K)\,. \tag{C.93}$$

The output factor is

$$P_{\mathrm{out},K}(\boldsymbol{f}_K; Q_I) = \exp\{-\eta_{\mathrm{th}} \ell_K(\boldsymbol{f}_K; Q_I)\}\,, \tag{C.94}$$

where ℓ_K is given in (B.33). It has no additional explicit dependence on Q_Γ; Q_Γ enters through the Gaussian law of the fields.

Substituting (C.92)–(C.94) into the generic replicated output factor gives

$$\Psi^{(s)}_{K,\mathrm{out}} := \mathbb{E}_\lambda \mathbb{E}_{(\boldsymbol{f}_K^1, \dots, \boldsymbol{f}_K^s)|\lambda} \left[\prod_{\mathsf{a}=1}^{s} \exp\{-\eta_{\mathrm{th}} \ell_K(\boldsymbol{f}_K^{\mathsf{a}}; Q_I^{\mathsf{aa}})\} \right], \tag{C.95}$$

where

$$(\boldsymbol{f}_K^1, \dots, \boldsymbol{f}_K^s) \mid \lambda \Longrightarrow \mathcal{N}_{(K+1)s}\Big(\sqrt{\beta}\lambda(\boldsymbol{M}_s \otimes \boldsymbol{e}_0), \mathsf{Q}_{\Gamma,s} \otimes \Omega_K \Big)\,. \tag{C.96}$$

Saddle-point specialization. Let $\boldsymbol{f}_\star$ denote the finite-K output proximal associated with ℓ_K through the generic Moreau problem (C.78), and let $\boldsymbol{F}_K$ denote the corresponding force defined in (C.80). Since the signal is loaded only along the full-field direction,

$$A_\star = \boldsymbol{e}_0^\top \boldsymbol{F}_K. \tag{C.97}$$

Moreover,

$$\partial_{q_I} \ell_K(\boldsymbol{f}_\star; q_I) = \frac{\rho}{K} \sum_{r=1}^{K} \sigma((1-\rho)h_\star + \delta_{r,\star})^2\,. \tag{C.98}$$

Since ℓ_K has no explicit dependence on q_Γ, (C.84) reduces to

$$D_\Gamma(\boldsymbol{f}_\star) = -\frac{1}{2\sqrt{q_\Gamma}} \boldsymbol{F}_K^\top \Omega_K^{1/2} \boldsymbol{z}. \tag{C.99}$$

The conjugate saddle equations (C.88)–(C.90) therefore become

$$\widehat{m}(\gamma) = \alpha\sqrt{\beta}\, \mathbb{E}_{\lambda,\boldsymbol{z}} \left[\lambda\, \boldsymbol{e}_0^\top \boldsymbol{F}_K \right], \tag{C.100}$$

$$\widehat{q}(\gamma) = \alpha\gamma\, \mathbb{E}_{\lambda,\boldsymbol{z}} \left[\boldsymbol{F}_K^\top \Omega_K \boldsymbol{F}_K \right], \tag{C.101}$$

$$\widehat{v}(\gamma) = 2\alpha\, \mathbb{E}_{\lambda,\boldsymbol{z}} \left[\frac{\rho}{K} \sum_{r=1}^{K} \sigma((1-\rho)h_\star + \delta_{r,\star})^2 - \frac{\gamma}{2\sqrt{q_\Gamma}} \boldsymbol{F}_K^\top \Omega_K^{1/2} \boldsymbol{z} \right]. \tag{C.102}$$

Together with the input-side equations (C.74) and the global closures (C.91), these equations close the finite-K saddle.

Static masking is the special case $K = 1$.

C.2.2 Dynamic-mask specialization

After the mask average has been taken, the dynamic-mask loss (B.47) depends explicitly only on the full preactivation h. The generic channel therefore reduces to

$$L = 1, \qquad \boldsymbol{f} = h, \qquad \boldsymbol{\kappa} = 1, \qquad \Omega = 1. \tag{C.103}$$

The corresponding output factor is

$$P_{\mathrm{dyn,out}}(h; Q_I, Q_\Gamma) = \exp\{-\eta_{\mathrm{th}} \ell_{\mathrm{dyn}}(h; Q_I, Q_\Gamma)\}\,. \tag{C.104}$$

Thus the generic replicated output factor becomes

$$\Psi^{(s)}_{\mathrm{dyn,out}} = \mathbb{E}_\lambda \mathbb{E}_{\boldsymbol{h}_s|\lambda}\left[\prod_{\mathsf{a}=1}^{s} \exp\{-\eta_{\mathrm{th}} \ell_{\mathrm{dyn}}(h^{\mathsf{a}}; Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}})\}\right], \tag{C.105}$$

where

$$\boldsymbol{h}_s \mid \lambda \sim \mathcal{N}_s\Big(\sqrt{\beta}\lambda \boldsymbol{M}_s, \mathsf{Q}_{\Gamma,s}\Big)\,. \tag{C.106}$$

Saddle-point specialization. Let $h_\star$ denote the scalar output proximal associated with ℓ_{dyn} through (C.78), and let F be the corresponding scalar force. Since $\boldsymbol{\kappa} = 1$,

$$A_\star = F. \tag{C.107}$$

The explicit derivative with respect to the Euclidean overlap is

$$\partial_{q_I} \ell_{\mathrm{dyn}}(h; q_I, q_\Gamma) = \rho \mathbb{E}_\xi\left[\sigma\Big((1-\rho)h + \sqrt{\rho(1-\rho)q_\Gamma}\,\xi\Big)^2\right]. \tag{C.108}$$

Unlike the finite-K loss, ℓ_{dyn} depends explicitly on q_Γ. Hence

$$D_\Gamma(h_\star) = \partial_{q_\Gamma} \ell_{\mathrm{dyn}}(h_\star; q_I, q_\Gamma) - \frac{Fz}{2\sqrt{q_\Gamma}}, \tag{C.109}$$

where

$$\partial_{q_\Gamma} \ell_{\mathrm{dyn}}(h; q_I, q_\Gamma) = \sqrt{\frac{\rho(1-\rho)}{q_\Gamma}}\,\mathbb{E}_\xi\left[\xi\left(\rho q_I \sigma(y)\sigma'(y) - t\,\sigma'(y) + \sigma(y)\right)\right], \tag{C.110}$$

with, only within this equation,

$$y = (1-\rho)h + \sqrt{\rho(1-\rho)q_\Gamma}\,\xi, \qquad t = \rho h - \sqrt{\rho(1-\rho)q_\Gamma}\,\xi. \tag{C.111}$$

The conjugate saddle equations are therefore

$$\widehat{m}(\gamma) = \alpha\sqrt{\beta}\,\mathbb{E}_{\lambda,z}\left[\lambda F\right], \tag{C.112}$$

$$\widehat{q}(\gamma) = \alpha\gamma\,\mathbb{E}_{\lambda,z}\left[F^2\right], \tag{C.113}$$

$$\widehat{v}(\gamma) = 2\alpha\,\mathbb{E}_{\lambda,z}\left[\partial_{q_I}\ell_{\mathrm{dyn}}(h_\star; q_I, q_\Gamma) + \gamma\left(\partial_{q_\Gamma}\ell_{\mathrm{dyn}}(h_\star; q_I, q_\Gamma) - \frac{Fz}{2\sqrt{q_\Gamma}}\right)\right]. \tag{C.114}$$

Together with the same input-side equations (C.74) and global closures (C.91), these equations close the dynamic-mask saddle.

The expressions containing $q_\Gamma^{-1/2}$ are understood through their continuous limit when $q_\Gamma \downarrow 0$.

C.2.3 The limit $K \to \infty$ at the replicated output level

It remains to connect the two specializations. For fixed positive integer s, write

$$\boldsymbol{h}_s = (h^1, \ldots, h^s)^\top, \qquad \boldsymbol{\delta}_r = (\delta_r^1, \ldots, \delta_r^s)^\top. \tag{C.115}$$

Under the finite-K Gaussian channel,

$$\boldsymbol{h}_s \mid \lambda \sim \mathcal{N}_s\Big(\sqrt{\beta}\lambda \boldsymbol{M}_s, \mathsf{Q}_{\Gamma,s}\Big)\,, \tag{C.116}$$

while

$$\boldsymbol{\delta}_r \overset{\mathrm{iid}}{\sim} \mathcal{N}_s(\mathbf{0}, \rho(1-\rho)\mathsf{Q}_{\Gamma,s})\,, \qquad r = 1, \ldots, K, \tag{C.117}$$

and the $\boldsymbol{\delta}_r$ are independent of $\boldsymbol{h}_s$. The components of a fixed $\boldsymbol{\delta}_r$ are generally correlated across replicas, but the fluctuation vectors are i.i.d. across the mask index r.

Define

$$q_{\mathsf{a}}(h,\delta) := \rho Q_I^{\mathsf{aa}}\sigma((1-\rho)h+\delta)^2 - 2\sigma((1-\rho)h+\delta)\,(\rho h-\delta). \tag{C.118}$$

Then

$$\prod_{\mathsf{a}=1}^{s} P_{\mathrm{out},K}(\boldsymbol{f}_K^{\mathsf{a}}; Q_I^{\mathsf{aa}}) = \exp\left\{-\eta_{\mathrm{th}}\frac{1}{K}\sum_{r=1}^{K}\sum_{\mathsf{a}=1}^{s} q_{\mathsf{a}}(h^{\mathsf{a}},\delta_r^{\mathsf{a}})\right\}. \tag{C.119}$$

Conditioning on $\boldsymbol{h}_s$ and applying the law of large numbers across the independent mask index gives

$$\frac{1}{K}\sum_{r=1}^{K}\sum_{\mathsf{a}=1}^{s} q_{\mathsf{a}}(h^{\mathsf{a}},\delta_r^{\mathsf{a}}) \xrightarrow[K\to\infty]{} \sum_{\mathsf{a}=1}^{s}\mathbb{E}_{\delta^{\mathsf{a}}} q_{\mathsf{a}}(h^{\mathsf{a}},\delta^{\mathsf{a}})\,. \tag{C.120}$$

Only the marginal distribution of each replica component appears on the right-hand side. From (C.117),

$$\delta^{\mathsf{a}} \sim \mathcal{N}(0,\rho(1-\rho)Q_\Gamma^{\mathsf{aa}})\,, \tag{C.121}$$

and therefore

$$\mathbb{E}_{\delta^{\mathsf{a}}} q_{\mathsf{a}}(h^{\mathsf{a}},\delta^{\mathsf{a}}) = \ell_{\mathrm{dyn}}(h^{\mathsf{a}}; Q_I^{\mathsf{aa}}, Q_\Gamma^{\mathsf{aa}})\,. \tag{C.122}$$

Consequently, under the usual integrability conditions that allow the limit to pass through the remaining Gaussian expectation,

$$\Psi_{K,\mathrm{out}}^{(s)} \xrightarrow[K\to\infty]{} \Psi_{\mathrm{dyn,out}}^{(s)}. \tag{C.123}$$

This establishes the dynamic-mask limit directly inside the same generic replica framework. The field dimension does not need to be sent formally from $K+1$ to infinity inside the generic derivation. Rather, the derivation is applied at each fixed K, and the resulting finite-K output functional converges to the $L=1$ dynamic output functional.

The sequential interpretation is therefore

$$\begin{gathered} d,n\to\infty \quad \text{at fixed } K,\\ \text{generic replica reduction with } L=K+1,\\ K\to\infty: \quad \Psi_{K,\mathrm{out}}^{(s)} \longrightarrow \Psi_{\mathrm{dyn,out}}^{(s)},\\ \text{followed by the usual replica continuation } s\to 0. \end{gathered} \tag{C.124}$$

Since the weight prior and the definitions of M, Q_I, and Q_Γ are unchanged, the remainder of the saddle-point calculation is common to both finite-K and dynamic masking; only the model-specific output functional changes.

C.2.4 Full-reconstruction autoencoder

We derive the output channel for the full-reconstruction objective (5), in which both the encoder input and the reconstruction targets include all coordinates. For one sample, the reconstruction loss is

$$\sum_{i=1}^{d}\left[x_i - \frac{w_i}{\sqrt{d}}\sigma(h)\right]^2, \qquad h = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} w_i x_i. \tag{C.125}$$

Dropping the term independent of $\boldsymbol{w}$ gives the effective output loss

$$\ell_{\mathrm{full}}(h; Q_I) = Q_I\sigma(h)^2 - 2h\sigma(h). \tag{C.126}$$

This is the full-reconstruction counterpart of the masked output loss (B.23): the encoder and target fields are both h, and the squared decoder norm is Q_I rather than the masked fraction ρQ_I. In particular, the cross term is $-2h\sigma(h)$ instead of $-2h_{r,\mathrm{mask}}\sigma(h_{r,\mathrm{vis}})$, making explicit the difference between reconstructing a field from itself and predicting its hidden part from its visible part.

The generic output channel therefore specializes to

$$L=1, \qquad \boldsymbol{f}=h, \qquad \boldsymbol{\kappa}=1, \qquad \Omega=1. \tag{C.127}$$

The corresponding output factor is

$$P_{\mathrm{full,out}}(h; Q_I) = \exp\{-\eta_{\mathrm{th}}\ell_{\mathrm{full}}(h; Q_I)\}\,. \tag{C.128}$$

Hence the replicated output contribution is

$$\Psi^{(s)}_{\text{full,out}} := \mathbb{E}_\lambda \mathbb{E}_{\boldsymbol{h}_s|\lambda}\left[\prod_{\mathsf{a}=1}^{s} \exp\{-\eta_{\text{th}}\ell_{\text{full}}(h^{\mathsf{a}}; Q_I^{\mathsf{aa}})\}\right], \tag{C.129}$$

where

$$\boldsymbol{h}_s \mid \lambda \sim \mathcal{N}_s\Big(\sqrt{\beta}\lambda \boldsymbol{M}_s, \mathsf{Q}_{\Gamma,s}\Big). \tag{C.130}$$

Saddle-point specialization. Let $h_\star$ denote the scalar output proximal obtained from the generic Moreau problem (C.78), and let F denote the corresponding scalar force defined in (C.80). Since $\boldsymbol{\kappa} = 1$,

$$A_\star = F. \tag{C.131}$$

The first-order condition of the proximal is

$$F + 2q_I\sigma(h_\star)\sigma'(h_\star) - 2\sigma(h_\star) - 2h_\star\sigma'(h_\star) = 0. \tag{C.132}$$

The loss depends explicitly on q_I through

$$\partial_{q_I}\ell_{\text{full}}(h_\star; q_I) = \sigma(h_\star)^2. \tag{C.133}$$

It has no explicit dependence on q_Γ. Therefore (C.84) reduces to

$$D_\Gamma(h_\star) = -\frac{Fz}{2\sqrt{q_\Gamma}}, \tag{C.134}$$

where the common field is

$$h_0 = \sqrt{\beta}\,\lambda M + \sqrt{q_\Gamma}\,z, \qquad z \sim \mathcal{N}(0,1). \tag{C.135}$$

Substitution into (C.88)–(C.90) gives

$$\widehat{m}(\gamma) = \alpha\sqrt{\beta}\,\mathbb{E}_{\lambda,z}\left[\lambda F\right], \tag{C.136}$$

$$\widehat{q}(\gamma) = \alpha\gamma\,\mathbb{E}_{\lambda,z}\left[F^2\right], \tag{C.137}$$

$$\widehat{v}(\gamma) = 2\alpha\mathbb{E}_{\lambda,z}\left[\sigma(h_\star)^2 - \gamma\frac{Fz}{2\sqrt{q_\Gamma}}\right]. \tag{C.138}$$

Together with the common input-side equations (C.74) and the global closures (C.91), these equations close the full-reconstruction saddle.

D Replica recovery predictions and the role of mask diversity

This appendix explains how mask diversity changes the onset of feature recovery and how dynamic masking exploits heterogeneous coordinate noise. We derive the recovery criterion and its factorization in App. D.1, obtain the weak-signal cost of mask reuse in App. D.2, and examine its finite-signal behavior in App. H.11. The linear theory then identifies the noise-adaptation mechanism in App. D.4. These predictions complement the learning curves and simulation comparisons in Figs. 2 to 4. We use the replica-symmetric equations of App. C at zero ridge; the onset calculation follows the growth of an infinitesimal feature overlap from the uninformative solution.

D.1 Recovery criterion and fixed-signal factorization

This subsection derives the fixed-β recovery criterion for the finite-K masked autoencoder. We set $\lambda_R = 0$, keep $0 < \beta < \infty$ fixed, and work on a regular replica-symmetric uninformative branch with $M = 0$, $q_{\Gamma,0} > 0$, $\mathcal{X}_{K,0} > 0$, and finite output moments. The only small parameter in the instability calculation is the teacher overlap M.

D.1.1 FINITE-K MASKED UNINFORMATIVE SADDLE

Whiten the finite-K field using

$$\boldsymbol{g} := \Omega_K^{-1/2} \boldsymbol{f}_K = (h, x_1, \dots, x_K), \qquad \delta_r = \sqrt{\rho(1-\rho)}\, x_r. \tag{D.1}$$

The common field entering the output proximal is

$$\boldsymbol{g}_0 = \sqrt{\beta}\, \lambda M \boldsymbol{e}_0 + \sqrt{q_\Gamma}\, \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_{K+1}). \tag{D.2}$$

In whitened coordinates the proximal and force are

$$\boldsymbol{g}_\star = \operatorname*{arg\,min}_{\boldsymbol{g} \in \mathbb{R}^{K+1}} \left\{ \ell_K(\boldsymbol{g}; q_I) + \frac{1}{2v_\Gamma} \|\boldsymbol{g} - \boldsymbol{g}_0\|_2^2 \right\}, \tag{D.3}$$

$$\widetilde{\boldsymbol{F}}_K := \Omega_K^{1/2} \boldsymbol{F}_K = \frac{\boldsymbol{g}_\star - \boldsymbol{g}_0}{v_\Gamma}. \tag{D.4}$$

At $M = 0$ write

$$\boldsymbol{g}_0^{(0)} = \sqrt{q_{\Gamma,0}}\, \boldsymbol{z}. \tag{D.5}$$

All quantities below are evaluated at the corresponding uninformative proximal. Define

$$\mathcal{E}_{K,0} := \mathbb{E} \|\widetilde{\boldsymbol{F}}_{K,0}\|_2^2, \tag{D.6}$$

$$\mathcal{X}_{K,0} := \frac{d}{d\varepsilon} \mathbb{E}_{\boldsymbol{z}} \left[\widetilde{F}_{K,h}\left(\sqrt{q_{\Gamma,0}}\, \boldsymbol{z} + \varepsilon \boldsymbol{e}_0\right) \right] \Big|_{\varepsilon=0}, \tag{D.7}$$

$$\mathcal{A}_{K,0} := \mathbb{E}[\partial_{q_I} \ell_K(\boldsymbol{g}_\star; q_{I,0})] = \rho\, \mathbb{E}\left[\frac{1}{K} \sum_{r=1}^{K} \sigma(y_{r,\star})^2 \right] \geq 0, \tag{D.8}$$

$$\mathcal{B}_{K,0} := \mathbb{E}[D_{\Gamma,K}]. \tag{D.9}$$

In (D.7), all other order parameters are held fixed. Differentiating the shifted Gaussian density gives the score expression (D.36); $q_{\Gamma,0} > 0$ and the finite force second moment suffice for this step. This averaged response includes changes in the selected proximal minimizer. It agrees with the expectation of the ordinary force derivative when differentiation can be interchanged with Gaussian averaging. At finite K, the loss has no explicit dependence on q_Γ. Using (C.99) in whitened coordinates gives

$$\mathcal{B}_{K,0} = -\frac{1}{2\sqrt{q_{\Gamma,0}}} \mathbb{E}\left[\widetilde{\boldsymbol{F}}_{K,0}^{\mathsf{T}} \boldsymbol{z} \right]. \tag{D.10}$$

The finite-K conjugate equations (C.101)–(C.102) reduce at $M = 0$ to

$$\widehat{q}_0(\gamma) = \alpha\gamma \mathcal{E}_{K,0}, \qquad \widehat{v}_0(\gamma) = 2\alpha \left(\mathcal{A}_{K,0} + \gamma \mathcal{B}_{K,0} \right). \tag{D.11}$$

The ridgeless input equations therefore give

$$v_0(\gamma) = \frac{1}{2\alpha(\mathcal{A}_{K,0} + \gamma \mathcal{B}_{K,0})}, \tag{D.12}$$

$$q_0(\gamma) = \alpha\gamma \mathcal{E}_{K,0} v_0(\gamma)^2. \tag{D.13}$$

Consequently,

$$q_{I,0} = \alpha \mathcal{E}_{K,0} \mathbb{E}_\gamma[\gamma v_0(\gamma)^2], \tag{D.14}$$

$$q_{\Gamma,0} = \alpha \mathcal{E}_{K,0} \mathbb{E}_\gamma[\gamma^2 v_0(\gamma)^2], \tag{D.15}$$

$$v_{\Gamma,0} = \mathbb{E}_\gamma[\gamma v_0(\gamma)]. \tag{D.16}$$

A regular ridgeless saddle requires

$$\mathcal{A}_{K,0} + \gamma \mathcal{B}_{K,0} > 0 \qquad P_\gamma\text{-a.s.} \tag{D.17}$$

D.1.2 LINEAR INSTABILITY IN THE TEACHER OVERLAP

Recovery begins when the response to a small feature alignment amplifies that alignment. Perturb the uninformative saddle by an infinitesimal M:

$$\boldsymbol{g}_0 = \boldsymbol{g}_0^{(0)} + \sqrt{\beta}\,\lambda M \boldsymbol{e}_0. \tag{D.18}$$

The remaining uninformative order parameters change only at order M^2. Expanding the Gaussian-averaged informative force at fixed remaining order parameters gives

$$\mathbb{E}_{\boldsymbol{z}}\left[\widetilde{F}_{K,h}(\boldsymbol{g}_0^{(0)} + \varepsilon \boldsymbol{e}_0)\right] = \mathbb{E}_{\boldsymbol{z}}\left[\widetilde{F}_{K,h}(\boldsymbol{g}_0^{(0)})\right] + \varepsilon \mathcal{X}_{K,0} + o(\varepsilon). \tag{D.19}$$

Substituting $\varepsilon = \sqrt{\beta}\,\lambda M$ and averaging over λ, using its independence from the $M = 0$ channel, $\mathbb{E}[\lambda] = 0$, and $\mathbb{E}[\lambda^2] = 1$, gives

$$\mathbb{E}[\lambda \widetilde{F}_{K,h}] = \sqrt{\beta}\, M \mathcal{X}_{K,0} + o(M). \tag{D.20}$$

Equation (C.100) then yields the teacher conjugate

$$\widehat{m} = \alpha\beta \mathcal{X}_{K,0} M + o(M). \tag{D.21}$$

The input relation $m(\gamma) = v_0(\gamma)\widehat{m} + o(M)$ gives

$$M_{\mathrm{new}} = \alpha\beta \mathcal{X}_{K,0} \mathbb{E}_\gamma[v_0(\gamma)] M + o(M). \tag{D.22}$$

Thus the exact continuous-instability condition is

$$1 = \alpha_c^{\mathrm{mask},K}\, \beta\, \mathcal{X}_{K,c} \mathbb{E}_\gamma[v_{K,c}(\gamma)]. \tag{D.23}$$

Equivalently, along the regular $M = 0$ branch define

$$\Lambda_K(\alpha;\beta) := \alpha\beta \mathcal{X}_{K,0}(\alpha) \mathbb{E}_\gamma[v_{K,0}(\gamma;\alpha)]. \tag{D.24}$$

The crossing of Λ_K through one predicts the continuous onset of feature recovery on this branch.

D.1.3 EXACT FIXED-β FACTORIZATION

The threshold separates into an output-efficiency factor and a factor describing the response to heterogeneous coordinate noise. The latter is

$$\mathcal{R}_\gamma(\tau) := \frac{\mathbb{E}_\gamma\left[\dfrac{\gamma^2}{(1+\tau\gamma)^2}\right]}{\left(\mathbb{E}_\gamma\left[\dfrac{1}{1+\tau\gamma}\right]\right)^2}, \qquad 1 + \tau\gamma > 0 \quad P_\gamma\text{-a.s.} \tag{D.25}$$

It collects the effect of the variance distribution on the recovery threshold. Define

$$H_{0,K} := \mathbb{E}_\gamma[v_{K,c}(\gamma)], \qquad H_{2,K} := \mathbb{E}_\gamma[\gamma^2 v_{K,c}(\gamma)^2]. \tag{D.26}$$

At criticality, (D.15) reads

$$q_{\Gamma,c} = \alpha_c^{\mathrm{mask},K} \mathcal{E}_{K,c} H_{2,K}. \tag{D.27}$$

Introduce the output-efficiency factor

$$\eta_{K,c}^2 := \frac{\mathcal{E}_{K,c}}{q_{\Gamma,c} \mathcal{X}_{K,c}^2}. \tag{D.28}$$

Substituting $\mathcal{E}_{K,c} = \eta_{K,c}^2 q_{\Gamma,c} \mathcal{X}_{K,c}^2$ into (D.27) and cancelling $q_{\Gamma,c} > 0$ gives

$$1 = \alpha_c^{\mathrm{mask},K} \eta_{K,c}^2 \mathcal{X}_{K,c}^2 H_{2,K}. \tag{D.29}$$

On the other hand, (D.23) gives

$$\mathcal{X}_{K,c} = \frac{1}{\alpha_c^{\mathrm{mask},K} \beta H_{0,K}}. \tag{D.30}$$

Eliminating $\mathcal{X}_{K,c}$ yields

$$\alpha_c^{\mathrm{mask},K} = \frac{\eta_{K,c}^2}{\beta^2}\frac{H_{2,K}}{H_{0,K}^2}. \tag{D.31}$$

Assume $\mathcal{A}_{K,c} > 0$ and define

$$\tau_{K,c} := \frac{\mathcal{B}_{K,c}}{\mathcal{A}_{K,c}}. \tag{D.32}$$

Equation (D.12) becomes

$$v_{K,c}(\gamma) = \frac{1}{2\alpha_c^{\mathrm{mask},K}\mathcal{A}_{K,c}}\frac{1}{1+\tau_{K,c}\gamma}. \tag{D.33}$$

The overall scale cancels from (D.31), giving

$$\alpha_c^{\mathrm{mask},K} = \frac{\eta_{K,c}^2}{\beta^2}\mathcal{R}_\gamma(\tau_{K,c}), \tag{D.34}$$

with $\mathcal{R}_\gamma$ defined in (D.25). No $\beta \downarrow 0$ limit has been used.

Dynamic masking. The same derivation applies after the mask average. One keeps $\mathcal{E}_{\infty,0} = \mathbb{E}[F_{\mathrm{dyn},0}^2]$ and $\mathcal{X}_{\infty,0} = q_{\Gamma,0}^{-1/2}\mathbb{E}[zF_{\mathrm{dyn},0}]$ with the same averaged-response convention, while $\mathcal{B}_{\infty,0}$ includes the explicit derivative of the dynamic loss,

$$\mathcal{B}_{\infty,0} = \mathbb{E}\left[\partial_{q_\Gamma}\ell_{\mathrm{dyn}}(h_\star; q_{I,0}, q_{\Gamma,0}) - \frac{F_{\mathrm{dyn}}z}{2\sqrt{q_{\Gamma,0}}}\right]. \tag{D.35}$$

Since $\widehat{v}_0(\gamma) = 2\alpha(\mathcal{A}_{\infty,0} + \gamma\mathcal{B}_{\infty,0})$, the factorization (D.34) carries over with $K = \infty$ whenever the regular dynamic saddle exists.

D.1.4 Output-efficiency bound

Differentiating the Gaussian density in the averaged response (D.7), evaluated at criticality, gives

$$\mathcal{X}_{K,c} = \frac{1}{\sqrt{q_{\Gamma,c}}}\mathbb{E}[z_h\widetilde{F}_{K,h}]. \tag{D.36}$$

This expression also evaluates the response for a discontinuous proximal force, without differentiating within a selected minimizer branch. Therefore

$$q_{\Gamma,c}\mathcal{X}_{K,c}^2 = \left(\mathbb{E}[z_h\widetilde{F}_{K,h}]\right)^2 \le \mathbb{E}[\widetilde{F}_{K,h}^2] \le \mathbb{E}\|\widetilde{\boldsymbol{F}}_{K,c}\|_2^2 = \mathcal{E}_{K,c}. \tag{D.37}$$

Hence

$$\eta_{K,c}^2 \ge 1. \tag{D.38}$$

Equality requires that the output force have no component in the mask-specific directions and that its informative component be proportional to z_h in L^2.

D.1.5 Monotonicity and bounds for the heteroskedastic factor

Let

$$a_\tau(\gamma) := \frac{1}{1+\tau\gamma}, \qquad X_\tau(\gamma) := \frac{\gamma}{1+\tau\gamma}. \tag{D.39}$$

Then $a_\tau = 1 - \tau X_\tau$ and

$$\mathcal{R}_\gamma(\tau) = \frac{\mathbb{E}_\gamma[X_\tau^2]}{\mathbb{E}_\gamma[a_\tau]^2}. \tag{D.40}$$

The derivatives are

$$\frac{\mathrm{d}X_\tau}{\mathrm{d}\tau} = -X_\tau^2, \qquad \frac{\mathrm{d}a_\tau}{\mathrm{d}\tau} = -X_\tau a_\tau, \tag{D.41}$$

so

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\log\mathcal{R}_\gamma(\tau) = -\frac{\mathbb{E}_\gamma[X_\tau^3]}{\mathbb{E}_\gamma[X_\tau^2]} + \frac{\mathbb{E}_\gamma[X_\tau a_\tau]}{\mathbb{E}_\gamma[a_\tau]}. \tag{D.42}$$

Let $\gamma' \sim P_\gamma$ be independent of γ, and set $X'_\tau = X_\tau(\gamma')$. Symmetrizing the last expression gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mathcal{R}_\gamma(\tau) = -\frac{\mathbb{E}_{\gamma,\gamma'}\left[(X_\tau - X'_\tau)^2(X_\tau + X'_\tau - \tau X_\tau X'_\tau)\right]}{2\,\mathbb{E}_\gamma[X_\tau^2]\,\mathbb{E}_\gamma[a_\tau]} \le 0. \tag{D.43}$$

The sign follows because

$$X_\tau + X'_\tau - \tau X_\tau X'_\tau = \frac{\gamma + \gamma' + \tau\gamma\gamma'}{(1+\tau\gamma)(1+\tau\gamma')} > 0 \tag{D.44}$$

throughout the regularity domain. The inequality is strict for a nondegenerate variance law. At $\tau = 0$,

$$\mathcal{R}_\gamma(0) = \mathbb{E}_\gamma[\gamma^2] = \kappa_2. \tag{D.45}$$

If $\mathbb{E}_\gamma[\gamma^{-1}] < \infty$, dominated convergence gives

$$\lim_{\tau\to+\infty} \mathcal{R}_\gamma(\tau) = \frac{1}{\mathbb{E}_\gamma[\gamma^{-1}]^2}. \tag{D.46}$$

Thus every regular critical saddle obeys

$$\mathcal{R}_\gamma(\tau_{K,c}) \ge \frac{1}{\mathbb{E}_\gamma[\gamma^{-1}]^2}. \tag{D.47}$$

Combining this with (D.38) gives

$$\alpha_c^{\mathrm{mask},K} \ge \frac{1}{\beta^2\mathbb{E}_\gamma[\gamma^{-1}]^2}. \tag{D.48}$$

Finally,

$$\mathbb{E}_\gamma[\gamma^{-2}] \ge \mathbb{E}_\gamma[\gamma^{-1}]^2 \tag{D.49}$$

by Cauchy–Schwarz, and hence

$$\alpha_c^{\mathrm{mask},K} \ge \frac{1}{\beta^2\mathbb{E}_\gamma[\gamma^{-1}]^2} \ge \alpha_c^{\mathrm{BO}}. \tag{D.50}$$

For unbounded variance support, the positivity condition in (D.25) requires $\tau_{K,c} \ge 0$. Hence the factorization gives

$$\alpha_c^{\mathrm{mask},K} \le \frac{\eta_{K,c}^2\kappa_2}{\beta^2}, \tag{D.51}$$

whenever the critical solution has finite output efficiency and $\kappa_2 < \infty$. The linear dynamic case is developed in App. D.3.

D.2 Mask diversity in the weak-signal regime

D.2.1 Recovery predictions and weak-signal regime

The weak-signal limit makes the sample cost of mask reuse explicit. Starting from the high-dimensional equations, we take

$$\lambda_R = 0, \qquad \beta \downarrow 0, \qquad \alpha = O(\beta^{-2}), \qquad K = O(1), \tag{D.52}$$

with $\alpha\beta^2 = O(1)$ and K held fixed while the weak-signal limit is taken. The dynamic-mask coefficient is obtained afterward through the sequential limit $K \to \infty$. We denote

$$\kappa_2 := \mathbb{E}_\gamma[\gamma^2]. \tag{D.53}$$

The analysis concerns the regular continuous branch on which $q_I, q_\Gamma, v_\Gamma \to 0$ as $\beta \downarrow 0$.

Weak-signal recovery predictions. Fix $0 < \rho < 1$ and finite K.

For the smooth class, assume $\sigma \in C^1(\mathbb{R})$, with both σ and σ' globally Lipschitz, and

$$\sigma(0) = 0, \qquad \sigma(x) = s_\sigma x + O(x^2), \qquad 0 < s_\sigma < \infty, \tag{D.54}$$

the continuous weak-recovery threshold is

$$\alpha_{\mathrm{c,sm}}^{(K)} = \frac{\kappa_2}{\beta^2}\left[1 + \frac{1 - 2\rho(1-\rho)}{2K\rho(1-\rho)}\right][1 + o_\beta(1)]\,. \tag{D.55}$$

These sufficient regularity conditions justify the Gaussian moment and proximal expansions below; they include linear, tanh, and the standard ELU $\sigma(x) = x$ for $x \geq 0$, $\sigma(x) = e^x - 1$ for $x < 0$. The second derivative need only exist almost everywhere and be bounded, so the jump in the ELU second derivative at zero is allowed. Activations satisfying these conditions belong to the same weak-signal universality class. Activations with $\sigma(0) \neq 0$, vanishing linear response, or discontinuities require a different expansion.

For ReLU,

$$\sigma(x) = x_+ := \max\{x, 0\}, \tag{D.56}$$

define the off-diagonal force covariance

$$\mathcal{D}_{\mathrm{R}}(\rho) = \rho^2(1-\rho)^2\left[1 + \frac{2}{\pi}\arcsin(1-\rho)\right] + \frac{\rho(1-\rho)}{2\pi}\sqrt{\frac{\rho}{2-\rho}}. \tag{D.57}$$

The ReLU threshold is

$$\alpha_{\mathrm{c,R}}^{(K)} = \frac{\kappa_2}{\beta^2}\left[\frac{\mathcal{D}_{\mathrm{R}}(\rho)}{\rho^2(1-\rho)^2} + \frac{1}{K}\left(\frac{1}{\rho(1-\rho)} - \frac{\mathcal{D}_{\mathrm{R}}(\rho)}{\rho^2(1-\rho)^2}\right)\right][1 + o_\beta(1)]\,. \tag{D.58}$$

For both activation classes, the leading weak-signal threshold is strictly decreasing with the number of independent masks in the fixed collection. The following ordering refers to these leading coefficients, with the dynamic coefficient obtained in the sequential limit specified above:

$$\alpha_{\mathrm{c}}^{(1)} > \alpha_{\mathrm{c}}^{(2)} > \cdots > \alpha_{\mathrm{c}}^{(\infty)}. \tag{D.59}$$

Since $K = 1$ corresponds to static masking and $K \to \infty$ to dynamic masking, the leading dynamic coefficient is strictly smaller than the static one. The fixed-K expansion does not establish uniform finite-signal monotonicity for K increasing jointly with β^{-1}.

At half masking, $\rho = 1/2$, the leading coefficients multiplying κ_2/β^2 are

	$K = 1$	$K = 2$	$K \to \infty$
Smooth	2	$3/2$	1
ReLU	4	$\frac{8}{3} + \frac{1}{\pi\sqrt{3}}$	$\frac{4}{3} + \frac{2}{\pi\sqrt{3}}$

(D.60)

ReLU has a distinct coefficient because its two one-sided slopes differ.

Excess over the dynamic weak-signal reference. For the centered smooth class, let $\alpha_c^{(\infty)} = \kappa_2\beta^{-2}[1 + o_\beta(1)]$ denote the sequential dynamic reference. The finite-K expansion can equivalently be written as

$$\alpha_c^{(K)} - \alpha_c^{(\infty)} = \frac{\alpha_c^{(1)} - \alpha_c^{(\infty)}}{K} + o_\beta(\beta^{-2}). \tag{D.61}$$

Thus each additional independent mask reduces the excess sample requirement, with diminishing returns. The full fixed-signal criterion of App. D.1 is evaluated numerically in App. H.11.

D.2.2 Common weak-signal reduction

Finite-K output channel. Use the whitened field, common Gaussian center, proximal, and force from (D.1)–(D.5). At $M = 0$, $\boldsymbol{g}_0 \sim \mathcal{N}(\mathbf{0}, q_\Gamma \mathbf{I}_{K+1})$. The visible and masked fields of view r are therefore

$$y_r := h_{r,\mathrm{vis}} = (1-\rho)h + \sqrt{\rho(1-\rho)}\,x_r, \qquad t_r := h_{r,\mathrm{mask}} = \rho h - \sqrt{\rho(1-\rho)}\,x_r. \tag{D.62}$$

The finite-K output loss becomes

$$\ell_K(\boldsymbol{g}; q_I) = \frac{1}{K}\sum_{r=1}^{K}\left[\rho q_I \sigma(y_r)^2 - 2\sigma(y_r)t_r\right]. \tag{D.63}$$

At differentiable proximal minimizers, $\widetilde{\boldsymbol{F}}_K = -\nabla_{\boldsymbol{g}}\ell_K(\boldsymbol{g}_\star; q_I)$; at ReLU kinks the displacement in (D.4) defines the force. Since the signal loads on the common component,

$$A_\star = \widetilde{F}_{K,h}, \qquad \boldsymbol{F}_K^\top \Omega_K \boldsymbol{F}_K = \|\widetilde{\boldsymbol{F}}_K\|_2^2. \tag{D.64}$$

The different views share the field h; their force covariances must be retained when computing the effect of K.

For the two activation classes considered below, take the weak-signal limit of the averaged response (D.7), using its Gaussian-score form (D.36), and define

$$\mathcal{X}_K := \lim_{\substack{q_I, q_\Gamma, v_\Gamma \to 0 \\ M=0}} \frac{1}{\sqrt{q_\Gamma}}\mathbb{E}_{\boldsymbol{z}}\left[z_h \widetilde{F}_{K,h}(\sqrt{q_\Gamma}\,\boldsymbol{z})\right], \tag{D.65}$$

$$T_K := \lim_{\substack{q_I, q_\Gamma, v_\Gamma \to 0 \\ M=0}} \frac{1}{q_\Gamma}\mathbb{E}\left[\|\widetilde{\boldsymbol{F}}_K\|_2^2\right]. \tag{D.66}$$

We also require

$$U_K := \lim_{q_\Gamma \downarrow 0} \frac{1}{q_\Gamma}\mathbb{E}\left[\partial_{q_I}\ell_K(\boldsymbol{g}_0; q_I)|_{q_I=0, M=0}\right]. \tag{D.67}$$

The coefficient U_K controls the norm of the uninformative solution but will cancel from the recovery threshold.

Generic reduction of the weak-recovery condition. Taking the weak-signal limit of the response in (D.19) gives, at leading order,

$$\widehat{m} = \alpha\beta\mathcal{X}_K M + o(M). \tag{D.68}$$

At $M = 0$,

$$\widehat{q}(\gamma) = \alpha\gamma q_\Gamma T_K + o(\alpha q_\Gamma). \tag{D.69}$$

It remains to expand $\widehat{v}(\gamma)$. At fixed finite K, the loss (D.63) has no explicit dependence on q_Γ. Thus, at $M = 0$,

$$D_\Gamma = -\widetilde{\boldsymbol{F}}_K^\top \partial_{q_\Gamma}\boldsymbol{g}_0 = -\frac{1}{2q_\Gamma}\widetilde{\boldsymbol{F}}_K^\top \boldsymbol{g}_0. \tag{D.70}$$

Euler identity and the nonlinear remainder. In this paragraph write $q = q_\Gamma$, $Q = q_I$, and $v = v_\Gamma$. All limits have $q, Q, v \downarrow 0$ at fixed K and ρ; constants may depend on K, ρ, and the activation, but not on q, Q, v. For the smooth class the loss decomposes as

$$\ell_K(\boldsymbol{g}; Q) = \ell_K^{(2)}(\boldsymbol{g}) + QG_K^{(2)}(\boldsymbol{g}) + \mathcal{E}_K(\boldsymbol{g}; Q), \qquad |\mathcal{E}_K(\boldsymbol{g}; Q)| \le C(1+Q)\|\boldsymbol{g}\|^3, \tag{D.71}$$

where $\ell_K^{(2)} = -2s_\sigma K^{-1}\sum_j y_j t_j$ and $G_K^{(2)} = \rho s_\sigma^2 K^{-1}\sum_j y_j^2$ are homogeneous of degree two. At the uninformative Gaussian field, y_j and t_j are independent and centered for each view. Thus

$$\mathbb{E}[\ell_K^{(2)}(\boldsymbol{g}_0)] = 0, \qquad \mathbb{E}[G_K^{(2)}(\boldsymbol{g}_0)] = U_K q, \qquad \mathbb{E}[\ell_K(\boldsymbol{g}_0; Q)] = QU_K q + O(Qq^{3/2}). \tag{D.72}$$

The last estimate uses the exact cancellation $\mathbb{E}[\sigma(y_j)t_j] = 0$, rather than just the pointwise cubic bound.

Define the Euler defect of the full loss by

$$\mathcal{R}_K(\boldsymbol{g}; Q) := \boldsymbol{g}^\top \nabla\ell_K(\boldsymbol{g}; Q) - 2\ell_K(\boldsymbol{g}; Q). \tag{D.73}$$

Proximal stationarity gives the exact identity

$$\widetilde{\boldsymbol{F}}_K^\top \boldsymbol{g}_0 = -2\ell_K(\boldsymbol{g}_\star; Q) - v\|\widetilde{\boldsymbol{F}}_K\|^2 - \mathcal{R}_K(\boldsymbol{g}_\star; Q). \tag{D.74}$$

The defect vanishes for the linear loss. For a general activation in the smooth class, put $d_\sigma(y) = y\sigma'(y) - \sigma(y)$. Direct differentiation gives

$$\mathcal{R}_K(\boldsymbol{g}; Q) = \frac{2}{K}\sum_{j=1}^{K}[\rho Q\sigma(y_j) - t_j]d_\sigma(y_j). \tag{D.75}$$

The assumptions on σ imply $|\sigma(y)| \le C|y|$, $|d_\sigma(y)| \le Cy^2$, and $|d'_\sigma(y)| \le C|y|$ almost everywhere. Independence within each pair (y_j, t_j) consequently yields

$$\mathbb{E}[\mathcal{R}_K(\boldsymbol{g}_0; Q)] = O(Qq^{3/2}), \qquad \|\nabla\mathcal{R}_K(\boldsymbol{g}; Q)\| \le C(1+Q)\|\boldsymbol{g}\|^2 \quad \text{a.e.} \tag{D.76}$$

To transfer this estimate to the proximal, note that $\ell_K(\boldsymbol{g}; Q) \ge -C\|\boldsymbol{g}\|^2$ and $\|\nabla\ell_K(\boldsymbol{g}; Q)\| \le C(1+Q)\|\boldsymbol{g}\|$. For sufficiently small v the proximal objective is coercive, and stationarity of any minimizer gives, for small Q, v,

$$\|\boldsymbol{g}_\star\| \le C\|\boldsymbol{g}_0\|, \qquad \|\boldsymbol{g}_\star - \boldsymbol{g}_0\| \le Cv\|\boldsymbol{g}_0\|. \tag{D.77}$$

Integrating the defect gradient along this displacement and using $\mathbb{E}\|\boldsymbol{g}_0\|^3 = O(q^{3/2})$ proves

$$|\mathbb{E}[\mathcal{R}_K(\boldsymbol{g}_\star; Q)]| \le C(Q+v)q^{3/2} = o\big(q(Q+v)\big). \tag{D.78}$$

This is an expectation bound: cancellation of the hidden-field term is essential, and a pointwise cubic remainder alone would not suffice.

The same regularity assumptions give $\|\nabla^2\ell_K(\boldsymbol{g}; Q)\| \le C(1+Q)(1+\|\boldsymbol{g}\|)$ almost everywhere. Taylor expansion between $\boldsymbol{g}_\star$ and $\boldsymbol{g}_0$, together with (D.77), therefore gives

$$\mathbb{E}[\ell_K(\boldsymbol{g}_\star; Q)] = \mathbb{E}[\ell_K(\boldsymbol{g}_0; Q)] - v\mathbb{E}\|\widetilde{\boldsymbol{F}}_K\|^2 + o(vq). \tag{D.79}$$

Indeed the Taylor remainder is $O(v^2(q+q^{3/2})) = o(vq)$. Moreover, $\nabla\ell_K(\boldsymbol{g}; Q) = \nabla\ell_K^{(2)}(\boldsymbol{g}) + O(Q\|\boldsymbol{g}\| + \|\boldsymbol{g}\|^2)$ as $Q \downarrow 0$. The displacement estimate and Gaussian moments then show that $q^{-1}\mathbb{E}\|\widetilde{\boldsymbol{F}}_K\|^2 \to T_K$, as evaluated below.

Positive homogeneity of ReLU. For ReLU, (D.71) holds with $\mathcal{E}_K = 0$, using the ReLU homogeneous terms. The full loss is positively homogeneous of degree two, so the Euler defect is zero. At a kink the force is defined by the proximal displacement; varying $\boldsymbol{g}_\star$ along its radial direction gives (D.74) with zero defect there as well. The bounds (D.77) still hold, since the loss is locally Lipschitz and every generalized gradient has norm at most $C(1+Q)\|\boldsymbol{g}\|$.

For completeness, rescale $\boldsymbol{g}_0 = \sqrt{q}\,\boldsymbol{z}$ and $\boldsymbol{g}_\star = \sqrt{q}\,\boldsymbol{z}_\star$. At almost every Gaussian $\boldsymbol{z}$, all visible fields are nonzero. For small enough v the proximal remains in the same activation region, where the loss is quadratic, and the first-order energy change divided by vq tends to $-\|\nabla\ell_K^{(2)}(\boldsymbol{z})\|^2$ as $Q, v \downarrow 0$. The displacement and gradient bounds dominate this quotient, and the squared force divided by q, by $C\|\boldsymbol{z}\|^2$. Dominated convergence proves (D.79) and, more precisely,

$$q^{-1/2}\widetilde{\boldsymbol{F}}_K(\sqrt{q}\,\boldsymbol{z}) \longrightarrow -\nabla\ell_K^{(2)}(\boldsymbol{z}) \qquad \text{in } L^2(\boldsymbol{z}).$$

This convergence yields both $q^{-1}\mathbb{E}\|\widetilde{\boldsymbol{F}}_K\|^2 \to T_K$ and the Gaussian-score response in (D.65), including the effect of activation-boundary crossings. Also $\mathbb{E}[\ell_K(\boldsymbol{g}_0; Q)] = QU_Kq$ exactly.

Susceptibility expansion. For either activation class, substitute (D.72), (D.74), and (D.79) into (D.70). The smooth Euler defect contributes $o(Q+v)$ after division by $2q$; it is identically zero in the ReLU case. Restoring the original notation yields

$$\mathbb{E}[D_\Gamma] = U_Kq_I - \frac{1}{2}v_\Gamma T_K + o(q_I + v_\Gamma). \tag{D.80}$$

Likewise, the displacement bound and the Gaussian moment estimate give $\mathbb{E}[\partial_{q_I}\ell_K(\boldsymbol{g}_\star; q_I)] = U_Kq_\Gamma + o(q_\Gamma)$ for both activation classes. The conjugate susceptibility therefore becomes

$$\widehat{v}(\gamma) = 2\alpha\left[U_Kq_\Gamma + \gamma\left(U_Kq_I - \frac{1}{2}v_\Gamma T_K\right)\right] + o(\alpha(q_I + q_\Gamma + v_\Gamma)). \tag{D.81}$$

The regular uninformative branch has the scaling

$$q_I = \beta\bar{q}_I + o(\beta), \qquad q_\Gamma = \beta\bar{q}_\Gamma + o(\beta), \qquad v(\gamma) = \beta\bar{v}(\gamma) + o(\beta), \qquad v_\Gamma = \beta\bar{v}_\Gamma + o(\beta), \tag{D.82}$$

with $\bar{q}_\Gamma > 0$.

At $M = 0$, the ridgeless input equations together with (D.69) give

$$q(\gamma) = \alpha\beta^3\gamma\bar{q}_\Gamma T_K \bar{v}(\gamma)^2 + o(\beta). \tag{D.83}$$

Consequently,

$$1 = \alpha\beta^2 T_K \mathbb{E}_\gamma\left[\gamma^2\bar{v}(\gamma)^2\right], \tag{D.84}$$

$$\bar{q}_I = \alpha\beta^2\bar{q}_\Gamma T_K \mathbb{E}_\gamma\left[\gamma\bar{v}(\gamma)^2\right]. \tag{D.85}$$

On the other hand, $v(\gamma) = 1/\widehat{v}(\gamma)$ in the ridgeless problem. Using (D.81),

$$\frac{1}{\bar{v}(\gamma)} = 2\alpha\beta^2\left[U_K\bar{q}_\Gamma + \gamma\left(U_K\bar{q}_I - \frac{1}{2}\bar{v}_\Gamma T_K\right)\right]. \tag{D.86}$$

Multiplying by $\gamma\bar{v}(\gamma)^2$, averaging over γ, and using (D.84) and (D.85) gives

$$\bar{v}_\Gamma = \frac{4U_K}{T_K}\bar{q}_I - \bar{v}_\Gamma. \tag{D.87}$$

Hence

$$2U_K\bar{q}_I = T_K\bar{v}_\Gamma. \tag{D.88}$$

The coefficient multiplying γ in (D.86) therefore vanishes. Thus the leading input susceptibility is forced to be independent of the variance type,

$$\bar{v}(\gamma) = \bar{v}. \tag{D.89}$$

Using $\mathbb{E}_\gamma[\gamma] = 1$ and $\kappa_2 = \mathbb{E}_\gamma[\gamma^2]$, the uninformative saddle then satisfies

$$1 = \alpha\beta^2\kappa_2 T_K\bar{v}^2. \tag{D.90}$$

Moreover,

$$\frac{\bar{q}_I}{\bar{q}_\Gamma} = \alpha\beta^2 T_K\bar{v}^2 = \frac{1}{\kappa_2}, \qquad \bar{v}_\Gamma = \bar{v}. \tag{D.91}$$

Thus the full variance distribution enters the leading weak-signal threshold only through its second moment κ_2.

Finally, the magnetization equation gives

$$M = \alpha\beta^2\mathcal{X}_K\bar{v}\,M + o(M). \tag{D.92}$$

A continuous instability occurs when

$$1 = \alpha\beta^2\mathcal{X}_K\bar{v}. \tag{D.93}$$

Eliminating $\bar{v}$ between (D.90) and (D.93) gives

$$\alpha_{\mathrm{c}}^{(K)} = \frac{\kappa_2}{\beta^2}\frac{T_K}{\mathcal{X}_K^2}\left[1 + o_\beta(1)\right]. \tag{D.94}$$

The activation and the number of masks therefore enter only through the output ratio $T_K/\mathcal{X}_K^2$.

D.2.3 Centered smooth activations with nonzero local slope

Under (D.54), the leading output loss is

$$\ell_{K,\mathrm{sm}}^{(2)} = -\frac{2s_\sigma}{K}\sum_{r=1}^{K} y_r t_r. \tag{D.95}$$

The positive reconstruction term does not contribute to the leading force, but its q_I derivative is required in (D.81). Since

$$\mathrm{Var}(y_r) = (1-\rho)q_\Gamma, \tag{D.96}$$

we obtain

$$U_{K,\mathrm{sm}} = \rho(1-\rho)s_\sigma^2. \tag{D.97}$$

Expanding the product of the visible and masked fields,

$$y_r t_r = \rho(1-\rho)h^2 + (2\rho-1)\sqrt{\rho(1-\rho)}\,hx_r - \rho(1-\rho)x_r^2. \tag{D.98}$$

The Hessian of the leading loss in the coordinates $(h, x_1, \dots, x_K)$ is therefore

$$\mathsf{H}_{K,\mathrm{sm}} = \begin{pmatrix} -4s_\sigma\rho(1-\rho) & \dfrac{2s_\sigma(1-2\rho)\sqrt{\rho(1-\rho)}}{K}\mathbf{1}_K^\top \\ \dfrac{2s_\sigma(1-2\rho)\sqrt{\rho(1-\rho)}}{K}\mathbf{1}_K & \dfrac{4s_\sigma\rho(1-\rho)}{K}\mathbf{I}_K \end{pmatrix}. \tag{D.99}$$

The preceding displacement and moment bounds give, in L^2,

$$\widetilde{\boldsymbol{F}}_{K,\mathrm{sm}} = -\mathsf{H}_{K,\mathrm{sm}}\boldsymbol{g}_0 + o(\sqrt{q_\Gamma}). \tag{D.100}$$

The informative susceptibility is therefore

$$\mathcal{X}_{K,\mathrm{sm}} = 4s_\sigma\rho(1-\rho). \tag{D.101}$$

Moreover,

$$\begin{aligned} T_{K,\mathrm{sm}} &= \mathrm{Tr}\left(\mathsf{H}_{K,\mathrm{sm}}^2\right) \\ &= 16s_\sigma^2\rho^2(1-\rho)^2 + \frac{8s_\sigma^2\rho(1-\rho)}{K}\left[(1-2\rho)^2 + 2\rho(1-\rho)\right] \\ &= 16s_\sigma^2\rho^2(1-\rho)^2 + \frac{8s_\sigma^2\rho(1-\rho)[1-2\rho(1-\rho)]}{K}. \end{aligned} \tag{D.102}$$

Hence

$$\frac{T_{K,\mathrm{sm}}}{\mathcal{X}_{K,\mathrm{sm}}^2} = 1 + \frac{1-2\rho(1-\rho)}{2K\rho(1-\rho)}. \tag{D.103}$$

The slope s_σ cancels, and substitution into (D.94) gives (D.55).

For consecutive values of K,

$$\alpha_{\mathrm{c,sm}}^{(K)} - \alpha_{\mathrm{c,sm}}^{(K+1)} = \frac{\kappa_2}{\beta^2}\frac{1-2\rho(1-\rho)}{2\rho(1-\rho)K(K+1)}\left[1 + o_\beta(1)\right] > 0. \tag{D.104}$$

Indeed, $\rho(1-\rho) \le 1/4$, so $1 - 2\rho(1-\rho) \ge 1/2$.

D.2.4 ReLU activation

For ReLU, positive homogeneity gives the leading loss

$$\ell_{K,\mathrm{R}}^{(2)} = -\frac{2}{K}\sum_{r=1}^{K}(y_r)_+ t_r. \tag{D.105}$$

At $M = 0$, $y_r \sim \mathcal{N}(0, (1-\rho)q_\Gamma)$, and hence

$$U_{K,\mathrm{R}} = \rho\frac{\mathbb{E}[(y_r)_+^2]}{q_\Gamma} = \frac{\rho(1-\rho)}{2}. \tag{D.106}$$

Let $\boldsymbol{e}_0$ denote the h direction and $\boldsymbol{e}_r$ the x_r direction. The force contribution of view r can be written as

$$\boldsymbol{\psi}_r := \mathbf{1}_{\{y_r>0\}}\left[\big((1-\rho)t_r + \rho y_r\big)\boldsymbol{e}_0 + \sqrt{\rho(1-\rho)}(t_r - y_r)\boldsymbol{e}_r\right]. \tag{D.107}$$

The leading force is then

$$\widetilde{\boldsymbol{F}}_{K,\mathrm{R}} = \frac{2}{K}\sum_{r=1}^{K}\boldsymbol{\psi}_r. \tag{D.108}$$

Informative susceptibility. The common component of $\boldsymbol{\psi}_r$ is

$$(\boldsymbol{\psi}_r)_h = \mathbf{1}_{\{y_r>0\}}\left[(1-\rho)t_r + \rho y_r\right]. \tag{D.109}$$

Evaluate (D.65) with the leading force. At $M=0$, $h = y_r + t_r$, while y_r and t_r are independent centered Gaussians with variances $(1-\rho)q_\Gamma$ and ρq_Γ. Consequently,

$$\mathbb{E}[h(\boldsymbol{\psi}_r)_h] = (1-\rho)\mathbb{E}[t_r^2]\mathbb{P}(y_r>0) + \rho\mathbb{E}[y_r^2\mathbf{1}_{\{y_r>0\}}] = \rho(1-\rho)q_\Gamma.$$

Since $h = \sqrt{q_\Gamma}z_h$, summing the K view contributions in (D.108) and dividing by q_Γ gives

$$\mathcal{X}_{K,\mathrm{R}} = 2\rho(1-\rho). \tag{D.110}$$

Force variance: diagonal term. A direct expansion gives

$$\|\boldsymbol{\psi}_r\|_2^2 = \mathbf{1}_{\{y_r>0\}}\left[(1-\rho)t_r^2 + \rho y_r^2\right]. \tag{D.111}$$

At $q_\Gamma = 1$,

$$y_r \sim \mathcal{N}(0, 1-\rho), \qquad t_r \sim \mathcal{N}(0,\rho), \tag{D.112}$$

and the two variables are independent. Therefore

$$\mathbb{E}\|\boldsymbol{\psi}_r\|_2^2 = \rho(1-\rho). \tag{D.113}$$

Force variance: off-diagonal term. For $r \neq s$, the mask-specific directions $\boldsymbol{e}_r$ and $\boldsymbol{e}_s$ are orthogonal, so only the common components contribute to $\boldsymbol{\psi}_r^\top \boldsymbol{\psi}_s$. We have

$$(1-\rho)t_r + \rho y_r = 2\rho(1-\rho)h + (2\rho-1)\sqrt{\rho(1-\rho)}\,x_r. \tag{D.114}$$

Conditioning on $h = H$, with $H \sim \mathcal{N}(0,1)$, the variables x_r and x_s are independent.

Let

$$\varphi(u) := \frac{e^{-u^2/2}}{\sqrt{2\pi}}, \qquad \Phi_{\mathrm{G}}(u) := \int_{-\infty}^{u}\varphi(v)\,\mathrm{d}v. \tag{D.115}$$

The conditional expectation of the common force component is

$$\begin{aligned}&\mathbb{E}_x\left[\mathbf{1}_{\{(1-\rho)H+\sqrt{\rho(1-\rho)}\,x>0\}} \times \left(2\rho(1-\rho)H + (2\rho-1)\sqrt{\rho(1-\rho)}\,x\right)\right] \\ &= 2\rho(1-\rho)H\Phi_{\mathrm{G}}\left(\sqrt{\frac{1-\rho}{\rho}}\,H\right) + (2\rho-1)\sqrt{\rho(1-\rho)}\,\varphi\left(\sqrt{\frac{1-\rho}{\rho}}\,H\right).\end{aligned} \tag{D.116}$$

Consequently,

$$\mathcal{D}_{\mathrm{R}}(\rho) = \mathbb{E}[\boldsymbol{\psi}_r^\top \boldsymbol{\psi}_s], \qquad r \neq s, \tag{D.117}$$

is the expectation over H of the square of (D.116).

The required Gaussian identities are

$$\mathbb{E}\left[\varphi\left(\sqrt{\frac{1-\rho}{\rho}}H\right)^2\right] = \frac{1}{2\pi}\sqrt{\frac{\rho}{2-\rho}}, \tag{D.118}$$

$$\mathbb{E}\left[H\Phi_{\mathrm{G}}\left(\sqrt{\frac{1-\rho}{\rho}}H\right)\varphi\left(\sqrt{\frac{1-\rho}{\rho}}H\right)\right] = \frac{\sqrt{\rho(1-\rho)}}{2\pi}\sqrt{\frac{\rho}{2-\rho}}, \tag{D.119}$$

$$\mathbb{E}\left[H^2\Phi_{\mathrm{G}}\left(\sqrt{\frac{1-\rho}{\rho}}H\right)^2\right] = \frac{1}{4} + \frac{\arcsin(1-\rho)}{2\pi} + \frac{1-\rho}{\pi}\sqrt{\frac{\rho}{2-\rho}}. \tag{D.120}$$

Substituting these identities into (D.117) gives the closed expression (D.57).

Combining the diagonal and off-diagonal contributions,

$$T_{K,\mathrm{R}} = 4\left[\mathcal{D}_{\mathrm{R}}(\rho) + \frac{\rho(1-\rho) - \mathcal{D}_{\mathrm{R}}(\rho)}{K}\right]. \tag{D.121}$$

Together with (D.110),

$$\frac{T_{K,\mathrm{R}}}{\mathcal{X}_{K,\mathrm{R}}^2} = \frac{\mathcal{D}_{\mathrm{R}}(\rho)}{\rho^2(1-\rho)^2} + \frac{1}{K}\left[\frac{1}{\rho(1-\rho)} - \frac{\mathcal{D}_{\mathrm{R}}(\rho)}{\rho^2(1-\rho)^2}\right]. \tag{D.122}$$

Substitution into (D.94) gives (D.58).

To establish strict monotonicity, Cauchy–Schwarz gives

$$\mathcal{D}_{\mathrm{R}}(\rho) = \mathbb{E}[\boldsymbol{\psi}_r^\top \boldsymbol{\psi}_s] \le \sqrt{\mathbb{E}\|\boldsymbol{\psi}_r\|_2^2 \mathbb{E}\|\boldsymbol{\psi}_s\|_2^2} = \rho(1-\rho). \tag{D.123}$$

The inequality is strict for $0 < \rho < 1$, because independently masked views do not generate identical force vectors almost surely. Hence

$$\alpha_{\mathrm{c,R}}^{(K)} - \alpha_{\mathrm{c,R}}^{(K+1)} = \frac{\kappa_2}{\beta^2} \frac{\rho(1-\rho) - \mathcal{D}_{\mathrm{R}}(\rho)}{\rho^2(1-\rho)^2 K(K+1)} \left[1 + o_\beta(1)\right] > 0. \tag{D.124}$$

D.3 Dynamic linear masking at fixed signal strength

For linear activation, the fixed-signal equations give both the recovery threshold and the feature alignment above it. We first identify when dynamic masking favors cleaner coordinates, then recover the homogeneous PCA learning curve. The comparison with static masking is developed in App. D.4. Throughout, $\sigma(h) = h$ and $\lambda_R = 0$.

D.3.1 Quadratic output and the critical equations

Write $Q = q_I$, $C = q_\Gamma$, and $\Delta = v_\Gamma$. The exact dynamic effective loss for the linear activation is

$$\ell_{\mathrm{dyn}}^{\mathrm{lin}}(h; Q, C) = ah^2 + bC, \qquad a = \rho(1-\rho)((1-\rho)Q - 2), \qquad b = \rho(1-\rho)(\rho Q + 2). \tag{D.125}$$

For $D = 1 + 2a\Delta > 0$, the output proximal and force are

$$h_\star = h_0/D, \qquad F = \mathcal{X} h_0, \qquad \mathcal{X} = -2a/D, \qquad h_0 = \sqrt{\beta} M\lambda + \sqrt{C} z. \tag{D.126}$$

Set $T = \beta M^2 + C$. The explicit derivatives of the loss give

$$\begin{aligned} U &= \rho(1-\rho)\left((1-\rho)T/D^2 + \rho C\right), \qquad H = b - \mathcal{X}/2, \\ \widehat{m} &= \alpha\beta\mathcal{X} M, \qquad \widehat{q}(\gamma) = \alpha\gamma\mathcal{X}^2 T, \qquad \widehat{v}(\gamma) = 2\alpha(U + \gamma H). \end{aligned} \tag{D.127}$$

In particular, H includes the explicit derivative $\partial_C \ell = b$; discarding it would impose the negative variance coefficient characteristic of the full-reconstruction susceptibility. On a nontrivial regular saddle, $U > 0$ and

$$v(\gamma) = \frac{1}{2\alpha U(1+\tau\gamma)}, \qquad \tau = H/U, \qquad 1 + \tau\gamma > 0 \quad P_\gamma\text{-a.s.} \tag{D.128}$$

Define, whenever the indicated moments are finite,

$$I_{j,\ell}(\tau) := \mathbb{E}_\gamma\left[\frac{\gamma^j}{(1+\tau\gamma)^\ell}\right]. \tag{D.129}$$

At a regular continuous critical point with $M = 0$ and $C > 0$, the variance closure and the magnetization instability condition are

$$1 = \alpha\mathcal{X}^2 \mathbb{E}_\gamma[\gamma^2 v(\gamma)^2], \qquad 1 = \alpha\beta\mathcal{X}\mathbb{E}_\gamma[v(\gamma)]. \tag{D.130}$$

Consequently $\mathcal{X} > 0$, $0 < D < 1$, and

$$\begin{aligned} \alpha_{c,\mathrm{RS}}^{\mathrm{dyn}} &= \frac{I_{2,2}(\tau_c)}{\beta^2 I_{0,1}(\tau_c)^2} = \frac{\mathcal{R}_\gamma(\tau_c)}{\beta^2}, \\ \frac{Q}{C} &= \frac{I_{1,2}}{I_{2,2}}, \qquad D = \left(1 + \frac{I_{1,1}}{\sqrt{\alpha I_{2,2}}}\right)^{-1}. \end{aligned} \tag{D.131}$$

This is the fixed-signal factorization with output-efficiency factor one; no weak-signal expansion is involved.

D.3.2 Upper bound and the masking condition

A positive response parameter favors cleaner coordinates and lowers the threshold below the second-moment scale κ_2/β^2. We now relate this improvement to the masking rate. The monotonicity argument in App. D.1 gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mathcal{R}_\gamma(\tau) = -2\frac{I_{3,3}}{I_{2,2}} + 2\frac{I_{1,2}}{I_{0,1}} \le 0, \qquad \mathcal{R}_\gamma(0) = \kappa_2. \tag{D.132}$$

For finite κ_2, continuity at zero extends this monotonicity bound even when $\kappa_3 = \infty$. When the inverse moments are finite, the same calculation gives

$$\frac{1}{\beta^2\mathbb{E}_\gamma[\gamma^{-2}]} \le \frac{1}{\beta^2\mathbb{E}_\gamma[\gamma^{-1}]^2} \le \alpha_{c,\mathrm{RS}}^{\mathrm{dyn}} \le \frac{\kappa_2}{\beta^2}, \qquad \tau_c \ge 0. \tag{D.133}$$

The last inequality needs only the finite second moment.

We now determine the sign of τ_c. At $M = 0$, write $A = 1 - \rho + \rho D^2$. Using $H = \rho(1-\rho)Q - \mathcal{X}(1-D)/2$, $U = \rho(1-\rho)CA/D^2$, and (D.130), one obtains

$$\tau I_{2,2} + D I_{1,1} = \frac{D^2 I_{1,2}}{A}, \qquad \frac{I_{1,2}}{I_{1,1}} = \frac{(1+D)A}{A+D^2}. \tag{D.134}$$

Here $I_{1,1} - I_{1,2} = \tau I_{2,2}$. Since $0 < D < 1$,

$$\tau \ge 0 \quad\Longleftrightarrow\quad A \le D \quad\Longleftrightarrow\quad \rho \ge \frac{1}{1+D}. \tag{D.135}$$

At fixed α, $D(\tau)$ in (D.131) is nonincreasing, because

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\log\frac{I_{1,1}}{\sqrt{I_{2,2}}} = \frac{I_{1,1}I_{3,3} - I_{2,2}^2}{I_{1,1}I_{2,2}} \ge 0. \tag{D.136}$$

The numerator is nonnegative by Cauchy–Schwarz. At zero, $D(0) = \sqrt{\alpha\kappa_2}/(1+\sqrt{\alpha\kappa_2})$. Comparing the positive and negative sides of zero in (D.135) therefore yields

$$\tau_c \ge 0 \quad\Longleftrightarrow\quad \rho \ge \frac{1+\sqrt{\alpha_c\kappa_2}}{1+2\sqrt{\alpha_c\kappa_2}}. \tag{D.137}$$

For a nondegenerate law, strict monotonicity of $\mathcal{R}_\gamma$ implies that $\tau_c > 0$ corresponds to $\alpha_c < \kappa_2/\beta^2$, and $\tau_c < 0$ to the reverse inequality. The right-hand side of (D.137) decreases with α_c. Evaluating it at κ_2/β^2 gives the equivalent fixed-signal statement

$$\tau_c \ge 0 \quad\Longleftrightarrow\quad \rho \ge \frac{\beta+\kappa_2}{\beta+2\kappa_2} \quad\Longrightarrow\quad \alpha_{c,\mathrm{RS}}^{\mathrm{dyn}} \le \frac{\kappa_2}{\beta^2}. \tag{D.138}$$

Equality holds at $\tau_c = 0$. For the homogeneous law the threshold is β^{-2} for every regular critical solution, irrespective of the sign of τ_c.

For unbounded support, positivity in (D.128) rules out $\tau_c < 0$, so the masking condition also bounds the domain of this continuous-onset calculation. It does not imply that recovery is impossible at smaller masking rates: under the Gaussian model with variances bounded away from zero and finite second moment, Theorem E.7 guarantees recovery at sufficiently large sample ratio for every $\rho \in (0,1)$. The transition mechanism at those smaller rates remains to be determined.

D.3.3 Feature alignment above threshold and the homogeneous benchmark

The full linear dynamic equations can also be reduced without linearizing in M. To avoid confusing the cosine $\theta_{\boldsymbol{u}^\star}$ with a ratio of second moments, define $z_M = M^2/C$ and $r_Q = Q/C$. For $M \neq 0$, the magnetization equation gives $\mathcal{X}/(2U) = 1/(\beta I_{0,1})$, hence

$$q(\gamma) = \frac{M^2 + \gamma T/(\alpha\beta^2)}{I_{0,1}^2(1+\tau\gamma)^2}. \tag{D.139}$$

The remaining equations reduce to

$$z_M = \frac{I_{0,1}^2 - I_{2,2}/(\alpha\beta^2)}{I_{1,2} + I_{2,2}/(\alpha\beta)},$$
$$r_Q = \frac{z_M I_{0,2} + (1+\beta z_M) I_{1,2}/(\alpha\beta^2)}{I_{0,1}^2},$$
$$D = \left(1 + \frac{I_{1,1}}{\alpha\beta I_{0,1}}\right)^{-1},$$
$$\Phi(\tau,\alpha) := \tau - \frac{r_Q}{\rho + (1-\rho)(1+\beta z_M)D^{-2}} + \frac{1-D}{\beta I_{0,1}} = 0. \tag{D.140}$$

All $I_{j,\ell}$ are evaluated at τ; β, ρ, P_γ are fixed in the notation $\Phi(\tau,\alpha)$. A root must have $z_M > 0$ and satisfy the regularity conditions. Its scale is recovered from

$$C = \frac{2}{(1-\rho)r_Q + [\rho + (1-\rho)(1+\beta z_M)D^{-2}]D/(\beta I_{0,1})}, \tag{D.141}$$
$$Q = r_Q C, \qquad M^2 = z_M C.$$

The squared alignment is $\theta^2_{\boldsymbol{u}^\star} = z_M/r_Q$. For $P_\gamma = \delta_1$, all susceptibility factors cancel, $r_Q = 1$, and this becomes

$$\theta^2_{\boldsymbol{u}^\star} = \frac{\alpha\beta^2 - 1}{\alpha\beta^2 + \beta}, \qquad \alpha > \beta^{-2}, \tag{D.142}$$

which recovers the PCA learning curve throughout the informative regime, consistent with the agreement observed in Fig. 2(a).

D.4 Mask diversity changes how the learner uses heterogeneous noise

The weak-signal law in (7) describes an additional sample requirement caused by mask reuse. At finite signal strength, the replica equations reveal a further distinction: static and dynamic masking can use the coordinate noise profile differently when learning first becomes possible. Static masking has a noise-independent response to an emerging feature at its regular linear recovery transition. Dynamic masking can instead give cleaner coordinates a larger response, lowering the threshold through an adaptation that the static critical solution lacks. Thus the protocols differ not only in how many samples they require, but also in how they extract the feature from those samples.

We develop this interpretation using the fixed-signal factorization in App. D.1.3 and the dynamic linear equations in App. D.3. Throughout this subsection, $\sigma(h) = h$, $\lambda_R = 0$, $\mathbb{E}_\gamma \gamma = 1$, and $\kappa_2 = \mathbb{E}_\gamma \gamma^2 < \infty$; the concrete examples have finite positive variance support and Gaussian latent priors. We evaluate the response at the continuous recovery onset defined above.

D.4.1 Coordinate response separates noise adaptation from mask reuse

The coordinate susceptibility $v_c(\gamma)$ describes the response to an infinitesimal informative field near recovery. In particular, the input equation gives $m(\gamma) = \widehat{m} v(\gamma)$, with the same $\widehat{m}$ at every coordinate variance. Ratios of susceptibilities therefore determine where the emerging feature alignment is concentrated. Using (D.33), for two variances $0 < \gamma_1 < \gamma_2$,

$$\frac{v_c(\gamma_1)}{v_c(\gamma_2)} = \frac{1 + \tau_c \gamma_2}{1 + \tau_c \gamma_1}. \tag{D.143}$$

For $\tau_c > 0$, this ratio exceeds one: the response favors cleaner coordinates. At $\tau_c = 0$ it is uniform, while negative τ_c favors noisier coordinates on its admissible domain. This is a statement about the response to the shared feature, not equality or ordering of the realized coordinate weights. The noise levels are not supplied to the autoencoder; their influence on the response emerges from the reconstruction objective and the data.

The general finite-K factorization (D.34) separates two contributions: $\eta^2_{K,c}$ measures output efficiency, while $\mathcal{R}_\gamma(\tau_{K,c})$ measures how coordinate reweighting changes the effective noise burden. Both depend on mask diversity through the full equations. The static and dynamic endpoints below show how the response itself changes, beyond the sample penalty captured by the weak-signal law.

D.4.2 WHY STATIC MASKING CANNOT ADAPT ITS RESPONSE AT THE TRANSITION

For $K = 1$, the regular uninformative equations force $\mathcal{B}_{1,0} = 0$. We give the short elimination because this identity is the source of the distinction, rather than an imposed approximation. Write $Q = q_{I,0}$, $C = q_{\Gamma,0}$, and $\Delta = v_{\Gamma,0}$, and use the visible and hidden fields y, t of the $K = 1$ channel. At $M = 0$, their Gaussian centers y_0, t_0 are independent, with variances $(1-\rho)C$ and ρC. The quadratic loss is $\rho Q y^2 - 2yt$. Eliminating t in the proximal gives

$$y_\star = \frac{y_0/[(1-\rho)\Delta] + 2t_0}{\mathcal{D}_{\rm stat}}, \qquad \mathcal{D}_{\rm stat} = \frac{1}{(1-\rho)\Delta} + 2\rho(Q - 2\Delta) > 0. \tag{D.144}$$

The last inequality is the remaining proximal-curvature condition. Evaluating the definitions in (D.8), (D.10), and (D.6) gives

$$\begin{aligned}
\mathcal{A}_{1,0} &= \frac{\rho C}{\mathcal{D}_{\rm stat}^2}\left[\frac{1}{(1-\rho)\Delta^2} + 4\rho\right], \\
\mathcal{B}_{1,0} &= \frac{1}{2\Delta} - \frac{1}{2\mathcal{D}_{\rm stat}}\left[\frac{1}{(1-\rho)\Delta^2} + 4\rho\right], \\
\mathcal{E}_{1,0} &= \left[\frac{1}{(1-\rho)\Delta^2} + 4\rho\right]\frac{\mathcal{A}_{1,0}}{\rho} + \frac{C}{\Delta^2} - \frac{2C}{(1-\rho)\Delta^3\mathcal{D}_{\rm stat}}.
\end{aligned} \tag{D.145}$$

The input equations (D.12)–(D.16) imply

$$\mathcal{E}_{1,0}\Delta = 2\mathcal{A}_{1,0}Q + 2\mathcal{B}_{1,0}C. \tag{D.146}$$

Substituting the explicit output expressions into the difference of the two sides yields the identity

$$\mathcal{E}_{1,0}\Delta - 2\mathcal{A}_{1,0}Q - 2\mathcal{B}_{1,0}C = -\frac{4C\mathcal{B}_{1,0}}{(1-\rho)\Delta\mathcal{D}_{\rm stat}}. \tag{D.147}$$

Since $C, \Delta, \mathcal{D}_{\rm stat} > 0$ and $0 < \rho < 1$, the input closure forces $\mathcal{B}_{1,0} = 0$. Consequently,

$$\tau_{1,c} = 0, \qquad v_0(\gamma) = \Delta, \qquad Q = 4\Delta, \qquad C = \kappa_2 Q. \tag{D.148}$$

Here the constant susceptibility follows from $\mathbb{E}_\gamma \gamma = 1$; the relations for Q, C follow from (D.145) and the variance closures. In fact $q_0(\gamma) = 4\gamma\Delta$, so noisy coordinates still contribute more to the uninformative weight fluctuations.

The remaining closure gives

$$\Delta = \frac{1}{2\sqrt{\rho(1-\rho)(2\alpha\kappa_2 - 1)}}, \qquad 2\alpha\kappa_2 > 1. \tag{D.149}$$

Evaluating the teacher-response gain (D.23) then gives

$$\alpha\beta\mathcal{X}_{1,0}\mathbb{E}_\gamma v_0 = \frac{\beta}{\kappa_2}\left[\sqrt{\rho(1-\rho)(2\alpha\kappa_2 - 1)} + \frac{2\rho - 1}{2}\right]. \tag{D.150}$$

Setting this gain to one yields the regular static threshold

$$\alpha_c^{\rm stat} = \frac{1}{2\kappa_2}\left[1 + \frac{(\kappa_2/\beta + 1/2 - \rho)^2}{\rho(1-\rho)}\right], \qquad \frac{\kappa_2}{\beta} + \frac{1}{2} - \rho > 0. \tag{D.151}$$

Thus two variance laws with the same mean and second moment have the same regular static threshold, even if their clean coordinates occur with very different frequencies. This insensitivity is specific to the uninformative saddle and its recovery transition: it does not imply that static masking cannot develop noise-dependent feature alignment after recovery.

D.4.3 DYNAMIC MASKING EXPLOITS CLEANER COORDINATES AT RECOVERY ONSET

Mask averaging changes the output channel. In the linear case, its explicit dependence on C contributes to H in (D.127); the static cancellation above no longer applies. The resulting susceptibility and threshold are already derived in (D.128) and (D.131). Their interpretation follows from the monotonicity proved in App. D.1.5: positive τ_c downweights high-noise coordinates and strictly lowers $\mathcal{R}_\gamma$ below its neutral value κ_2 for a nondegenerate variance law.

The masking condition (D.138) makes the relevant regime explicit:

$$\rho > \frac{\beta + \kappa_2}{\beta + 2\kappa_2} \quad \Longrightarrow \quad \tau_c > 0, \qquad \alpha_c^{\mathrm{dyn}} < \frac{\kappa_2}{\beta^2}, \tag{D.152}$$

whenever the regular critical saddle exists and the variance law is nonconstant. The learner can therefore exploit the heterogeneous reliability of coordinates while the informative direction is only beginning to emerge. Static masking, whose response is uniform at this point, cannot obtain the same threshold reduction by such reweighting.

For $0 < \beta < 2\kappa_2$, optimizing (D.151) gives

$$\min_{\rho} \alpha_c^{\mathrm{stat}} = \frac{2\kappa_2}{\beta^2}, \qquad \rho_{\mathrm{stat}}^{\star} = \frac{1}{2} + \frac{\beta}{4\kappa_2}. \tag{D.153}$$

Hence a dynamic critical saddle with $\tau_c > 0$ requires less than half the sample ratio of even the best regular static crossing in this regime. The factor-of-two weak-signal comparison is therefore not the whole advantage: finite-signal noise adaptation can increase it. The signal range above is the domain of this optimized-static comparison. The dynamic masking boundary also depends on β; its advantage at fixed ρ therefore need not increase with signal strength.

D.4.4 Matched second moments expose the different mechanisms

Consider the three mean-one laws

$$\begin{aligned} P_\gamma^{\text{balanced}} &= \tfrac{1}{2}\delta_{1-\sqrt{0.45}} + \tfrac{1}{2}\delta_{1+\sqrt{0.45}}, \\ P_\gamma^{\text{rare high}} &= \tfrac{320}{329}\delta_{71/80} + \tfrac{9}{329}\delta_{5}, \\ P_\gamma^{\text{very clean}} &= \tfrac{180}{541}\delta_{1/20} + \tfrac{361}{541}\delta_{28/19}. \end{aligned} \tag{D.154}$$

Each has $\kappa_2 = 29/20$, but the distributions of reliable and unreliable coordinates are different. The balanced law divides its mass equally between two noise levels. The second assigns a small fraction of coordinates a high variance. The third contains a substantial fraction of particularly clean coordinates.

Table 2: **Identical static thresholds, different dynamic adaptation.** Regular linear replica thresholds for the laws in (D.154) at $\beta = 1.5$, $\rho = 0.95$, and zero ridge. All three laws have unit mean and $\kappa_2 = 1.45$. The response ratio compares the lowest- and highest-variance coordinates using (D.143). Entries are computed numerically from the replica equations.

Variance law	α_c^{stat}	α_c^{dyn}	τ_c^{dyn}	$v_c(\gamma_{\min})/v_c(\gamma_{\max})$
Balanced	2.28272	0.52222	0.19380	1.24442
Rare high variance	2.28272	0.43843	0.26078	1.87089
Very clean component	2.28272	0.54817	0.17987	1.25380

Table 2 shows that matching the static threshold does not match dynamic learning. In this example, resampling is particularly effective for the rare-high-variance law, for which the response most strongly suppresses the noisy component. The common optimized regular static threshold is 1.28889, so the dynamic benefit also survives optimizing the static masking ratio. The comparison isolates dependence on the shape of P_γ beyond its second moment, rather than a change in average noise power or signal strength.

The entries are obtained from the linear critical equations, with numerical checks reported in App. H.12.

The rare-high-variance law even has $\alpha_c^{\mathrm{dyn}} = 0.43843 < 1/\beta^2 = 0.44444$, the homogeneous threshold at the same mean variance and signal strength. Conversely, the very clean component does not give the smallest dynamic threshold: the benefit depends on how the reconstruction objective uses the entire noise profile. Dynamic adaptation still differs from fully noise-aware inference, as quantified by the inverse-moment bounds in (D.133) and the Bayes comparison in App. H.4.

E PROOFS OF THE LINEAR RECONSTRUCTION SEPARATION

Assumption E.1. *We collect the precise hypotheses for Theorem 2.1. The activation is linear, $\sigma(h) = h$, and $\lambda_R = 0$. The noise law is fixed as the dimension grows; the independence assumptions are those of (1).*

(A1) (Model and scaling.) *Data are generated by (1) with $P_u = P_\lambda = \mathcal{N}(0,1)$; the signal strength $\beta > 0$, the masking ratio $\rho \in (0,1)$ and the sample complexity $\alpha = n/d$ are held fixed as $n, d \to \infty$.*

(A2) (Coordinate noise.) *P_γ is a mean-one distribution supported in $[a, \infty)$ for some $a > 0$, with $\kappa_2 = \mathbb{E}_\gamma[\gamma^2] < \infty$. We write $\gamma_+ = \operatorname{ess\,sup} \gamma$, allowing $\gamma_+ = \infty$.*

(A3) (Bounded failure conditions.) *In addition to (A1) and (A2), $\gamma_+ < \infty$ and*

$$\beta\, \mathbb{E}_\gamma \frac{1}{\gamma_+ - \gamma} < 1, \qquad \mathbb{E}_\gamma \frac{\gamma^2}{(\gamma_+ - \gamma)^2} = \infty. \tag{E.1}$$

We establish the equivalence of full linear reconstruction and PCA, prove the PCA obstructions, and give finite sufficient sample bounds for dynamic masking. The bounded-noise proof is followed by its extension to variance laws without an upper support bound.

E.1 PCA OBSTRUCTIONS AND BOUNDED-NOISE MASKING RECOVERY

This subsection establishes the PCA obstructions and bounded-noise masking recovery used in Theorem 2.1; App. E.2 extends masking recovery beyond bounded variance support. The PCA statements concern the leading empirical principal component. The masking statements concern every global minimizer of the exact mask-averaged objective, with no replica assumption and no claim about convergence of a particular optimization algorithm. The sufficient sample ratio is deliberately non-sharp.

E.1.1 MODEL, PCA REDUCTION, AND THE HOMOGENEOUS BENCHMARK

Use (B.3) with independent standard Gaussian $u_i^\star, \lambda_\mu, z_{\mu i}$, independent positive $\gamma_i^\star \overset{\text{iid}}{\sim} P_\gamma$, and $n/d \to \alpha \in (0, \infty)$. Define

$$\boldsymbol{e} = \frac{\boldsymbol{u}^\star}{\|\boldsymbol{u}^\star\|}, \qquad \beta_d = \frac{\beta \|\boldsymbol{u}^\star\|^2}{d}, \qquad \Sigma_d = \Gamma^\star + \beta_d \boldsymbol{e}\boldsymbol{e}^\top, \qquad \widehat{\Sigma} = \frac{1}{n}\sum_\mu \boldsymbol{x}_\mu \boldsymbol{x}_\mu^\top. \tag{E.2}$$

Conditionally on the quenched variables $\boldsymbol{u}^\star, \Gamma^\star$, the samples are independent $\mathcal{N}(0, \Sigma_d)$ vectors. With the sample-first matrices X, Z of App. B.1, write

$$Y := \frac{X}{\sqrt{n}} = N + \sqrt{\beta}\, \frac{\boldsymbol{\lambda}}{\sqrt{n}} \frac{(\boldsymbol{u}^\star)^\top}{\sqrt{d}}, \qquad N = \frac{1}{\sqrt{n}} Z (\Gamma^\star)^{1/2}. \tag{E.3}$$

Both Y and N are $n \times d$, and $\widehat{\Sigma} = Y^\top Y$. The PCA direction $\widehat{\boldsymbol{r}}_{\mathrm{PCA}}$ is the leading right singular vector of Y; its squared overlap is $|\boldsymbol{e}^\top \widehat{\boldsymbol{r}}_{\mathrm{PCA}}|^2$.

For a linear tied autoencoder, $\boldsymbol{w} = \sqrt{dQ}\, \boldsymbol{r}$, $\|\boldsymbol{r}\| = 1$, produces $\widehat{\boldsymbol{x}} = Q \boldsymbol{r}\boldsymbol{r}^\top \boldsymbol{x}$. Therefore (5), divided by n and with the target-only constant $n^{-1}\sum_\mu \|\boldsymbol{x}_\mu\|^2$ subtracted, is exactly

$$R_{\mathrm{full}}(Q, \boldsymbol{r}) = (Q^2 - 2Q) \boldsymbol{r}^\top \widehat{\Sigma} \boldsymbol{r}. \tag{E.4}$$

Since the largest sample eigenvalue is positive, the global minimum has $Q = 1$ and $\boldsymbol{r}$ in its top eigenspace. This proves the finite-dimensional equivalence with PCA, including the radial optimization.

For $P_\gamma = \delta_1$, the rectangular rank-one singular-vector transition (Benaych-Georges & Nadakuditi, 2012) gives

$$|\boldsymbol{e}^\top \widehat{\boldsymbol{r}}_{\mathrm{PCA}}|^2 \longrightarrow \begin{cases} 0, & \alpha\beta^2 \le 1, \\ \dfrac{1 - 1/(\alpha\beta^2)}{1 + 1/(\alpha\beta)}, & \alpha\beta^2 > 1, \end{cases} \tag{E.5}$$

in probability. App. D.3.3 derives the matching regular dynamic linear RS formula.

E.1.2 PCA FAILURE FOR UNBOUNDED COORDINATE NOISE

Proposition E.2 (Unbounded variance support)**.** *Under* (E.2)*, if P_γ has unbounded support and $\mathbb{E}_\gamma \gamma = 1$, then for every fixed $\beta > 0$ and $\alpha > 0$, $|\boldsymbol{e}^\top \widehat{\boldsymbol{r}}_{\mathrm{PCA}}|^2 \to 0$ in probability.*

Proof. Let i_d maximize $\gamma_i^\star$. Since P_γ is unbounded, $\gamma_{i_d}^\star \to \infty$ in probability. The selected column of Z is independent of this selection, and its squared norm divided by n converges in probability to one. Thus

$$\|N\|_{\mathrm{op}} \geq \sqrt{\gamma_{i_d}^\star}\, \frac{\|Z_{:,i_d}\|}{\sqrt{n}} \longrightarrow \infty. \tag{E.6}$$

The signal term in (E.3) has operator norm $\sqrt{\beta}\|\boldsymbol{u}^\star\|\|\boldsymbol{\lambda}\|/\sqrt{dn} = O_{\mathbb{P}}(1)$. Hence the top singular value $s_1(Y)$ also diverges in probability. On the other hand,

$$\mathbb{E}\|Y\boldsymbol{e}\|^2 = \mathbb{E}[\boldsymbol{e}^\top \Sigma_d \boldsymbol{e}] = 1 + \beta, \tag{E.7}$$

because $\boldsymbol{e}$ is independent of $\Gamma^\star$ and $\mathbb{E}\beta_d = \beta$. $\|Y\boldsymbol{e}\|$ is consequently bounded in probability. Let $\widehat{\boldsymbol{v}}_1 \in \mathbb{R}^n$ be the leading left singular vector paired with $\widehat{\boldsymbol{r}}_{\mathrm{PCA}}$. Then $Y^\top \widehat{\boldsymbol{v}}_1 = s_1(Y)\widehat{\boldsymbol{r}}_{\mathrm{PCA}}$ gives

$$|\boldsymbol{e}^\top \widehat{\boldsymbol{r}}_{\mathrm{PCA}}| = \frac{|\widehat{\boldsymbol{v}}_1^\top Y\boldsymbol{e}|}{s_1(Y)} \leq \frac{\|\widehat{\boldsymbol{v}}_1^\top\|\|Y\boldsymbol{e}\|}{s_1(Y)} \longrightarrow 0. \tag{E.8}$$

This argument needs no assumptions about localization of the top eigenvector and no finite second moment of γ. □

E.1.3 A BOUNDED-NOISE OBSTRUCTION AT EVERY SAMPLE RATIO

The following lemma identifies when the population covariance has an eigenvalue separated from the variance support.

Lemma E.3 (Population outlier equation)**.** *Under* (E.2)*, suppose P_γ is supported in $[a, \gamma_+] \subset (0, \infty)$, with γ_+ its essential upper endpoint. Let $m_+ = \mathbb{E}_\gamma[(\gamma_+ - \gamma)^{-1}] \in (0, \infty]$. If $\beta m_+ > 1$, the largest eigenvalue of Σ_d converges in probability to the unique $t > \gamma_+$ satisfying*

$$1 = \beta \mathbb{E}_\gamma \frac{1}{t - \gamma}. \tag{E.9}$$

If $\beta m_+ \leq 1$, the largest eigenvalue instead converges in probability to γ_+.

Proof. Write $t_d = \lambda_{\max}(\Sigma_d)$. Almost surely, $t_d > \max_i \gamma_i^\star$: the Rayleigh quotient at a coordinate maximizing $\gamma_i^\star$ is strictly larger than that maximum, since every Gaussian teacher coordinate is nonzero almost surely. For a corresponding eigenvector $\boldsymbol{v} \neq 0$,

$$(t_d \mathbf{I} - \Gamma^\star)\boldsymbol{v} = \beta_d \boldsymbol{e}(\boldsymbol{e}^\top \boldsymbol{v}).$$

The matrix on the left is invertible, and $\boldsymbol{e}^\top \boldsymbol{v} \neq 0$; otherwise $\boldsymbol{v} = 0$. Multiplying by $\boldsymbol{e}^\top (t_d \mathbf{I} - \Gamma^\star)^{-1}$ and cancelling gives the exact finite-dimensional equation

$$1 = \beta_d \sum_{i=1}^d \frac{e_i^2}{t_d - \gamma_i^\star} = \frac{\beta}{d} \sum_{i=1}^d \frac{(u_i^\star)^2}{t_d - \gamma_i^\star}.$$

For this proof, define the random function

$$F_d(t) = \frac{\beta}{d} \sum_{i=1}^d \frac{(u_i^\star)^2}{t - \gamma_i^\star}, \qquad t > \max_i \gamma_i^\star,$$

and its deterministic limit

$$F(t) = \beta \mathbb{E}_\gamma \frac{1}{t - \gamma}, \qquad t > \gamma_+.$$

Thus $F_d(t_d) = 1$, and F_d is almost surely strictly decreasing on its domain. For each fixed deterministic $t > \gamma_+$, independence and the law of large numbers yield

$$F_d(t) \xrightarrow{\mathbb{P}} F(t).$$

Indeed, the summands are i.i.d. and bounded above by $(u_i^\star)^2/(t-\gamma_+)$, which is integrable, and $\mathbb{E}[(u_i^\star)^2]=1$. The function F is continuous and strictly decreasing on (γ_+,∞), with limits βm_+ at the left endpoint and zero at infinity. Thus it equals one at a unique point in this interval exactly when $\beta m_+ > 1$.

It remains to pass from convergence at deterministic arguments to convergence of the random solution t_d of $F_d(t_d)=1$; the pointwise law of large numbers alone does not justify substituting the random argument t_d. Monotonicity allows us to locate this solution by evaluating F_d at deterministic arguments on either side of its proposed limit.

Suppose first that $\beta m_+ > 1$, and denote the unique solution of $F(t)=1$ by $t_\star > \gamma_+$. Fix $0<\varepsilon<t_\star-\gamma_+$ and set $t_-=t_\star-\varepsilon$ and $t_+=t_\star+\varepsilon$. These two deterministic arguments lie above γ_+ and satisfy

$$F(t_-) > 1 > F(t_+).$$

Convergence in probability at each of them implies $\Pr(F_d(t_-)\le 1)\to 0$ and $\Pr(F_d(t_+)\ge 1)\to 0$. Whenever $F_d(t_-) > 1 > F_d(t_+)$, strict monotonicity and $F_d(t_d)=1$ force $t_- < t_d < t_+$. Consequently, by a union bound,

$$\Pr\big(|t_d-t_\star|\ge\varepsilon\big)\le \Pr\big(F_d(t_-)\le 1\big)+\Pr\big(F_d(t_+)\ge 1\big)\longrightarrow 0.$$

Since this holds for every sufficiently small $\varepsilon>0$, it proves $t_d \xrightarrow{\mathbb{P}} t_\star$.

If $\beta m_+ \le 1$, then $F(\gamma_+ + \varepsilon) < 1$ for every $\varepsilon>0$, including when $\beta m_+ = 1$, by strict monotonicity. The law of large numbers and the same monotonicity argument give

$$\Pr(t_d \ge \gamma_+ + \varepsilon)\le \Pr\big(F_d(\gamma_+ + \varepsilon)\ge 1\big)\longrightarrow 0.$$

For the lower bound, $t_d \ge \max_i \gamma_i^\star$, and this maximum converges in probability to γ_+ because the variances are i.i.d. with essential upper endpoint γ_+. Combining these upper and lower bounds proves $t_d \xrightarrow{\mathbb{P}} \gamma_+$. □

We next provide the eigenvector argument needed when this condition fails.

Proposition E.4 (Compact variance laws with a regular upper sample edge)**.** *Suppose* $\mathbb{E}_\gamma \gamma = 1$, $\mathrm{supp}(P_\gamma)\subset[a,\gamma_+]$, γ_+ *is the upper support endpoint, and*

$$\beta\mathbb{E}_\gamma\frac{1}{\gamma_+-\gamma}<1, \qquad \mathbb{E}_\gamma\frac{\gamma^2}{(\gamma_+-\gamma)^2}=\infty. \tag{E.10}$$

Then PCA has zero asymptotic squared overlap with $\mathbf{e}$ *for every fixed* $\alpha\in(0,\infty)$.

Proof. The companion Stieltjes transform of $NN^\top$ satisfies the Silverstein–Pastur equation (Silverstein & Choi, 1995), with convention $\underline{m}(z)=\int(\lambda-z)^{-1}\underline{F}(\mathrm{d}\lambda)$:

$$\underline{m}(z)=-\left(z-\frac{1}{\alpha}\mathbb{E}_\gamma\frac{\gamma}{1+\gamma\underline{m}(z)}\right)^{-1}. \tag{E.11}$$

On the upper outside-support branch, setting $t=-1/\underline{m}(z)>\gamma_+$ gives

$$z=\psi_\alpha(t):=t\left(1+\frac{1}{\alpha}\mathbb{E}_\gamma\frac{\gamma}{t-\gamma}\right),\quad \psi_\alpha'(t)=1-\frac{1}{\alpha}\mathbb{E}_\gamma\frac{\gamma^2}{(t-\gamma)^2}. \tag{E.12}$$

The second condition in (E.10) ensures a unique $t_e>\gamma_+$ with $\psi_\alpha'(t_e)=0$ for every finite $\alpha>0$. Furthermore $\psi_\alpha''(t_e)=2\alpha^{-1}\mathbb{E}_\gamma[\gamma^2/(t_e-\gamma)^3]>0$. The support characterization of Silverstein & Choi (1995) then gives a regular upper covariance edge $b_\alpha^2=\psi_\alpha(t_e)$ with a square-root density. The empirical noise spectrum converges to this law, and $s_1(N)\to b_\alpha$: the bounded population variances have no outliers beyond their support, so the no-eigenvalues-outside-the-support theorem applies (Bai & Silverstein, 1998). A bounded Gaussian matrix norm controls the remaining unbounded part of the spectral axis.

For a singular-value argument $s>b_\alpha$, let $g_d(s)$ and $g_n(s)$ be the limits of the normalized traces of $s(s^2I_d-N^\top N)^{-1}$ and $s(s^2I_n-NN^\top)^{-1}$, respectively. Their product is the rectangular D-transform in Benaych-Georges & Nadakuditi (2012); using both traces automatically includes the zero singular values on the larger side. At $s^2=\psi_\alpha(t)$,

$$g_d(s)=\frac{t}{s}\mathbb{E}_\gamma\frac{1}{t-\gamma},\qquad g_n(s)=\frac{s}{t},\qquad D_\alpha(s)=g_d(s)g_n(s)=\mathbb{E}_\gamma\frac{1}{t-\gamma}. \tag{E.13}$$

In particular,

$$\beta D_\alpha(b_\alpha^+) = \beta \mathbb{E}_\gamma \frac{1}{t_e - \gamma} < \beta m_+ < 1. \tag{E.14}$$

The perturbation in (E.3) is exactly the independent Gaussian rank-one i.i.d. spike model of Benaych-Georges & Nadakuditi (2012), with strength $\sqrt{\beta}$. Its singular-vector transition theorem gives vanishing projections onto the spike vectors below the threshold, provided the singular resolvent derivative diverges at the edge. The square-root density just verified implies this derivative condition. Both left and right projections vanish, so the conclusion also holds when $\alpha > 1$ and the matrix is transposed to place the smaller dimension first. Normalizing $\boldsymbol{u}^\star/\sqrt{d}$ to $\boldsymbol{e}$ changes the overlap by a factor converging to one. This proves the eigenvector conclusion, rather than merely absence of a separated eigenvalue. □

For reference, if (E.9) has a root $t > \gamma_+$, define $J_j = \mathbb{E}_\gamma[\gamma^j/(t-\gamma)^2]$ and $B = \mathbb{E}_\gamma[\gamma/(t-\gamma)]$. On the regular separated sample branch, (E.12) requires $\alpha > J_2$, and the same singular-vector formula gives

$$\alpha_c^{\mathrm{PCA}} = J_2, \qquad \theta^2_{\boldsymbol{u}^\star,\mathrm{PCA}} = \frac{1 - J_2/\alpha}{\beta^2 J_0(1 + B/\alpha)}, \qquad \alpha > J_2. \tag{E.15}$$

Explicitly, that formula is $-2g_d(s)/(\beta D'_\alpha(s))$ at the outlier. Using $D'_\alpha(s) = -2sJ_0/\psi'_\alpha(t)$ and $g_d(s) = t/(\beta s)$ gives (E.15). Substitution of $P_\gamma = \delta_1$, $t = 1 + \beta$ recovers (E.5).

E.1.4 Exact mask average and radial reduction

For the masking proof assume only that P_γ is supported in $[a, \gamma_+] \subset (0, \infty)$ and has mean one. Fix $\rho \in (0,1)$. For a generic mask, suppress the sample and view indices and write $\chi_i \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}(\rho)$, with $\chi_i = 1$ marking a hidden coordinate. Define $\boldsymbol{r}_{\mathrm{vis}} = ((1-\chi_i)r_i)_i$, $\boldsymbol{r}_{\mathrm{mask}} = (\chi_i r_i)_i$, and $m_{\mathrm{mask}} = \|\boldsymbol{r}_{\mathrm{mask}}\|^2$. The hidden predictions from the linear encoder are $Q\boldsymbol{r}_{\mathrm{mask}}(\boldsymbol{r}_{\mathrm{vis}}^\top \boldsymbol{x})$. Expanding the squared loss and averaging over masks gives exactly

$$\begin{aligned} \widehat{R}_{\mathrm{dyn}}(Q, \boldsymbol{r}) &= \widehat{H}(\boldsymbol{r})Q^2 - 2\widehat{\mathcal{C}}(\boldsymbol{r})Q, \\ \widehat{H}(\boldsymbol{r}) &= \mathbb{E}_\chi[m_{\mathrm{mask}} \boldsymbol{r}_{\mathrm{vis}}^\top \widehat{\Sigma} \boldsymbol{r}_{\mathrm{vis}}], \qquad \widehat{\mathcal{C}}(\boldsymbol{r}) = \mathbb{E}_\chi[\boldsymbol{r}_{\mathrm{vis}}^\top \widehat{\Sigma} \boldsymbol{r}_{\mathrm{mask}}], \end{aligned} \tag{E.16}$$

after subtracting the mask-averaged target-only constant. For $\widehat{H} > 0$, optimizing $Q \geq 0$ yields

$$Q_*(\boldsymbol{r}) = \frac{[\widehat{\mathcal{C}}(\boldsymbol{r})]_+}{\widehat{H}(\boldsymbol{r})}, \qquad \inf_{Q \geq 0} \widehat{R}_{\mathrm{dyn}}(Q, \boldsymbol{r}) = -\widehat{S}(\boldsymbol{r}), \qquad \widehat{S}(\boldsymbol{r}) = \frac{[\widehat{\mathcal{C}}(\boldsymbol{r})]_+^2}{\widehat{H}(\boldsymbol{r})}. \tag{E.17}$$

Coordinate directions have $\widehat{H} = \widehat{\mathcal{C}} = 0$ and score zero. On the positive-definite covariance event used below, these are the only directions with zero denominator. No replacement of a mask mass by its mean is used in this reduction.

E.1.5 Population score gap, including localized directions

Let $H, \mathcal{C}, S$ be the conditional population versions at fixed $\boldsymbol{u}^\star, \Gamma^\star$, obtained by replacing $\widehat{\Sigma}$ with Σ_d. Set

$$c = \boldsymbol{e}^\top \boldsymbol{r}, \quad I = \sum_i r_i^4, \quad J = \sum_i r_i^2 e_i^2, \quad \delta_d = \max_i e_i^2. \tag{E.18}$$

The disjoint visible and hidden sets make $\boldsymbol{r}_{\mathrm{vis}}^\top \Gamma^\star \boldsymbol{r}_{\mathrm{mask}} = 0$ exactly. Thus

$$\mathcal{C}(\boldsymbol{r}) = \rho(1-\rho)\beta_d(c^2 - J), \qquad H(\boldsymbol{r}) \geq a\mathbb{E}_\chi[m_{\mathrm{mask}}(1 - m_{\mathrm{mask}})] = a\rho(1-\rho)(1-I). \tag{E.19}$$

The first identity isolates the cross-coordinate signal; the noise remains in the positive quartic coefficient H.

Lemma E.5 (Control of localization)**.** *For unit vectors $\boldsymbol{r}, \boldsymbol{e}$, with c, I, J as above and $\delta = \max_i e_i^2$,*

$$\frac{[c^2 - J]_+^2}{1 - I} \leq 8(c^2 + \delta), \tag{E.20}$$

where the left side is defined to be zero at coordinate directions.

Proof. Only $D_1 = c^2 - J > 0$ needs consideration. If $m = \max_i r_i^2 \le 1/2$, then $I \le 1/2$ and $D_1^2/(1-I) \le 2c^4 \le 2c^2$. Otherwise let $r_k^2 = m > 1/2$ and $R = 1 - m$. Then $1 - I \ge 2mR \ge R$. Writing $z = r_k e_k$, $y = \sum_{i \ne k} r_i e_i$ gives $D_1 = 2zy + y^2 - \sum_{i \ne k} r_i^2 e_i^2$ and $D_1 \le 2\sqrt{\delta R} + R$. If $R \ge c^2$, use $D_1 \le c^2$ to obtain $D_1^2/(1-I) \le c^2$. If $R < c^2$, use $D_1^2/(1-I) \le (2\sqrt{\delta} + \sqrt{R})^2 \le 8\delta + 2R \le 8(\delta + c^2)$. At $R = 0$ the direction is a coordinate vector and $D_1 = 0$. □

Combining (E.19) and (E.20) gives the uniform estimate

$$S(\boldsymbol{r}) \le \frac{8\beta_d^2 \rho(1-\rho)}{a}(c^2 + \delta_d). \tag{E.21}$$

For the Gaussian teacher, $\beta_d \to \beta$, $\delta_d \to 0$, and $\boldsymbol{e}^\top \Gamma^\star \boldsymbol{e} \to 1$ in probability. The teacher score is bounded away from zero: its cross coefficient tends to $\rho(1-\rho)\beta$, and its quartic coefficient has noise part

$$\rho(1-\rho)\left(\boldsymbol{e}^\top \Gamma^\star \boldsymbol{e} - \sum_i \gamma_i^\star e_i^4\right) \to \rho(1-\rho). \tag{E.22}$$

For the signal part, $\boldsymbol{e}_{\mathrm{vis}}^\top \boldsymbol{e} = 1 - m_{\mathrm{mask}}$ and $\mathrm{Var}_\chi(m_{\mathrm{mask}}) = \rho(1-\rho)\sum_i e_i^4 \to 0$, so that $\beta_d \mathbb{E}_\chi[m_{\mathrm{mask}}(1-m_{\mathrm{mask}})^2] \to \beta\rho(1-\rho)^2$. Consequently

$$S(\boldsymbol{e}) \longrightarrow g := \frac{\rho(1-\rho)\beta^2}{1 + \beta(1-\rho)} > 0. \tag{E.23}$$

E.1.6 Uniform empirical comparison and an explicit sample bound

Define the relative covariance error

$$\varepsilon_d = \|\Sigma_d^{-1/2} \widehat{\Sigma} \Sigma_d^{-1/2} - \mathbf{I}\|_{\mathrm{op}}. \tag{E.24}$$

Conditionally on the teacher and variances, the whitened covariance is standard Gaussian Wishart. Its extreme eigenvalue limits (Bai & Yin, 1993) give, for $\alpha > 1$,

$$\varepsilon_d \longrightarrow \varepsilon_\alpha = 2/\sqrt{\alpha} + 1/\alpha \quad \text{in probability.} \tag{E.25}$$

On $\varepsilon_d < 1$, the relative quadratic-form bound immediately gives

$$(1 - \varepsilon_d) H(\boldsymbol{r}) \le \widehat{H}(\boldsymbol{r}) \le (1 + \varepsilon_d) H(\boldsymbol{r}). \tag{E.26}$$

For $\Delta_\Sigma = \widehat{\Sigma} - \Sigma_d$, the corresponding bilinear inequality is

$$|\boldsymbol{x}^\top \Delta_\Sigma \boldsymbol{y}| \le \varepsilon_d \sqrt{\boldsymbol{x}^\top \Sigma_d \boldsymbol{x}} \sqrt{\boldsymbol{y}^\top \Sigma_d \boldsymbol{y}}. \tag{E.27}$$

Apply it to $\boldsymbol{r}_{\mathrm{vis}}, \boldsymbol{r}_{\mathrm{mask}}$. For $m_{\mathrm{mask}} > 0$, rewrite the product as $\sqrt{m_{\mathrm{mask}} \boldsymbol{r}_{\mathrm{vis}}^\top \Sigma_d \boldsymbol{r}_{\mathrm{vis}}} \sqrt{\boldsymbol{r}_{\mathrm{mask}}^\top \Sigma_d \boldsymbol{r}_{\mathrm{mask}} / m_{\mathrm{mask}}}$. The second factor is at most $\sqrt{\|\Sigma_d\|_{\mathrm{op}}}$; when $m_{\mathrm{mask}} = 0$ the bilinear form is zero. Cauchy–Schwarz over masks therefore yields

$$|\widehat{\mathcal{C}}(\boldsymbol{r}) - \mathcal{C}(\boldsymbol{r})| \le \varepsilon_d \sqrt{H(\boldsymbol{r}) \|\Sigma_d\|_{\mathrm{op}}}. \tag{E.28}$$

These estimates hold simultaneously for all directions and imply

$$\begin{aligned} \sqrt{\widehat{S}(\boldsymbol{r})} &\le \frac{\sqrt{S(\boldsymbol{r})} + \varepsilon_d \sqrt{\|\Sigma_d\|_{\mathrm{op}}}}{\sqrt{1 - \varepsilon_d}}, \\ \sqrt{\widehat{S}(\boldsymbol{r})} &\ge \frac{\sqrt{S(\boldsymbol{r})} - \varepsilon_d \sqrt{\|\Sigma_d\|_{\mathrm{op}}}}{\sqrt{1 + \varepsilon_d}} \quad \text{if } \mathcal{C}(\boldsymbol{r}) > 0 \text{ and the numerator is positive.} \end{aligned} \tag{E.29}$$

Theorem E.6 (Global dynamic linear recovery at finite sample ratio)**.** *For every mean-one law supported in $[a, \gamma_+] \subset (0, \infty)$, fixed $\beta > 0$, and $\rho \in (0,1)$, define*

$$\begin{aligned} g &= \frac{\rho(1-\rho)\beta^2}{1 + \beta(1-\rho)}, & \eta &= \frac{a}{512[1 + \beta(1-\rho)]}, \\ \varepsilon_* &= \min\left\{\frac{1}{4}, \frac{\sqrt{g}}{16\sqrt{\gamma_+ + \beta + 1}}\right\}, & A_{\mathrm{suff}} &= \frac{1}{(\sqrt{1+\varepsilon_*} - 1)^2}. \end{aligned} \tag{E.30}$$

If $n/d \to \alpha > A_{\mathrm{suff}}$, then with probability tending to one the exact ridgeless dynamic linear objective has a global minimizer, and every global minimizer has $|\boldsymbol{e}^\top \widehat{\boldsymbol{r}}|^2 \ge \eta$.

Proof. Put $C_0 = 8\beta^2\rho(1-\rho)/a$, so that $\eta = g/(64C_0)$. The population estimates imply, with probability tending to one,

$$\sup_{c^2\le\eta} S(\boldsymbol{r}) \le g/16, \qquad S(\boldsymbol{e}) \ge 3g/4, \qquad \|\Sigma_d\|_{\mathrm{op}} \le \gamma_+ + \beta + 1. \tag{E.31}$$

Solving $2/\sqrt{\alpha} + 1/\alpha < \varepsilon_*$ gives exactly $\alpha > A_{\mathrm{suff}}$. By (E.25), on an event of probability tending to one, $\varepsilon_d < \varepsilon_*$ as well. The uniform comparisons then yield

$$\sup_{c^2\le\eta} \sqrt{\widehat{S}(\boldsymbol{r})} \le \frac{(1/4+1/16)\sqrt{g}}{\sqrt{3/4}} < \frac{(\sqrt{3/4}-1/16)\sqrt{g}}{\sqrt{5/4}} \le \sqrt{\widehat{S}(\boldsymbol{e})}. \tag{E.32}$$

Thus no maximizing direction can have squared overlap at most η.

To verify attainment, for large d every coordinate direction has squared teacher overlap at most $\delta_d < \eta$. On the compact set $\{\|\boldsymbol{r}\| = 1 : c^2 \ge \eta\}$ the positive lower bound for H and the relative comparison ensure $\widehat{H} > 0$. The score is continuous there and attains its maximum. The strict gap (E.32) makes this a maximum over all directions. Its score is positive, so its radial optimum in (E.17) is finite and positive. The zero vector and coordinate directions cannot be global minimizers. The radial identity proves the conclusion for every global weight minimizer. □

This proof uses only a population score gap and relative covariance control. It does not require the exact population maximizing direction, uniqueness of a resolvent equation for that direction, or a uniform delocalization assumption on candidate empirical weights. The factor $\|\Sigma_d\|_{\mathrm{op}}$ in (E.29) explains why this sufficient bound does not extend directly to unbounded variances. App. E.2 removes that factor by normalizing the off-diagonal error with the actual masking penalty, and proves recovery for variance laws bounded away from zero with finite second moment.

Proof of Theorem 2.1. Under (A1) and (A2), Theorem E.7 applies whether or not the variance support is bounded. It gives existence of global minimizers and the positive-overlap guarantee with constants depending only on a, β, ρ, as stated in the main theorem. For bounded support, Theorem E.6 also supplies the alternative sufficient constants in (E.30).

If the variance support is unbounded, Proposition E.2 gives zero PCA overlap for every fixed $\alpha > 0$. If instead the bounded-noise conditions (A3) hold, Proposition E.4 gives the same conclusion. The exact identity (E.4) transfers this failure to full linear reconstruction. Combining the applicable PCA obstruction with masked recovery proves the separation on the same Gaussian data model and variance law. □

E.1.7 Examples within the bounded-support class

Choose P_γ from the family

$$P_\gamma(\mathrm{d}\gamma) = \frac{r+1}{(\gamma_+ - a)^{r+1}}(\gamma_+ - \gamma)^r \mathbf{1}_{\{a<\gamma<\gamma_+\}}\,\mathrm{d}\gamma, \qquad 0 < a < 1, \quad 0 < r \le 1, \qquad \gamma_+ = (r+2) - (r+1)a. \tag{E.33}$$

Direct integration gives $\mathbb{E}_\gamma\gamma = a + (\gamma_+ - a)/(r+2) = 1$ and

$$m_+ = \frac{r+1}{r(\gamma_+ - a)}, \qquad \mathbb{E}_\gamma \frac{\gamma^2}{(\gamma_+ - \gamma)^2} = \infty. \tag{E.34}$$

The latter follows from the endpoint power $(\gamma_+ - \gamma)^{r-2}$. Consequently, if

$$0 < \beta < \frac{r(r+2)(1-a)}{r+1}, \tag{E.35}$$

Proposition E.4 proves PCA failure at every fixed sample ratio, while Theorem E.6 proves dynamic masking recovery for every fixed $\rho \in (0,1)$ at sufficiently large finite sample ratio. This is a class result; no particular density is needed for the dynamic theorem itself.

For a concrete numerical illustration, take $a = 1/4$, $r = 1$, $\gamma_+ = 5/2$, and $\beta = 1$. The resulting law is

$$P_\gamma(\mathrm{d}\gamma) = \frac{32}{81}\left(\frac{5}{2} - \gamma\right)\mathbf{1}_{\{1/4<\gamma<5/2\}}\,\mathrm{d}\gamma. \tag{E.36}$$

This variance law is the bounded-noise example in Figure 2(b); the general theorem does not depend on this choice. Its moments are

$$\mathbb{E}_\gamma \gamma = 1, \qquad \kappa_2 = 41/32, \qquad \kappa_3 = 617/320, \qquad m_+ = 8/9. \tag{E.37}$$

The strict signal condition is $1 < 9/8$, so this law is one instance of Theorem 2.1. For an entirely explicit sufficient choice, take $\rho = 1/2$. Then $g = 1/6$, $\eta = 1/3072$, and $\varepsilon_* = 1/(48\sqrt{3})$. Every $\alpha \geq 30000$ satisfies

$$\varepsilon_\alpha \leq \frac{1}{50\sqrt{3}} + \frac{1}{30000} < \frac{1}{48\sqrt{3}} = \varepsilon_*. \tag{E.38}$$

These conservative constants certify a finite proportional sample size.

The failure criterion is not universal over heterogeneous laws. For example, an atom at the upper endpoint makes $m_+ = \infty$, so a mean-one two-point law has a population outlier for every $\beta > 0$ and PCA eventually recovers as α increases. The separation concerns ordinary PCA and the specified full-reconstruction architecture, rather than spectral estimators designed to correct heterogeneous noise.

E.2 Rigorous dynamic linear recovery with unbounded coordinate noise

The upper bound on the variance support in Theorem E.6 can be removed. The essential improvement is to control the off-diagonal sampling error relative to the noise penalty already present in the exact masking objective. This avoids the factor $\|\Sigma_d\|_{\mathrm{op}}$ in (E.29). All results in this subsection concern the original tied linear architecture and the exact mask-averaged empirical objective; no clipping, reweighting, or restriction on the candidate weights is introduced.

E.2.1 Statement and assumptions

Use the Gaussian model and notation of (E.2). Suppose

$$\gamma_i^\star \overset{\text{iid}}{\sim} P_\gamma, \qquad P_\gamma([a,\infty)) = 1 \quad \text{for some } a > 0, \qquad \mathbb{E}_\gamma \gamma = 1, \qquad \mathbb{E}_\gamma \gamma^2 < \infty. \tag{E.39}$$

The variance law is fixed as d grows, is independent of the Gaussian teacher and sample variables, and may have unbounded support. Fix $\beta > 0$, $\rho \in (0,1)$, and $\lambda_R = 0$.

Theorem E.7 (Recovery without an upper variance bound)**.** *Under (E.39), there is a finite deterministic $A(a,\beta,\rho)$ such that, whenever $n/d \to \alpha > A(a,\beta,\rho)$, with probability tending to one the exact dynamic linear objective has a global minimizer and every global minimizer satisfies*

$$\theta_{\boldsymbol{u}^\star}^2 \geq \eta := \frac{a}{512[1+\beta(1-\rho)]} > 0. \tag{E.40}$$

If P_γ also has unbounded support, full linear reconstruction and PCA have squared overlap converging to zero for every fixed $\alpha > 0$ and $\beta > 0$.

The second assertion follows from Proposition E.2. The first is proved below. The sufficient sample ratio is non-sharp and the result does not assert convergence of an optimization algorithm. Neither the law nor its support needs to change with α, β, or ρ. The constants in the asymptotic sufficient bound below do not depend on an upper support endpoint; the dimension required to reach its high-probability regime can depend on the variance law.

E.2.2 A variance-sensitive lower bound on the score denominator

Recall the exact radial reduction in (E.16)–(E.17): writing $\boldsymbol{w} = \sqrt{dQ}\,\boldsymbol{r}$ with $\|\boldsymbol{r}\| = 1$ and minimizing over $Q \geq 0$ reduces the recovery problem to maximizing the directional score

$$\widehat{S}(\boldsymbol{r}) = \frac{[\widehat{\mathcal{C}}(\boldsymbol{r})]_+^2}{\widehat{H}(\boldsymbol{r})}.$$

Here $\widehat{\mathcal{C}}$ and $\widehat{H}$ are the empirical coefficients defined in (E.16); their conditional population counterparts $\mathcal{C}, H$ are obtained by replacing $\widehat{\Sigma}$ by Σ_d, with the teacher and variances held fixed. The score is

defined to be zero at coordinate directions. These exact identities do not require an upper bound on the coordinate variances.

To transfer the population score gap to the empirical score, we will control $\widehat{H}$ relative to H and the numerator error $|\widehat{\mathcal{C}} - \mathcal{C}|/\sqrt{H}$; the square root appears because we compare $\sqrt{\widehat{S}}$ with $\sqrt{S}$. The purpose of this subsection is to retain a lower bound on H that grows with the variances in the candidate direction. This will allow large sampling errors to be controlled relative to the score denominator.

Write $\mathrm{off}(M) = M - \mathrm{diag}(M)$. For a unit direction $\boldsymbol{r}$, let

$$D_\Gamma(\boldsymbol{r}) := \sum_i \gamma_i^\star r_i^2(1 - r_i^2), \qquad I(\boldsymbol{r}) := \sum_i r_i^4. \tag{E.41}$$

This noise penalty satisfies $D_\Gamma \geq a(1 - I)$ and is positive except at coordinate directions. The exact mask identities give

$$\widehat{\mathcal{C}}(\boldsymbol{r}) = \rho(1-\rho)\, \boldsymbol{r}^\top \mathrm{off}(\widehat{\Sigma})\boldsymbol{r}, \qquad H(\boldsymbol{r}) \geq \rho(1-\rho) D_\Gamma(\boldsymbol{r}). \tag{E.42}$$

The second inequality retains the actual coordinate variances rather than replacing them by their common lower bound. Its noise contribution is exactly $\rho(1-\rho)D_\Gamma$; the signal contribution to H is nonnegative. Thus D_Γ is a lower bound on the population score denominator, up to the factor $\rho(1-\rho)$, rather than the denominator itself. It suffices to bound the off-diagonal covariance error after division by $\sqrt{D_\Gamma(\boldsymbol{r})}$. The next subsection establishes this bound for the noise contribution; (E.62) then gives the required control of $|\widehat{\mathcal{C}} - \mathcal{C}|/\sqrt{H}$ for the full covariance.

E.2.3 Uniform control of the noise sampling error

We first study the noise-noise part of the off-diagonal covariance error. Use the $n \times d$ matrix $Z = (z_{\mu i})$ from App. B.1 and set

$$E_d := \mathrm{off}\Big((\Gamma^\star)^{1/2} Z^\top Z (\Gamma^\star)^{1/2}/n\Big), \qquad \gamma_{+,d} = \max_i \gamma_i^\star, \qquad s_d = \frac{\mathrm{tr}\,\Gamma^\star}{n}. \tag{E.43}$$

Here $\gamma_{+,d}$ is the finite-sample maximum, distinct from the deterministic support endpoint γ_+, which may be infinite. The population off-diagonal noise covariance is zero. Finite second moment and the law of large numbers imply

$$s_d \to \alpha^{-1}, \qquad \mathrm{tr}[(\Gamma^\star)^2] = O_{\mathbb{P}}(d), \qquad \gamma_{+,d} = o_{\mathbb{P}}(\sqrt{d}), \qquad \frac{\gamma_{+,d} \log d}{n} \to 0. \tag{E.44}$$

For the third assertion, for each fixed $\varepsilon > 0$, $\Pr(\gamma_{+,d} > \varepsilon\sqrt{d}) \leq d \Pr(\gamma > \varepsilon\sqrt{d}) \to 0$, since $t^2 \Pr(\gamma > t) \to 0$ when $\mathbb{E}_\gamma \gamma^2 < \infty$.

Lemma E.8 (Noise error normalized by the masking penalty)**.** *There is an absolute finite constant C_1 such that, for each fixed $\alpha > 1$,*

$$\sup_{\|\boldsymbol{r}\|=1,\ D_\Gamma(\boldsymbol{r})>0} \frac{|\boldsymbol{r}^\top E_d \boldsymbol{r}|}{\sqrt{D_\Gamma(\boldsymbol{r})}} \leq C_1 \left(\alpha^{-1/2} + \frac{\alpha^{-1}}{\sqrt{a}}\right) + o_{\mathbb{P}}(1). \tag{E.45}$$

The supremum includes data-dependent and localized directions.

Proof. We first establish a bound using $\boldsymbol{r}^\top \Gamma^\star \boldsymbol{r}$ and then treat directions near a coordinate axis separately.

A uniform quadratic-form bound. For $t > 0$, define the diagonal matrices

$$P_t = (\Gamma^\star + t\mathbf{I})^{-1/2}, \qquad A_t = P_t \Gamma^\star P_t, \qquad \|A_t\|_{\mathrm{op}} \leq 1, \quad \mathrm{tr}\, A_t \leq \mathrm{tr}\,\Gamma^\star/t. \tag{E.46}$$

Fix $t > 0$ and condition on $\Gamma^\star$. For $\boldsymbol{z}_\mu = (z_{\mu 1}, \ldots, z_{\mu d})^\top$, the transformed samples $\boldsymbol{y}_\mu^{(t)} = P_t(\Gamma^\star)^{1/2}\boldsymbol{z}_\mu$ are conditionally i.i.d. $\mathcal{N}(0, A_t)$. Their sample covariance $\widehat{A}_t = n^{-1}\sum_\mu \boldsymbol{y}_\mu^{(t)}(\boldsymbol{y}_\mu^{(t)})^\top$ therefore satisfies

$$\widehat{A}_t - A_t = P_t \left((\Gamma^\star)^{1/2} \frac{Z^\top Z}{n} (\Gamma^\star)^{1/2} - \Gamma^\star\right) P_t. \tag{E.47}$$

By Jensen's inequality, $\big(\mathbb{E}[\|\boldsymbol{y}_\mu^{(t)}\| \mid \Gamma^\star]\big)^2 \le \mathbb{E}[\|\boldsymbol{y}_\mu^{(t)}\|^2 \mid \Gamma^\star] = \operatorname{tr} A_t$. Thus the effective rank in Corollary 2 of Koltchinskii & Lounici (2017) is at most $\operatorname{tr}(A_t)/\|A_t\|_{\mathrm{op}}$. Applying that corollary and bounding its maximum by a sum gives, for every confidence parameter $\omega \ge 1$, with conditional probability at least $1 - e^{-\omega}$,

$$\begin{aligned}\|\widehat{A}_t - A_t\|_{\mathrm{op}} \le C\Big[&\sqrt{\|A_t\|_{\mathrm{op}} \operatorname{tr}(A_t)/n} + \operatorname{tr}(A_t)/n \\ &+ \|A_t\|_{\mathrm{op}}\sqrt{\omega/n} + \|A_t\|_{\mathrm{op}}\omega/n\Big].\end{aligned} \tag{E.48}$$

Here C is universal. Using the bounds on $\|A_t\|_{\mathrm{op}}$ and $\operatorname{tr} A_t$ above, together with (E.47), yields

$$\begin{aligned}&\left\| P_t \left((\Gamma^\star)^{1/2} Z^\top Z (\Gamma^\star)^{1/2}/n - \Gamma^\star \right) P_t \right\|_{\mathrm{op}} \\ &\qquad \le C\left[\sqrt{s_d/t} + s_d/t + \sqrt{\omega/n} + \omega/n\right].\end{aligned} \tag{E.49}$$

For a symmetric matrix M, $\|\operatorname{off}(M)\|_{\mathrm{op}} \le 2\|M\|_{\mathrm{op}}$. Because P_t and $\Gamma^\star$ are diagonal, $\operatorname{off}(\widehat{A}_t - A_t) = P_t E_d P_t$, so the same bound with $2C$ controls $\|P_t E_d P_t\|_{\mathrm{op}}$.

To choose the preconditioning scale according to a possibly data-dependent direction, we first make the bound simultaneous over a finite grid. Use the grid $t_j = a2^j$, from a up to its first value at least $\gamma_{+,d}$, and take $\omega = 4\log d$. On $\gamma_{+,d} \le d$, which has probability tending to one, this grid has $O(\log d)$ points. The grid depends only on $\Gamma^\star$, which is fixed under the conditioning; it does not depend on Z or on the candidate direction. Its conditional failure probability is at most $O(\log d)d^{-4}$. A conditional union bound makes (E.49) simultaneous on the grid with probability tending to one. Averaging over $\Gamma^\star$ and adding $\Pr(\gamma_{+,d} > d) = o(1)$ gives the same unconditional conclusion. On this simultaneous event, for any unit $\boldsymbol{r}$, put $g_{\boldsymbol{r}} = \boldsymbol{r}^\top \Gamma^\star \boldsymbol{r} \in [a, \gamma_{+,d}]$ and choose a grid point $g_{\boldsymbol{r}} \le t_j \le 2g_{\boldsymbol{r}}$. Then $\boldsymbol{r}^\top(\Gamma^\star + t_j \mathbf{I})\boldsymbol{r} \le 3g_{\boldsymbol{r}}$, and therefore

$$\frac{|\boldsymbol{r}^\top E_d \boldsymbol{r}|}{\sqrt{g_{\boldsymbol{r}}}} \le 6C\left[\sqrt{s_d} + \frac{s_d}{\sqrt{a}} + \sqrt{\gamma_{+,d}}\left(\sqrt{\omega/n} + \omega/n\right)\right]. \tag{E.50}$$

By (E.44), the last term vanishes in probability. Consequently, with

$$T_d := \sup_{\|\boldsymbol{r}\|=1} \frac{|\boldsymbol{r}^\top E_d \boldsymbol{r}|}{\sqrt{\boldsymbol{r}^\top \Gamma^\star \boldsymbol{r}}}, \tag{E.51}$$

we have $T_d \le 6C(\alpha^{-1/2} + a^{-1/2}\alpha^{-1}) + o_{\mathbb{P}}(1)$. By homogeneity, every vector $\boldsymbol{v}$ obeys

$$|\boldsymbol{v}^\top E_d \boldsymbol{v}| \le T_d \|\boldsymbol{v}\| \sqrt{\boldsymbol{v}^\top \Gamma^\star \boldsymbol{v}}. \tag{E.52}$$

A simultaneous row bound. Let $\boldsymbol{f}_i$ denote coordinate vector i and let $\boldsymbol{z}_i = Z_{:,i} \in \mathbb{R}^n$ be column i of Z. Conditional on $\Gamma^\star, \boldsymbol{z}_i$, the variables $\boldsymbol{z}_i^\top \boldsymbol{z}_j/\|\boldsymbol{z}_i\|$, $j \ne i$, are independent standard Gaussians. Hence

$$\frac{\|E_d \boldsymbol{f}_i\|^2}{\gamma_i^\star} \overset{\text{law}}{=} \frac{\|\boldsymbol{z}_i\|^2}{n} \frac{1}{n} \sum_{j \ne i} \gamma_j^\star \xi_j^2, \tag{E.53}$$

where the equality in law is conditional on $\Gamma^\star, \boldsymbol{z}_i$. For independent standard Gaussians ξ_j, their moment-generating function gives

$$\Pr\left\{ \sum_j \gamma_j^\star(\xi_j^2 - 1) > 2\sqrt{\omega \sum_j (\gamma_j^\star)^2} + 2\gamma_{+,d}\omega \,\middle|\, \Gamma^\star \right\} \le e^{-\omega}. \tag{E.54}$$

Indeed, for $0 < s < (2\gamma_{+,d})^{-1}$, the logarithm of the centered moment-generating function is at most $s^2 \sum_j (\gamma_j^\star)^2/(1 - 2s\gamma_{+,d})$; Chernoff's inequality gives the displayed bound. Apply it to each row with $\omega = 4\log d$ and combine it with the analogous bound for $\|\boldsymbol{z}_i\|^2$. With conditional probability at least $1 - 2de^{-\omega}$, simultaneously for all rows,

$$W_d^2 \le \left(1 + 2\sqrt{\omega/n} + 2\omega/n\right)\left(s_d + \frac{2\sqrt{\omega \operatorname{tr}[(\Gamma^\star)^2]}}{n} + \frac{2\gamma_{+,d}\omega}{n}\right). \tag{E.55}$$

Here W_d is defined in the next display. This conditional union bound and (E.44) yield

$$W_d := \max_i \frac{\|E_d \boldsymbol{f}_i\|}{\sqrt{\gamma_i^\star}} \le \alpha^{-1/2} + o_{\mathbb{P}}(1). \tag{E.56}$$

No independence between different rows' events is required.

Including directions concentrated on a coordinate. If $\max_i r_i^2 \le 1/2$, then $D_\Gamma(\boldsymbol{r}) \ge g_{\boldsymbol{r}}/2$ and the normalized error is at most $\sqrt{2}T_d$. Otherwise, let $r_i^2 = m > 1/2$ and write $\boldsymbol{r} = r_i \boldsymbol{f}_i + \boldsymbol{v}$, where $v_i = 0$ and $R = \|\boldsymbol{v}\|^2 = 1 - m$. Put $g_{\boldsymbol{v}} = \boldsymbol{v}^\top \Gamma^\star \boldsymbol{v}$. Since $1 - v_j^2 \ge m$ for $j \ne i$,

$$D_\Gamma(\boldsymbol{r}) \ge \gamma_i^\star m R + m g_{\boldsymbol{v}}. \tag{E.57}$$

The diagonal of E_d is zero, so

$$|\boldsymbol{r}^\top E_d \boldsymbol{r}| \le 2W_d \sqrt{\gamma_i^\star m R} + T_d \sqrt{R g_{\boldsymbol{v}}}. \tag{E.58}$$

Dividing by $\sqrt{D_\Gamma(\boldsymbol{r})}$ bounds this by $2W_d + T_d$, because $R < m$. At coordinate directions the numerator is exactly zero. Thus the supremum is at most $2T_d + 2W_d$. The preceding bounds prove (E.45), for example with $C_1 = 12C + 2$. □

E.2.4 Adding the signal and transferring the score gap

The sample covariance decomposes exactly as

$$\begin{aligned} \widehat{\Sigma} - \Sigma_d &= (\Gamma^\star)^{1/2}(Z^\top Z/n - \mathbf{I})(\Gamma^\star)^{1/2} + \sqrt{\beta_d}(\boldsymbol{e}\boldsymbol{t}^\top + \boldsymbol{t}\boldsymbol{e}^\top) + \beta_d(q_\lambda - 1)\boldsymbol{e}\boldsymbol{e}^\top, \\ q_\lambda &= \|\boldsymbol{\lambda}\|^2/n, \qquad \boldsymbol{t} = (\Gamma^\star)^{1/2} Z^\top \boldsymbol{\lambda}/n. \end{aligned} \tag{E.59}$$

Here $q_\lambda \to 1$, $\beta_d \to \beta$, and $\|\boldsymbol{t}\|^2 \to \alpha^{-1}$ in probability. To see the last assertion, condition on $\Gamma^\star, \boldsymbol{\lambda}$: $\boldsymbol{t}$ is Gaussian with covariance $q_\lambda \Gamma^\star/n$. Its squared norm has conditional mean $q_\lambda s_d \to \alpha^{-1}$ and conditional variance $2q_\lambda^2 \operatorname{tr}[(\Gamma^\star)^2]/n^2 \to 0$.

For unit $\boldsymbol{r}$, the Frobenius-norm identity is $\|\operatorname{off}(\boldsymbol{r}\boldsymbol{r}^\top)\|_{\mathrm{F}} = \sqrt{1 - I(\boldsymbol{r})}$. Frobenius Cauchy–Schwarz and $D_\Gamma \ge a(1 - I)$ give

$$\begin{aligned} \frac{|\boldsymbol{r}^\top \operatorname{off}(\boldsymbol{e}\boldsymbol{t}^\top + \boldsymbol{t}\boldsymbol{e}^\top)\boldsymbol{r}|}{\sqrt{D_\Gamma(\boldsymbol{r})}} &\le \frac{2\|\boldsymbol{t}\|}{\sqrt{a}}, \\ \frac{|\boldsymbol{r}^\top \operatorname{off}(\boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{r}|}{\sqrt{D_\Gamma(\boldsymbol{r})}} &\le \frac{1}{\sqrt{a}}. \end{aligned} \tag{E.60}$$

Combining this with Lemma E.8 proves

$$\begin{aligned} \sup_{\|\boldsymbol{r}\|=1,\ D_\Gamma>0} \frac{|\boldsymbol{r}^\top \operatorname{off}(\widehat{\Sigma} - \Sigma_d)\boldsymbol{r}|}{\sqrt{D_\Gamma(\boldsymbol{r})}} &\le B_\alpha + o_{\mathbb{P}}(1), \\ B_\alpha &:= C_1\left(\alpha^{-1/2} + a^{-1/2}\alpha^{-1}\right) + 2\sqrt{\frac{\beta}{a\alpha}}. \end{aligned} \tag{E.61}$$

In particular, $B_\alpha \to 0$ as $\alpha \to \infty$, with no upper variance bound. The mask identities now imply, simultaneously over all non-coordinate directions,

$$\frac{|\widehat{\mathcal{C}}(\boldsymbol{r}) - \mathcal{C}(\boldsymbol{r})|}{\sqrt{H(\boldsymbol{r})}} \le \sqrt{\rho(1-\rho)}\,[B_\alpha + o_{\mathbb{P}}(1)]. \tag{E.62}$$

The relative Wishart bound (E.25) remains valid: conditionally on $\boldsymbol{u}^\star, \Gamma^\star$, the samples are Gaussian with covariance Σ_d. Hence $(1 - \varepsilon_d)H \le \widehat{H} \le (1 + \varepsilon_d)H$, with $\varepsilon_d \to 2/\sqrt{\alpha} + 1/\alpha$. Define the actual uniform cross-coefficient error

$$b_d := \sup_{\|\boldsymbol{r}\|=1,\ H(\boldsymbol{r})>0} \frac{|\widehat{\mathcal{C}}(\boldsymbol{r}) - \mathcal{C}(\boldsymbol{r})|}{\sqrt{H(\boldsymbol{r})}} \le \sqrt{\rho(1-\rho)} B_\alpha + o_{\mathbb{P}}(1). \tag{E.63}$$

The score comparison is therefore

$$\begin{aligned} \sqrt{\widehat{S}(\boldsymbol{r})} &\le \frac{\sqrt{S(\boldsymbol{r})} + b_d}{\sqrt{1 - \varepsilon_d}}, \\ \sqrt{\widehat{S}(\boldsymbol{e})} &\ge \frac{\sqrt{S(\boldsymbol{e})} - b_d}{\sqrt{1 + \varepsilon_d}}, \end{aligned} \tag{E.64}$$

where the lower bound is used when its numerator is positive.

Proof of Theorem E.7. The population localization bound (E.21) uses only $\Gamma^\star \succeq a\mathbf{I}$, so it remains valid. The teacher score limit (E.23) also remains valid: the law of large numbers gives $\boldsymbol{e}^\top\Gamma^\star\boldsymbol{e} \to 1$, and $\sum_i \gamma_i^\star e_i^4 \le (\max_i e_i^2)\boldsymbol{e}^\top\Gamma^\star\boldsymbol{e} \to 0$. Thus, with $g = \rho(1-\rho)\beta^2/(1+\beta(1-\rho))$ and η from (E.40), with probability tending to one,

$$\sup_{|\boldsymbol{e}^\top\boldsymbol{r}|^2\le\eta} S(\boldsymbol{r}) \le g/16, \qquad S(\boldsymbol{e}) \ge 3g/4. \tag{E.65}$$

For a finite sufficient bound, define

$$\begin{aligned} C_{a,\beta} &= C_1(1+a^{-1/2}) + 2\sqrt{\beta/a}, \\ A(a,\beta,\rho) &= \max\left\{100, \frac{1024C_{a,\beta}^2[1+\beta(1-\rho)]}{\beta^2}\right\}. \end{aligned} \tag{E.66}$$

For $\alpha > A$, one has $\varepsilon_\alpha < 1/4$ and $B_\alpha \le C_{a,\beta}/\sqrt{\alpha} < \beta/[32\sqrt{1+\beta(1-\rho)}]$. Consequently, with probability tending to one, $\varepsilon_d < 1/4$ and $b_d < \sqrt{g}/16$. Equations (E.64) and (E.65) yield the strict empirical gap

$$\sup_{|\boldsymbol{e}^\top\boldsymbol{r}|^2\le\eta} \sqrt{\widehat{S}(\boldsymbol{r})} \le \frac{(1/4+1/16)\sqrt{g}}{\sqrt{3/4}} < \frac{(\sqrt{3/4}-1/16)\sqrt{g}}{\sqrt{5/4}} \le \sqrt{\widehat{S}(\boldsymbol{e})}. \tag{E.67}$$

This excludes every low-overlap direction, including directions chosen adaptively from the data.

For attainment, $\max_i e_i^2 < \eta$ with probability tending to one. The compact set $\{\|\boldsymbol{r}\| = 1 : |\boldsymbol{e}^\top\boldsymbol{r}|^2 \ge \eta\}$ then contains no coordinate direction. On this set, $D_\Gamma > 0$, and the relative covariance event makes $\widehat{H} > 0$. The continuous score attains its maximum there. The strict empirical gap makes this a global maximum with positive score. Its radial optimizer $Q_* = [\widehat{\mathcal{C}}]_+/\widehat{H}$ is finite and positive by (E.17). Coordinate directions and the zero weight have zero centered improvement and cannot minimize globally. This proves existence and the overlap assertion for every global minimizer. □

The sufficient sample ratio in (E.66) has, up to universal multiplicative constants, the dependence $A(a,\beta,\rho) \asymp 1 + (1+\beta)[1+\beta(1-\rho)]/(a\beta^2)$, since $a \le 1$. Holding the other parameters fixed, this bound scales as β^{-2} for weak signal and as a^{-1} when the lower variance bound vanishes; it remains bounded as $\beta \to \infty$. These describe the chosen sufficient bound, not a sharp recovery threshold. Its conservatism reflects uniform control over all candidate directions, lower bounds on the masking penalty, and the numerical margins used to preserve the score gap. The variance law's second moment enters the vanishing remainder terms rather than this leading bound; the dimension needed for the high-probability guarantee can therefore depend on the law.

One fixed unbounded example Take

$$\gamma = (1+\xi_{\exp})/2, \qquad \xi_{\exp} \sim \mathrm{Exp}(1). \tag{E.68}$$

Then $a = 1/2$, $\mathbb{E}_\gamma\gamma = 1$, $\kappa_2 = 5/4$, and $\kappa_3 = 2$. This one variance law has unbounded support and satisfies all assumptions of Theorem E.7. For every fixed $\beta > 0$ and $\rho \in (0,1)$, dynamic linear masking recovers at every sufficiently large fixed α, whereas PCA and full linear reconstruction fail at every fixed $\alpha > 0$. The required sufficient α may depend on β and ρ; it is not uniform as $\beta \downarrow 0$.

F Bayes-optimal pretraining and AMP

The benchmark in this appendix conditions on the realized variance labels and uses priors matched to the data-generating model. We derive its scalar state evolution and specialize the bipartite AMP of Lesieur et al. (2017) in App. F.5. The calculation of variance consistency below is distinct from a proof that an estimated-variance algorithm attains the same benchmark.

F.1 Mapping to bipartite low-rank estimation

We map the heteroskedastic spiked model to the bipartite low-rank output-channel formulation of (Lesieur et al., 2015):

$$x_{\mu i} = \sqrt{\frac{\beta}{d}}\,\lambda_\mu u_i^\star + \sqrt{\gamma_i^\star}\,z_{\mu i}, \qquad \mu = 1,\dots,n, \quad i = 1,\dots,d, \tag{F.1}$$

with

$$\frac{n}{d} \longrightarrow \alpha, \qquad \lambda_\mu \overset{\text{iid}}{\sim} P_\lambda, \qquad u_i^\star \overset{\text{iid}}{\sim} P_u, \qquad \gamma_i^\star \overset{\text{iid}}{\sim} P_\gamma, \qquad z_{\mu i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{F.2}$$

We assume throughout that

$$\mathbb{E}[u^\star] = \mathbb{E}[\lambda] = 0, \qquad \mathbb{E}[(u^\star)^2] = \mathbb{E}[\lambda^2] = 1, \tag{F.3}$$

and that the moments needed below are finite. In particular, for the linearized weak-recovery calculation we assume

$$\mathbb{E}_\gamma[\gamma^{-2}] < \infty. \tag{F.4}$$

We retain n, d for the sample and coordinate counts, respectively, and the sample-first data matrix $X = (x_{\mu i})$. Relative to the cited bipartite formulation, its coordinate and sample counts correspond to d and n, and its observation matrix is $X^\top$. We identify

$$P_U = P_u, \qquad P_V = P_\lambda, \qquad v_\mu^\star = \lambda_\mu. \tag{F.5}$$

The rank-one variable entering the output channel is

$$w_{\mu i}^\star := \frac{u_i^\star v_\mu^\star}{\sqrt{d}}. \tag{F.6}$$

The SNR parameter β is absorbed into the output channel, which, conditional on the variance label γ, is

$$P_{\text{out},\gamma}(x \mid w) = \frac{1}{\sqrt{2\pi\gamma}} \exp\left\{-\frac{(x - \sqrt{\beta}\, w)^2}{2\gamma}\right\}. \tag{F.7}$$

Thus

$$x_{\mu i} = \sqrt{\beta}\, w_{\mu i}^\star + \sqrt{\gamma_i^\star}\, z_{\mu i}, \tag{F.8}$$

which is exactly Eq. (F.1) in the sample-first convention.

F.2 Asymptotic identifiability of the coordinate variances

The bipartite formulas of (Lesieur et al., 2017) are written for a known output channel. The state-evolution reference conditions on the realized coordinate variances. In the numerical AMP implementation, we instead estimate these variances from the unlabeled observations.

For a fixed coordinate i, define the empirical second moment

$$\widehat{\gamma}_i := \frac{1}{n} \sum_{\mu=1}^{n} x_{\mu i}^2. \tag{F.9}$$

Using the data model (F.1),

$$\widehat{\gamma}_i = \gamma_i^\star \frac{1}{n} \sum_{\mu=1}^{n} z_{\mu i}^2 + \frac{\beta (u_i^\star)^2}{d} \frac{1}{n} \sum_{\mu=1}^{n} \lambda_\mu^2 + 2\sqrt{\frac{\beta \gamma_i^\star}{d}}\, u_i^\star \frac{1}{n} \sum_{\mu=1}^{n} \lambda_\mu z_{\mu i}. \tag{F.10}$$

We now estimate the three terms separately. Since $z_{\mu i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$,

$$\frac{1}{n} \sum_{\mu=1}^{n} z_{\mu i}^2 = 1 + O_{\mathbb{P}}(n^{-1/2}) = 1 + O_{\mathbb{P}}(d^{-1/2}), \tag{F.11}$$

where the last equality uses $n/d \to \alpha \in (0, \infty)$ in the high-dimensional limit. Therefore the first term is

$$\gamma_i^\star \frac{1}{n} \sum_{\mu=1}^{n} z_{\mu i}^2 = \gamma_i^\star + O_{\mathbb{P}}(d^{-1/2}). \tag{F.12}$$

Because $\mathbb{E}[\lambda^2] = 1$, we have that

$$\frac{1}{n} \sum_{\mu=1}^{n} \lambda_\mu^2 = 1 + o(1). \tag{F.13}$$

For each fixed coordinate i, $u_i^\star = O_{\mathbb{P}}(1)$, and hence the second term is subleading

$$\frac{\beta(u_i^\star)^2}{d}\frac{1}{n}\sum_{\mu=1}^{n}\lambda_\mu^2 = O_{\mathbb{P}}(d^{-1}). \tag{F.14}$$

Finally, the variables $\lambda_\mu z_{\mu i}$ are centered and have unit variance, so

$$\frac{1}{n}\sum_{\mu=1}^{n}\lambda_\mu z_{\mu i} = O_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(d^{-1/2}). \tag{F.15}$$

Multiplying by the prefactor $d^{-1/2}$ shows that the cross term is

$$2\sqrt{\frac{\beta\gamma_i^\star}{d}}\,u_i^\star\frac{1}{n}\sum_{\mu=1}^{n}\lambda_\mu z_{\mu i} = O_{\mathbb{P}}(d^{-1}). \tag{F.16}$$

Combining Eqs. (F.12), (F.14), and (F.16), we obtain

$$\widehat{\gamma}_i = \gamma_i^\star + O_{\mathbb{P}}(d^{-1/2}). \tag{F.17}$$

In particular,

$$\widehat{\gamma}_i \xrightarrow{\mathbb{P}} \gamma_i^\star. \tag{F.18}$$

Thus each fixed coordinate variance is consistently estimable. In the Gaussian-prior specialization with $\gamma_i^\star \geq a > 0$, one can also obtain uniform relative consistency. Conditionally on the teacher and variances, $\widehat{\gamma}_i/[\gamma_i^\star + \beta(u_i^\star)^2/d]$ has law χ_n^2/n. A union bound and $\max_i(u_i^\star)^2 = O_{\mathbb{P}}(\log d)$ give

$$\max_i\left|\frac{\widehat{\gamma}_i}{\gamma_i^\star} - 1\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{d}\right) \longrightarrow 0. \tag{F.19}$$

Neither pointwise nor uniform parameter consistency alone establishes that substituting all $\widehat{\gamma}_i$ into the posterior or AMP preserves its limiting overlap and threshold. That substitution also introduces dependence through the data used to estimate the variances. We therefore state the following equations for the known-variance benchmark; equivalence of its plug-in version requires an additional stability argument.

F.3 Bayes-optimal output channel and Fisher information

Conditioned on $\gamma_i^\star$, define the output channel

$$g_i(x, w) := \log P_{\mathrm{out},\gamma_i^\star}(x \mid w). \tag{F.20}$$

The posterior can be written in the normalized form used in the bipartite low-rank results:

$$P(U, V \mid X, \boldsymbol{\gamma}^\star) \propto \prod_{i=1}^{d} P_U(\mathrm{d}u_i)\prod_{\mu=1}^{n} P_V(\mathrm{d}v_\mu) \times \exp\left\{\sum_{i=1}^{d}\sum_{\mu=1}^{n}\left[g_i\left(x_{\mu i}, \frac{u_i v_\mu}{\sqrt{d}}\right) - g_i(x_{\mu i}, 0)\right]\right\}. \tag{F.21}$$

Following the Fisher-score notation

$$S_i(x) := \left.\frac{\partial g_i(x, w)}{\partial w}\right|_{w=0} \tag{F.22}$$

$$R_i(x) := \left.\left[\left(\frac{\partial g_i(x, w)}{\partial w}\right)^2 + \frac{\partial^2 g_i(x, w)}{\partial w^2}\right]\right|_{w=0}. \tag{F.23}$$

In particular, for the Gaussian channel (F.7),

$$S_i(x) = \frac{\sqrt{\beta}}{\gamma_i^\star}x, \tag{F.24}$$

whereas

$$R_i(x) = \beta \left(\frac{x^2}{(\gamma_i^\star)^2} - \frac{1}{\gamma_i^\star} \right). \tag{F.25}$$

Under the null output law $x \sim \mathcal{N}(0, \gamma_i^\star)$,

$$\mathbb{E}[S_i(x)] = 0, \qquad \mathbb{E}[R_i(x)] = 0. \tag{F.26}$$

The local Fisher-information parameter is defined by

$$\frac{1}{\Delta_i} := \mathbb{E}_{x \sim P_{\mathrm{out},\gamma_i^\star}(\cdot|0)} \left[S_i(x)^2 \right]. \tag{F.27}$$

and using Eq. (F.24),

$$\frac{1}{\Delta_i} = \frac{\beta}{\gamma_i^\star}, \qquad \Delta_i = \Delta(\gamma_i^\star) = \frac{\gamma_i^\star}{\beta}. \tag{F.28}$$

Thus heteroskedasticity enters the Bayes-optimal low-rank equations only in the site-dependent Fisher information

$$\Delta(\gamma)^{-1} = \beta \gamma^{-1}. \tag{F.29}$$

F.4 Denoisers

For a prior P_U, define the scalar partition function

$$Z_U(A, B) := \int P_U(\mathrm{d}u) \exp\left\{ Bu - \frac{A}{2} u^2 \right\}, \tag{F.30}$$

and the posterior-mean denoiser

$$f_{\mathrm{in}}^U(A, B) := \frac{\partial}{\partial B} \log Z_U(A, B) = \frac{\int P_U(\mathrm{d}u)\, u \exp\{Bu - Au^2/2\}}{Z_U(A, B)}. \tag{F.31}$$

Define Z_V and f_{in}^V analogously with $P_V = P_\lambda$.

For a scalar Bayes channel of A introduce the overlap response functions

$$\mathcal{F}_U(A) := \mathbb{E}_{u^\star, W} \left[u^\star f_{\mathrm{in}}^U \left(A, Au^\star + \sqrt{A}\, W \right) \right], \tag{F.32}$$

$$\mathcal{F}_V(A) := \mathbb{E}_{v^\star, W} \left[v^\star f_{\mathrm{in}}^V \left(A, Av^\star + \sqrt{A}\, W \right) \right], \tag{F.33}$$

where

$$u^\star \sim P_U, \qquad v^\star \sim P_V, \qquad W \sim \mathcal{N}(0, 1), \tag{F.34}$$

independently. By the Nishimori identity,

$$\mathcal{F}_U(A) = \mathbb{E}_{u^\star, W} \left[f_{\mathrm{in}}^U \left(A, Au^\star + \sqrt{A}\, W \right)^2 \right], \tag{F.35}$$

$$\mathcal{F}_V(A) = \mathbb{E}_{v^\star, W} \left[f_{\mathrm{in}}^V \left(A, Av^\star + \sqrt{A}\, W \right)^2 \right]. \tag{F.36}$$

F.5 Explicit heteroskedastic bipartite AMP

We specialize the bipartite Low-RAMP equations (90)–(97) of Lesieur et al. (2017), Section II.4, using the sample-first observations $x_{\mu i}$, d coordinate variables, and n sample variables with $v_\mu^\star = \lambda_\mu$. The variance labels $\gamma_i^\star$ are held fixed and known in the matched benchmark. Define the coordinate-by-sample coupling and likelihood curvature

$$J_{i\mu} := \frac{\sqrt{\beta}\, x_{\mu i}}{\sqrt{d}\, \gamma_i^\star}, \qquad T_i := \frac{\beta}{d \gamma_i^\star}. \tag{F.37}$$

The Gaussian likelihood contributes exactly $J_{i\mu} u_i v_\mu - T_i u_i^2 v_\mu^2/2$ to the log posterior. In the score notation of (F.24) and (F.25), $S_{i\mu}/\sqrt{d} = J_{i\mu}$ and $R_{i\mu}/d = J_{i\mu}^2 - T_i$. These identities fix both the normalization and the signs of the precision updates below.

Let $\widehat{u}_i, \widehat{v}_\mu$ denote AMP means and $c_{u,i}, c_{v,\mu}$ their scalar posterior variances. A sweep first updates the coordinate side at time t, then the sample side to time $t+1$. Given $(\widehat{v}^t, c_v^t, \widehat{u}^{t-1})$, set

$$\begin{aligned} B_{U,i}^t &= \sum_{\mu=1}^n J_{i\mu}\widehat{v}_\mu^t - \widehat{u}_i^{t-1}\sum_{\mu=1}^n J_{i\mu}^2 c_{v,\mu}^t, \\ A_{U,i}^t &= \sum_{\mu=1}^n \left[T_i(\widehat{v}_\mu^t)^2 - (J_{i\mu}^2 - T_i)c_{v,\mu}^t\right], \\ \widehat{u}_i^t &= f_{\text{in}}^U(A_{U,i}^t, B_{U,i}^t), \qquad c_{u,i}^t = \partial_B f_{\text{in}}^U(A_{U,i}^t, B_{U,i}^t). \end{aligned} \tag{F.38}$$

Using these newly computed coordinate estimates, set

$$\begin{aligned} B_{V,\mu}^t &= \sum_{i=1}^d J_{i\mu}\widehat{u}_i^t - \widehat{v}_\mu^t\sum_{i=1}^d J_{i\mu}^2 c_{u,i}^t, \\ A_{V,\mu}^t &= \sum_{i=1}^d \left[T_i(\widehat{u}_i^t)^2 - (J_{i\mu}^2 - T_i)c_{u,i}^t\right], \\ \widehat{v}_\mu^{t+1} &= f_{\text{in}}^V(A_{V,\mu}^t, B_{V,\mu}^t), \qquad c_{v,\mu}^{t+1} = \partial_B f_{\text{in}}^V(A_{V,\mu}^t, B_{V,\mu}^t). \end{aligned} \tag{F.39}$$

In particular, the two Onsager terms multiply $\widehat{u}^{t-1}$ and $\widehat{v}^t$, respectively. This is the alternating schedule of the cited equations. These are AMP approximations to the posterior; the exactness of the Gaussian likelihood expansion does not make its finite-dimensional marginals factorized.

Fisher-averaged implementation. Under the regular state-evolution assumptions, the squared couplings in the summed corrections can be replaced by their null-channel means $\mathbb{E}_0[J_{i\mu}^2 \mid \gamma_i^\star] = T_i$; the centered curvature terms $J_{i\mu}^2 - T_i$ self-average. This is a site-dependent version of the Bayes simplification in Section II.3 of Lesieur et al. (2017). With $\alpha_d = n/d$, define

$$\begin{aligned} Q_v^t &= \frac{1}{n}\sum_\mu (\widehat{v}_\mu^t)^2, \qquad V_v^t = \frac{1}{n}\sum_\mu c_{v,\mu}^t, \\ Q_{u,-1}^t &= \frac{1}{d}\sum_i \frac{(\widehat{u}_i^t)^2}{\gamma_i^\star}, \quad V_{u,-1}^t = \frac{1}{d}\sum_i \frac{c_{u,i}^t}{\gamma_i^\star}. \end{aligned} \tag{F.40}$$

The resulting implementation uses

$$\begin{aligned} A_{U,i}^t &= \frac{\alpha_d \beta Q_v^t}{\gamma_i^\star}, \quad B_{U,i}^t = \sum_\mu J_{i\mu}\widehat{v}_\mu^t - \frac{\alpha_d\beta V_v^t}{\gamma_i^\star}\widehat{u}_i^{t-1}, \\ A_{V,\mu}^t &= \beta Q_{u,-1}^t, \quad B_{V,\mu}^t = \sum_i J_{i\mu}\widehat{u}_i^t - \beta V_{u,-1}^t \widehat{v}_\mu^t, \end{aligned} \tag{F.41}$$

followed by the same denoisers and the same update order. The factors $1/\gamma_i^\star$ remain inside the coordinate-side coefficients and sample-side averages; replacing them by a single mean variance changes the algorithm. The two finite-dimensional versions (F.38)–(F.39) and (F.41) need not give identical iterates.

For $P_u = P_\lambda = \mathcal{N}(0,1)$, both denoisers are explicit:

$$\begin{aligned} \widehat{u}_i^t &= \frac{B_{U,i}^t}{1 + A_{U,i}^t}, \quad c_{u,i}^t = \frac{1}{1 + A_{U,i}^t}, \\ \widehat{v}_\mu^{t+1} &= \frac{B_{V,\mu}^t}{1 + A_{V,\mu}^t}, \quad c_{v,\mu}^{t+1} = \frac{1}{1 + A_{V,\mu}^t}. \end{aligned} \tag{F.42}$$

The Fisher-averaged precisions are nonnegative, so these denominators are at least one. The unsimplified iteration requires $1 + A > 0$ for each Gaussian denoiser. For a Rademacher sample prior, only the sample-side denoiser changes to $\tanh B$ with variance $1 - \tanh^2 B$; it must not be used for Gaussian sample amplitudes.

Connection to the existing state evolution. Let M_v^t, Q_v^t be the limiting overlap and squared norm on the sample side, and $M_u^t(\gamma), Q_u^t(\gamma)$ their conditional coordinate-side counterparts. The formal Onsager-corrected scalar fields corresponding to the alternating AMP have

$$\begin{aligned} \mathcal{A}_U^t(\gamma) &= \frac{\alpha\beta Q_v^t}{\gamma}, \qquad \mathcal{B}_U^t(\gamma) = \frac{\alpha\beta M_v^t}{\gamma} u^\star + \sqrt{\mathcal{A}_U^t(\gamma)}\, Z_U, \\ \mathcal{A}_V^t &= \beta \mathbb{E}_\gamma \frac{Q_u^t(\gamma)}{\gamma}, \qquad \mathcal{B}_V^t = \beta \mathbb{E}_\gamma \frac{M_u^t(\gamma)}{\gamma}\, v^\star + \sqrt{\mathcal{A}_V^t}\, Z_V, \end{aligned} \tag{F.43}$$

where the Gaussian noises are independent of the respective scalar teachers. These are effective cavity laws, not independence assertions between finite-dimensional iterates and their training data. On a matched Nishimori trajectory, $Q_v^t = M_v^t$ and $Q_u^t(\gamma) = M_u^t(\gamma)$. Applying the scalar denoisers to (F.43) then gives exactly (F.49)–(F.51). This verifies the normalization of the explicit AMP against the existing Bayes equations. It does not supply a general convergence theorem from an arbitrary initialization or a uniform state-evolution theorem for all unbounded variance laws.

The formulas condition on the realized variances. Replacing them by (F.9) defines a plug-in AMP, whose performance requires the stability analysis discussed in App. F.2. When complete training samples are available, independent masking carries no additional information about the planted factors. The full-data Bayes/AMP benchmark therefore does not depend on the masking ratio ρ or the number of masks K.

F.6 State evolution

Let

$$M_v^t := \frac{1}{n} \sum_{\mu=1}^{n} \widehat{v}_\mu^t v_\mu^\star \tag{F.44}$$

denote the overlap on the latent variables side. On the coordinate side (spike), the effective channel depends on the quenched variances. For a site with label γ, the bipartite state-evolution field has to be rescaled with

$$A_U^t(\gamma) = \frac{\alpha M_v^t}{\Delta(\gamma)}. \tag{F.45}$$

Therefore the conditional coordinate overlap is

$$M_u^t(\gamma) = \mathcal{F}_U\left(\frac{\alpha M_v^t}{\Delta(\gamma)}\right). \tag{F.46}$$

and asymptotically

$$\overline{M}_u^t := \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \frac{M_{u,i}^t}{\Delta_i} = \mathbb{E}_\gamma\left[\frac{M_u^t(\gamma)}{\Delta(\gamma)}\right]. \tag{F.47}$$

The V-side update is consequently

$$M_v^{t+1} = \mathcal{F}_V\left(\overline{M}_u^t\right). \tag{F.48}$$

For the heteroskedastic Gaussian channel $\Delta(\gamma) = \gamma/\beta$, these equations become

$$M_u^t(\gamma) = \mathcal{F}_U\left(\frac{\alpha\beta}{\gamma} M_v^t\right), \tag{F.49}$$

$$\overline{M}_u^t = \beta\, \mathbb{E}_\gamma\left[\frac{M_u^t(\gamma)}{\gamma}\right], \tag{F.50}$$

$$M_v^{t+1} = \mathcal{F}_V\left(\overline{M}_u^t\right). \tag{F.51}$$

F.7 Fixed-point equations in the present notation

Returning to

$$v_\mu^\star = \lambda_\mu, \qquad M_v = q_\lambda, \qquad M_u(\gamma) = q_u(\gamma), \tag{F.52}$$

the coordinate-side scalar precision is

$$\widehat{q}_u(\gamma) = \frac{\alpha\beta}{\gamma} q_\lambda. \tag{F.53}$$

The sample-amplitude scalar precision is the Fisher-weighted coordinate overlap

$$\widehat{q}_\lambda = \overline{M}_u = \beta \mathbb{E}_\gamma \left[\frac{q_u(\gamma)}{\gamma} \right]. \tag{F.54}$$

Thus the fixed-point equations are

$$q_u(\gamma) = \mathbb{E}_{u^\star, W} \left[u^\star f_{\text{in}}^u \left(\widehat{q}_u(\gamma), \widehat{q}_u(\gamma) u^\star + \sqrt{\widehat{q}_u(\gamma)}\, W \right) \right], \tag{F.55}$$

$$q_\lambda = \mathbb{E}_{\lambda, W} \left[\lambda f_{\text{in}}^\lambda \left(\widehat{q}_\lambda, \widehat{q}_\lambda \lambda + \sqrt{\widehat{q}_\lambda}\, W \right) \right], \tag{F.56}$$

with conjugate relations

$$\widehat{q}_u(\gamma) = \frac{\alpha\beta}{\gamma} q_\lambda, \qquad \widehat{q}_\lambda = \beta \mathbb{E}_\gamma \left[\frac{q_u(\gamma)}{\gamma} \right]. \tag{F.57}$$

By Nishimori,

$$q_u(\gamma) = \mathbb{E}_{u^\star, W} \left[f_{\text{in}}^u \left(\widehat{q}_u(\gamma), \widehat{q}_u(\gamma) u^\star + \sqrt{\widehat{q}_u(\gamma)} W \right)^2 \right], \tag{F.58}$$

$$q_\lambda = \mathbb{E}_{\lambda, W} \left[f_{\text{in}}^\lambda \left(\widehat{q}_\lambda, \widehat{q}_\lambda \lambda + \sqrt{\widehat{q}_\lambda} W \right)^2 \right]. \tag{F.59}$$

F.8 Continuous weak-recovery threshold

The uninformative fixed point is

$$q_\lambda = 0, \qquad q_u(\gamma) = 0. \tag{F.60}$$

On the matched Nishimori trajectory, and under the scalar-response regularity used in the linearization, at small order

$$\mathcal{F}_U(A) = A + o(A), \qquad \mathcal{F}_V(A) = A + o(A). \tag{F.61}$$

Linearizing Eq. (F.49) around $M_v = 0$ gives

$$M_u^t(\gamma) = \frac{\alpha\beta}{\gamma} M_v^t + o(M_v^t). \tag{F.62}$$

Consequently,

$$\overline{M}_u^t = \beta \mathbb{E}_\gamma \left[\frac{M_u^t(\gamma)}{\gamma} \right] = \alpha\beta^2 \mathbb{E}_\gamma[\gamma^{-2}] M_v^t + o(M_v^t). \tag{F.63}$$

The V-side update then gives

$$M_v^{t+1} = \alpha\beta^2 \mathbb{E}_\gamma[\gamma^{-2}] M_v^t + o(M_v^t). \tag{F.64}$$

For Gaussian priors, dominated convergence justifies the integrated linear coefficient with only (F.4). A quadratic remainder generally requires an inverse third moment; the linear threshold does not require that stronger expansion. Therefore the uninformative fixed point is locally stable when

$$\alpha\beta^2 \mathbb{E}_\gamma[\gamma^{-2}] < 1 \tag{F.65}$$

and loses stability when

$$\alpha\beta^2 \mathbb{E}_\gamma[\gamma^{-2}] > 1. \tag{F.66}$$

The Bayes-optimal continuous weak-recovery threshold is consequently

$$\alpha_{\text{c}}^{\text{BO}} = \frac{1}{\beta^2 \mathbb{E}_\gamma[\gamma^{-2}]}. \tag{F.67}$$

This is the local continuous instability of the uninformative Bayes fixed point; for generic priors it does not by itself exclude an informative branch appearing discontinuously at a smaller sample complexity. For the homogeneous Gaussian noise-specialization $P_\gamma = \delta_1$, this reduces to

$$\alpha_{\text{c}}^{\text{BO}} = \frac{1}{\beta^2}, \tag{F.68}$$

F.9 Gaussian-prior specialization

If

$$P_u = P_\lambda = \mathcal{N}(0,1), \tag{F.69}$$

then

$$Z(A,B) = \frac{1}{\sqrt{1+A}} \exp\left\{\frac{B^2}{2(1+A)}\right\}, \tag{F.70}$$

so

$$f_{\text{in}}(A,B) = \frac{B}{1+A} \tag{F.71}$$

$$\mathcal{F}(A) = \frac{A}{1+A}. \tag{F.72}$$

The coordinate overlap is therefore

$$q_u(\gamma) = \frac{\alpha\beta q_\lambda}{\gamma + \alpha\beta q_\lambda}. \tag{F.73}$$

The sample-side precision is

$$\widehat{q}_\lambda = \beta\mathbb{E}_\gamma\left[\frac{q_u(\gamma)}{\gamma}\right] = \alpha\beta^2 q_\lambda \mathbb{E}_\gamma\left[\frac{1}{\gamma(\gamma+\alpha\beta q_\lambda)}\right], \tag{F.74}$$

$$q_\lambda = \frac{\widehat{q}_\lambda}{1+\widehat{q}_\lambda}. \tag{F.75}$$

Linearizing Eqs. (F.73)–(F.75) again yields Eq. (F.67).

Complete overlap curve. For $q_\lambda > 0$, the preceding equations are equivalent to the single scalar equation

$$\frac{1}{1-q_\lambda} = \alpha\beta^2\mathbb{E}_\gamma\frac{1}{\gamma(\gamma+\alpha\beta q_\lambda)}, \qquad 0 < q_\lambda < 1. \tag{F.76}$$

The left side strictly increases from one to infinity; the right side strictly decreases from $\alpha\beta^2\mathbb{E}_\gamma\gamma^{-2}$. Hence a positive fixed point exists if and only if $\alpha\beta^2\mathbb{E}_\gamma\gamma^{-2} > 1$, and it is unique within these Gaussian-prior equations. Below and at this threshold only the uninformative fixed point remains. This excludes a competing positive fixed point within this scalar reduction, not an algorithmic failure caused by its initialization or finite-dimensional dynamics.

The Bayes-optimal cosine similarity plotted in the figures is $\sqrt{\mathcal{O}_{\text{BO}}}$. For homogeneous noise, elimination of q_λ recovers $\mathcal{O}_{\text{BO}} = (\alpha\beta^2 - 1)/(\alpha\beta^2 + \beta)$ above threshold, in agreement with the PCA benchmark.

For Gaussian priors and known covariance with compact positive variance support, the weak-estimation threshold also has an independent rigorous justification through Proposition 4.1 and Corollary 4.3 of Zhang & Mondelli (2024), after transposing the model and identifying their Section 4 SNR with β and aspect ratio with α. That result concerns sign-invariant estimation and has covariance-support assumptions; it does not by itself justify plug-in AMP or the full unwhitened cosine curve for arbitrary unbounded laws.

G Additional observables and finite-shot downstream analysis

G.1 Training and test errors

Besides the cosine similarity defined in the main text, we can also track the training and test error of the pretraining for the autoencoder. Using the objective $\mathcal{R}_{K,d}$ in (3), we define the centered training reconstruction error per sample by removing the ridge penalty and subtracting the risk assuming $\widehat{\boldsymbol{w}} = \mathbf{0}$,

$$\Delta\mathcal{L}_{\text{train},d} := \frac{1}{n}\left[\mathcal{R}_{K,d}(\widehat{\boldsymbol{w}}) - \mathcal{R}_{K,d}(\mathbf{0}) - \frac{\lambda_R}{2}\|\widehat{\boldsymbol{w}}\|_2^2\right]. \tag{G.1}$$

For dynamic masking, the view average is replaced by an expectation over fresh masks on the training samples. The centered population test error uses an independent sample and an independent fresh mask:

$$\Delta\mathcal{L}_{\mathrm{test},d} := \mathbb{E}_{\boldsymbol{x}_{\mathrm{new}},\boldsymbol{\chi}_{\mathrm{new}}}\left[\left\|\boldsymbol{\chi}_{\mathrm{new}}\odot\left[\boldsymbol{x}_{\mathrm{new}}-\frac{\widehat{\boldsymbol{w}}}{\sqrt{d}}\sigma(h^{\mathrm{new}}_{\mathrm{vis}})\right]\right\|_2^2-\|\boldsymbol{\chi}_{\mathrm{new}}\odot\boldsymbol{x}_{\mathrm{new}}\|_2^2\right]. \tag{G.2}$$

Both quantities exclude the ridge penalty. Lower values are better. In the simulations, the test expectation is estimated using independent test samples and masks.

At the replica saddle, these observables are

$$\Delta\mathcal{L}_{\mathrm{train}} = \mathbb{E}_{\lambda,\boldsymbol{z}}[\ell_{\mathrm{out}}(\boldsymbol{f}_\star; q_I, q_\Gamma)], \tag{G.3}$$

where $\boldsymbol{f}_\star$ is the training-sample proximal field, and

$$\Delta\mathcal{L}_{\mathrm{test}} = \mathbb{E}_{\lambda,z}[\ell_{\mathrm{dyn}}(h_0; q_I, q_\Gamma)], \qquad h_0 = \sqrt{\beta}\lambda M + \sqrt{q_\Gamma}\,z, \quad z\sim\mathcal{N}(0,1). \tag{G.4}$$

Why center the reconstruction error? The population test loss for the null estimator is

$$\mathcal{L}_{0,\mathrm{test},d} = \rho\sum_{i=1}^{d}\gamma_i + \frac{\rho\beta}{d}\|\boldsymbol{u}^\star\|_2^2 \sim \rho d, \tag{G.5}$$

Under the order-one scaling of the replica solution, the reconstruction-dependent terms are only $O(1)$, so this $O(d)$ baseline dominates the raw test loss. Subtracting it reduces to:

$$\Delta\mathcal{L}_{\mathrm{test},d} = \mathbb{E}\left[\sum_i \chi_i^{\mathrm{new}}\left(\widehat{x}_i^2 - 2x_i^{\mathrm{new}}\widehat{x}_i\right)\right], \qquad \widehat{\boldsymbol{x}} = \frac{\widehat{\boldsymbol{w}}}{\sqrt{d}}\sigma(h^{\mathrm{new}}_{\mathrm{vis}}). \tag{G.6}$$

The centered error therefore isolates the $O(1)$ contribution that changes with the learned representation. The same cancellation applies to the training error, using its own samples and masks.

G.2 Downstream task

After self-supervised pretraining, we freeze the learned weight vector $\widehat{\boldsymbol{w}}$ and evaluate it on an independent labeled dataset

$$\mathcal{D}_{\mathrm{ds}} = \left\{(\boldsymbol{x}_\mu^{\mathrm{ds}}, y_\mu)\right\}_{\mu=1}^{n_{\mathrm{ds}}}, \qquad y_\mu = \mathrm{sign}(\lambda_\mu), \qquad n_{\mathrm{ds}} = O(1). \tag{G.7}$$

We assume $P_\lambda(\{0\}) = 0$, so labels lie in $\{-1,+1\}$. No masking is applied at downstream time.

For a fresh downstream sample, the frozen preactivation satisfies

$$H := \frac{1}{\sqrt{d}}\widehat{\boldsymbol{w}}^\top\boldsymbol{x}^{\mathrm{ds}} \xrightarrow[d\to\infty]{} \sqrt{\beta}\,M\lambda + \sqrt{q_\Gamma}\,Z, \qquad Z\sim\mathcal{N}(0,1). \tag{G.8}$$

We use this preactivation directly as the downstream representation. The downstream channel therefore depends only on the pretraining overlaps M and q_Γ: M controls the informative component and q_Γ the projected noise variance.

We fit a one-dimensional logistic probe,

$$\widehat{a} = \arg\min_{a\in\mathbb{R}}\sum_{\mu=1}^{n_{\mathrm{ds}}}\log(1+\exp[-y_\mu a H_\mu])\,. \tag{G.9}$$

Its orientation is determined by

$$\mathrm{sign}(\widehat{a}) = \mathrm{sign}\left(\sum_{\mu=1}^{n_{\mathrm{ds}}} y_\mu H_\mu\right). \tag{G.10}$$

If the unregularized probe is separable, this orientation is understood along a minimizing sequence. Adding a positive quadratic penalty makes the minimizer finite and preserves the same orientation rule.

Assume $M \neq 0$ in the following orientation formulas. The population-optimal orientation of the frozen feature is

$$s_\star = \operatorname{sign}\left(\mathbb{E}_{\lambda,Z}\left[\operatorname{sign}(\lambda)\left(\sqrt{\beta}\,M\lambda + \sqrt{q_\Gamma}\,Z\right)\right]\right) = \operatorname{sign}(M), \tag{G.11}$$

where we used

$$\mathbb{E}_{\lambda,Z}\left[\operatorname{sign}(\lambda)H\right] = \sqrt{\beta}\,M\,\mathbb{E}_\lambda[|\lambda|]. \tag{G.12}$$

The corresponding population classification error is

$$\epsilon_{\rm ds}^{\infty} = \mathbb{P}\left[s_\star\ \operatorname{sign}(\lambda)\left(\sqrt{\beta}\,M\lambda + \sqrt{q_\Gamma}\,Z\right) < 0\right]. \tag{G.13}$$

For $n_{\rm ds} = O(1)$ labeled examples, the probability that the fitted probe chooses the correct population orientation is

$$p_{n_{\rm ds}} = \mathbb{P}\left[s_\star \sum_{\mu=1}^{n_{\rm ds}} \operatorname{sign}(\lambda_\mu)\left(\sqrt{\beta}\,M\lambda_\mu + \sqrt{q_\Gamma}\,Z_\mu\right) > 0\right], \tag{G.14}$$

where (λ_μ, Z_μ) are i.i.d. copies of (λ, Z).

The expected few-shot classification error is therefore

$$\epsilon_{\rm ds}^{(n_{\rm ds})} = p_{n_{\rm ds}}\epsilon_{\rm ds}^{\infty} + (1 - p_{n_{\rm ds}})(1 - \epsilon_{\rm ds}^{\infty}). \tag{G.15}$$

At $M = 0$, both orientations give chance-level performance, and we set $\epsilon_{\rm ds}^{\infty} = \epsilon_{\rm ds}^{(n_{\rm ds})} = 1/2$. Predictions with zero score are resolved by a fair random choice, including when the frozen feature vanishes identically.

These downstream formulas are conditional on the limiting overlaps (M, q_Γ). For the autoencoders, the replica saddle supplies their predicted values; the subsequent calculation evaluates the specified logistic probe.

Gaussian amplitudes. For $P_\lambda = \mathcal{N}(0,1)$, write $A = \beta M^2/q_\Gamma$ and $k = n_{\rm ds} \geq 1$. The population error is

$$\epsilon_{\rm ds}^{\infty}(A) = \frac{1}{\pi}\arccos\sqrt{\frac{A}{1+A}}. \tag{G.16}$$

Conditioning the orientation statistic in (G.14) on the amplitudes gives

$$p_k(A) = \mathbb{E}_{\lambda_1,\ldots,\lambda_k}\Phi\left(\sqrt{\frac{A}{k}}\sum_{\mu=1}^{k}|\lambda_\mu|\right), \qquad \lambda_\mu \overset{\rm iid}{\sim} \mathcal{N}(0,1), \tag{G.17}$$

where Φ is the standard normal cumulative distribution function. Together with (G.15), these expressions give the finite-label accuracy curve. Set $A = 0$ when the feature vanishes; the error is then $1/2$.

Interpretation The downstream scalar channel,

$$H = \sqrt{\beta}\,M\lambda + \sqrt{q_\Gamma}\,Z, \tag{G.18}$$

is controlled by the downstream signal-to-noise ratio

$$\frac{\beta M^2}{q_\Gamma} = \frac{\beta\,\theta_{\boldsymbol{u}^\star}^2}{q_\Gamma/q_I}, \qquad \theta_{\boldsymbol{u}^\star} = \frac{|M|}{\sqrt{q_I}}. \tag{G.19}$$

The ratio

$$\frac{q_\Gamma}{q_I} = \frac{\sum_i \gamma_i^\star \widehat{w}_i^2}{\sum_i \widehat{w}_i^2} \tag{G.20}$$

can therefore be interpreted as the effective noise level seen by the learned representation. Whereas β specifies the signal-to-noise ratio in the data model, (G.19) describes the signal-to-noise ratio after projection onto the learned direction. Hence downstream performance depends on two distinct properties of the pretrained representation: its alignment with the planted feature, measured by $\theta_{\boldsymbol{u}^\star}$, and its effective exposure to heteroskedastic noise, measured by q_Γ/q_I. Changes in the number of masks K or in the activation function can improve downstream performance through either mechanism.

AMP pretraining reference with the same probe. At the matched state-evolution solution of App. F, the unweighted AMP feature has $|M_{\mathrm{AMP}}| = \mathbb{E}_\gamma q_u(\gamma)$ and $q_{\Gamma,\mathrm{AMP}} = \mathbb{E}_\gamma[\gamma q_u(\gamma)]$ by the Nishimori identities. Its downstream SNR is therefore

$$A_{\mathrm{AMP}} = \frac{\beta\big(\mathbb{E}_\gamma q_u(\gamma)\big)^2}{\mathbb{E}_\gamma[\gamma q_u(\gamma)]}. \tag{G.21}$$

For Gaussian priors, solve (F.76), obtain $q_u(\gamma)$ from (F.73), and substitute A_{AMP} into (G.16), (G.17), and (G.15). At the uninformative fixed point set $A_{\mathrm{AMP}} = 0$. This is the green downstream reference in Fig. 3: Bayes-optimal pretraining followed by the same finite-label logistic probe as the autoencoders. Its identification with AMP performance requires the corresponding state-evolution solution to be attained. Algorithmically, freeze the AMP estimate as $\widehat{\boldsymbol{w}} = \widehat{\boldsymbol{u}}$, form the unweighted scores (G.8), and fit (G.9).

H Experimental settings and reporting

We report the settings needed to interpret the comparisons below. Source code and per-run configurations are supplied to reviewers. In all figures, error bars and shaded bands indicate ± 1 sample standard deviation (SD) across seeds around the empirical mean for each condition. Single-run results are shown without error bars.

H.1 Masking and augmentation protocols

Mask diversity and training budget. For finite K, each example has K independently drawn masks, fixed throughout pretraining. Static masking is $K = 1$; dynamic masking draws fresh masks on repeated presentations. The number of underlying examples is unchanged. In the shallow experiments, each update averages over the fixed mask collection, or uses one fresh mask for dynamic training (App. H.12). The CNN training schedule is specified below.

CNN data and evaluation. Fig. 1(b) and Fig. 4(b) use CIFAR-10 (Krizhevsky, 2009) with five independent encoder-initialization and data-split seeds. Each seed uses a balanced $40{,}000$-image training subset for both pretraining and labeled frozen probing. Masked pretraining and linear probing use $10{,}000$ held-out images for validation. The separate official $10{,}000$-image test set supplies final accuracy. One probe is fitted per encoder; points report mean test accuracy and bars show ± 1 SD across the five seeds. The masked comparisons share the split and initialization within each seed.

CNN architecture and preprocessing. All CNN conditions use the same $854{,}563$-parameter autoencoder. Images are scaled to $[0, 1]$ and standardized using channel statistics from the pretraining subset. The encoder receives the three masked image channels and a binary mask channel, with hidden pixels set to zero after standardization. Its four convolutions have output widths $(32, 64, 128, 128)$, kernel sizes $(3, 4, 4, 4)$, strides $(1, 2, 2, 2)$, padding one, and ReLU activations. The decoder uses three stride-two, kernel-four transposed convolutions with widths $(128, 64, 32)$, padding one and ReLU, followed by a kernel-three, padding-one convolution to RGB. There are no skip connections or normalization layers. Global spatial averaging of the encoder's $128 \times 4 \times 4$ output gives the frozen representation. Images contain 64 nonoverlapping 4×4 patches; each mask hides $\mathrm{round}(64\rho)$ patches chosen uniformly without replacement.

CNN optimization. For the main CNN sweeps in Figs. 1, 4, 14 and 15, we use AdamW (Loshchilov & Hutter, 2019) with batch size 512, learning rate 10^{-3}, weight decay 10^{-4}, and gradient norm clipping at 5. A validation-based plateau scheduler halves the learning rate, with patience five epochs and a minimum of 10^{-5}.

For finite K, each epoch traverses the fixed collection of $40{,}000K$ image–mask pairs. Dynamic masking draws a fresh mask at every presentation throughout training. We group 32 presentations per image into one epoch for validation and scheduling, corresponding to $2{,}500$ optimizer updates. Under dynamic masking, each image therefore receives $32E$ independent mask draws over E epochs.

Validation uses one fixed mask per held-out image after each epoch. We retain the checkpoint with the lowest validation reconstruction MSE and stop after 15 epochs without improvement, with a minimum

of 20 and a ceiling of 460 epochs. Under these parameters, our setting is always early-stopped and not undertrained.

CNN linear probes. Probes receive unmasked, unaugmented images and use the frozen 128-dimensional representation, standardized coordinatewise using the probe training set. A linear ten-class head is fitted with cross-entropy, AdamW, batch size 512, constant learning rate 10^{-2}, and zero weight decay. Classifier initialization and minibatch-shuffling seeds are fixed across encoders. The head with lowest validation cross-entropy is retained; training stops after 25 epochs without improvement, with a minimum of 20 and a ceiling of 200 epochs. All reported masked-model probes satisfy this stopping criterion. The unmasked checkpoint sweep uses the same probe settings with a 300-epoch ceiling (App. H.8).

ViT architecture and data. Our implementation adapts the U-MAE code of Zhang et al. (2022b), which builds on MAE (He et al., 2022).[2] We use ViT-Base/16 on ImageNet-100, with 126,689 training and 5,000 validation images across 100 classes. Images have resolution 224×224, yielding 196 patches. The encoder has 12 blocks, width 768 and 12 attention heads; the decoder has 8 blocks, width 512 and 16 heads, with MLP expansion ratio 4 throughout. Pretraining minimizes MSE on masked patches, using targets standardized within each patch. U-MAE's uniformity penalty is disabled in all reported ViT runs; AdamW weight decay is retained. The inherited online classifier receives detached features and does not train the encoder. Reported accuracies use a separately fitted frozen probe.

ViT preprocessing and masking. Without stochastic cropping or flipping, images are resized bicubically to shorter side 256 and center-cropped to 224×224. The augmented condition instead uses random resized crops with area fraction in $[0.2, 1]$ and aspect ratio in $[3/4, 4/3]$, followed by horizontal flipping with probability $1/2$. Both conditions use ImageNet channel normalization. A random patch permutation retains $\lfloor 196(1-\rho) \rfloor$ visible patches. Static masking fixes this permutation per training image; dynamic masking redraws it at every presentation. With cropping and flipping, the static patch positions refer to the transformed image, whose content can change across epochs.

ViT pretraining. All masked runs in Fig. 1(c) and Fig. 5 use 100 epochs of AdamW with momentum parameters $(0.9, 0.95)$, weight decay 0.05, peak learning rate 7.5×10^{-5}, ten warmup epochs and cosine decay to zero. The effective batch size is 128: batches of 64 with two accumulation steps without cropping/flipping, and batches of 128 without accumulation in the augmented condition. Biases and normalization parameters are excluded from weight decay. We probe the final pretraining checkpoint.

The comparison uses one common seed, with identical training and validation splits, initial autoencoder weights, optimizer settings, and pretraining budget across conditions. A separate frozen probe is fitted to each encoder using the same probe initialization and training and evaluation protocol. In contrast, the unaugmented curves in Fig. 1(c) summarize three encoder initializations (seeds $0, 1, 2$), with means and ± 1 SD.

ViT linear probes. The frozen encoder processes all patches. Its 768-dimensional class-token representation feeds a non-affine batch-normalization layer and a linear 100-class head. The head is trained with cross-entropy and LARS (You et al., 2017), batch size 256, momentum 0.9, zero weight decay, peak learning rate 0.1, ten warmup epochs and cosine decay to zero over 50 epochs. Probe training uses random resized crops with area fraction in $[0.08, 1]$, aspect ratio in $[3/4, 4/3]$, and horizontal flipping with probability $1/2$ for every pretraining condition. Validation uses the deterministic resize and center crop above. Thus the no-cropping/no-flipping condition refers to pretraining. Probe seeds are fixed at zero, and we report validation top-5 accuracy at the final probe epoch. Unmasked checkpoint selection is described in App. H.8.

Table 4 ablates the use of stochastic augmentation during linear probing. We find that including or omitting stochastic augmentation yields quantitatively similar downstream performance.

Longer ViT training. The 100-epoch protocol above is used to compare masking conditions under a common training budget. To assess the attainable absolute performance with the same architecture,

[2] U-MAE implementation: https://github.com/zhangq327/U-MAE.

we also adapt the longer training schedules of U-MAE and MAE (Zhang et al., 2022b; He et al., 2022), retaining dynamic masking at $\rho = 0.75$, no stochastic cropping or flipping during pretraining, and no uniformity penalty. Table 3 reports these checks, all using initialization zero and ImageNet-100 pretraining only. The 200-epoch schedule uses effective batch 1024 (128×8 accumulation), 40 warmup epochs and peak learning rate 6×10^{-4}. The longer MAE schedule uses effective batch 4096 (64×64 accumulation), 40 warmup epochs and peak learning rate 2.4×10^{-3}; it was stopped early, and the evaluated checkpoint is after 641 epochs of the planned 1600-epoch cosine schedule. Other pretraining settings remain as above.

Each encoder is evaluated with the existing probe and with a 90-epoch probe using effective batch 16,384 (512×32 accumulation), peak learning rate 6.4 and ten warmup epochs, following the MAE linear-probing recipe. Both probes use cropping and flipping. The larger batch is implemented by gradient accumulation; batch normalization uses the physical batch, and incomplete accumulation windows at the end of an epoch do not produce an optimizer update. With the 200-epoch encoder, the longer probe reaches 61.50% top-1 and 86.04% top-5 accuracy without pretraining augmentation. For context, Zhang et al. (2022b) report 61.2% and 67.5% top-1 for MAE and regularized U-MAE, respectively, on ImageNet-100 with ViT-Base. These published results provide a reference under different training conditions. The additional runs show that the shorter protocol used for our masking comparisons does not exhaust the performance of the architecture; they vary the optimization schedule as well as batch size. See Baevski et al. (2023) for experiments showing the effect of the batch size on downstream performance.

Pretraining			Linear probe		Accuracy (%, ↑)	
Schedule	Epochs	Batch	Epochs	Batch	Top-1	Top-5
Main comparison	100	128	50	256	39.96	68.78
U-MAE	200	1024	50	256	47.56	76.14
U-MAE	200	1024	90	16384	61.50	86.04
MAE	641	4096	50	256	51.38	79.60
MAE	641	4096	90	16384	59.12	84.76

Table 3: **Higher accuracy with longer ViT training schedules.** Frozen-probe validation accuracy for dynamic masking at $\rho = 0.75$, without cropping or flipping during pretraining. Batch sizes are effective batches. All rows use the same initialization; repeated pretraining rows probe the same encoder checkpoint. Schedule names indicate adaptations of the published optimization recipes, with uniformity regularization disabled. Results use one run per configuration and the final probe epoch.

Removing augmentation from the ViT probe. We also repeat the 50-epoch probes of the static and dynamic 100-epoch encoders at $\rho = 0.75$, using initialization zero and deterministic resize/center crop during probe training. All other probe settings and the pretrained checkpoints are held fixed. Table 4 shows that removing cropping and flipping raises top-5 accuracy by 2.28 and 2.08 percentage points for dynamic and static masking, respectively. Their gap remains nearly unchanged (17.58 versus 17.78 points), indicating that probe augmentation does not account for the dynamic-masking advantage in this check.

	Probe with crop/flip		Probe without crop/flip	
Pretraining masks	Top-1	Top-5	Top-1	Top-5
Static	25.34	51.20	26.22	53.28
Dynamic	39.96	68.78	41.98	71.06

Table 4: **The masking advantage persists without probe augmentation.** ViT-Base/ImageNet-100 validation accuracy (%, ↑), using the same pretrained encoders and probe hyperparameters in each pair. Both encoders were pretrained without cropping or flipping at $\rho = 0.75$. Results use one initialization and one probe seed.

H.2 Finite-size effects of PCA

Finite-dimensional cosine similarity. Fig. 6 uses the same bounded-noise setting as Fig. 2(b), with $\gamma = \frac{1}{4} + \frac{9}{4}B$, $B \sim \mathrm{Beta}(1, 2)$, and $\beta = 1$, while varying the dimension d and sample complexity α. It examines the residual finite-dimensional PCA alignment visible in that main-text comparison. In Fig. 6, the observed cosine decreases as the dimension increases, including at the largest values of α shown, consistent with the high-dimensional scaling $A_\alpha d^{-1/2}$ and supporting its interpretation as a finite-size effect. We use uncentered, unstandardized PCA at $\beta = 1$, dimensions $d = 100 \times 2^j$ for $j = 0, \ldots, 8$, ten logarithmically spaced values of α from 1 to 50, and 30 seeds. The plotted observable is the mean absolute cosine, with ± 1 SD.

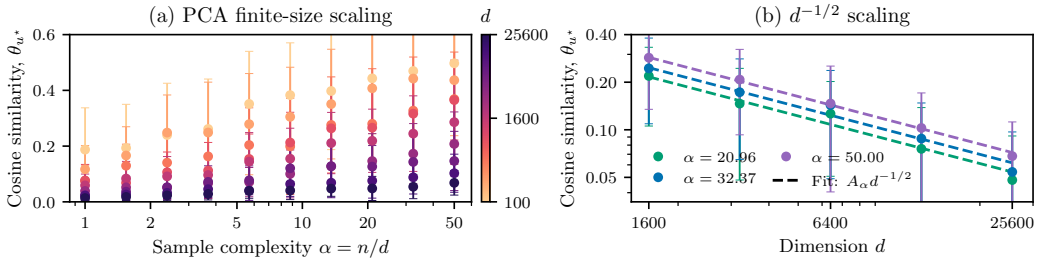


Figure 6: **Finite-dimensional PCA cosine similarity decreases with dimension.** Same bounded-noise setting as Fig. 2(b): $\gamma = \frac{1}{4} + \frac{9}{4}B$, $B \sim \mathrm{Beta}(1, 2)$, and $\beta = 1$. **(a)** Cosine similarity versus sample complexity, with color indicating dimension size d. **(b)** Dimension dependence at fixed α. Dashed lines fit $A_\alpha d^{-1/2}$ for $d \geq 1600$, with exponent fixed. Bars show ± 1 SD over 30 seeds.

H.3 Top-1 accuracy for ViT

Fig. 7 reports the top-1 counterparts of the ViT comparisons in Fig. 1(c) and Fig. 5, using the same pretrained encoders and frozen linear probes. The qualitative conclusions also hold for this metric: dynamic masking improves accuracy without cropping or flipping during pretraining, while these augmentations substantially improve static masking and narrow the gap to dynamic masking.

Table 3 puts these absolute accuracies in context. At $\rho = 0.75$, dynamic masking reaches 39.96% top-1 under the main comparison protocol, increasing to 61.50% and 59.12% with the longer U-MAE and MAE schedules and 90-epoch probes, respectively. All three settings use no cropping or flipping during pretraining. Thus, the shorter protocol used to compare masking conditions does not exhaust the architecture's performance; these additional runs vary both the training schedule and batch size.

Table 4 separately checks standard augmentation during linear probing. Removing probe cropping and flipping raises top-1 accuracy from 39.96% to 41.98% for dynamic masking and from 25.34% to 26.22% for static masking. The dynamic–static gap therefore persists (14.62 versus 15.76 percentage points), supporting the same conclusion with either probe protocol. Both tables report single-seed checks, distinct from the three-seed averages in Fig. 7(a).

H.4 Noise-aware spectral methods

Fig. 8 adds diagonal-deletion PCA, HeteroPCA, and estimated-whitening PCA to the bounded and unbounded settings of Fig. 2(b,c), with $\beta = 1$ and $\rho = 0.75$ for dynamic masking. Spectral estimators and autoencoders use $d = 2000$ and 20 seeds; AMP uses $d = 4000$ and ten seeds (App. H.14). Bars show ± 1 SD. The added spectral methods use the same observations and seeds as ordinary PCA and receive neither the spike nor the true coordinate variances. Covariances are centered in the unbounded setting and uncentered in the bounded setting.

Spectral estimators. Write $S = \widehat{\Sigma}$ the empirical covariance matrix, $\mathrm{diag}(S)$ for its diagonal matrix, and $\mathrm{off}(S) = S - \mathrm{diag}(S)$. In this section we will work with the following additional spectral estimators:

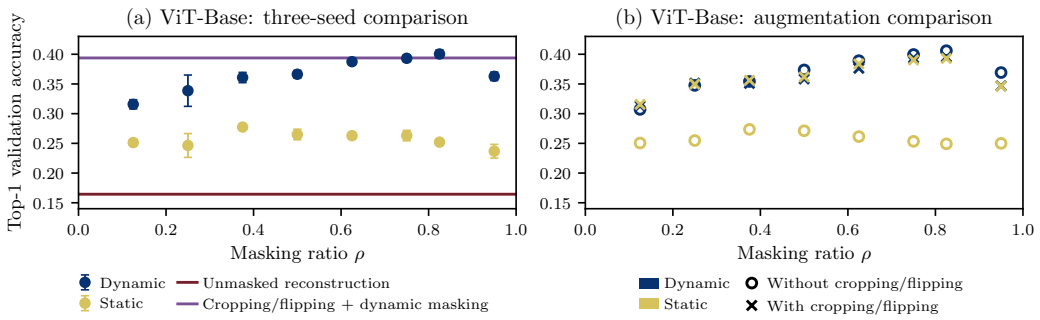


Figure 7: **The ViT masking comparison also holds for top-1 accuracy.** Frozen linear-probe validation accuracy on ImageNet-100. **(a)** Static and dynamic masking without cropping or flipping: mean ± 1 SD over three seeds. Each horizontal reference uses one seed. **(b)** Static and dynamic masking with and without cropping and flipping, using the common seed and protocol of Fig. 5; no error bars are shown. Training and evaluation settings are given in App. H.1.

Diagonal-deletion PCA removes the empirical diagonal before spectral extraction, using $\mathrm{off}(S)$. This baseline is studied alongside HeteroPCA by Zhang et al. (2022a) but was originally introduced by Florescu & Perkins (2016). We extract its largest-magnitude eigenpair, corresponding to a leading singular direction. This removes the direct diagonal noise contribution, but also discards the signal diagonal.

HeteroPCA iteratively imputes the diagonal from a rank-one approximation while retaining the observed off-diagonal entries (Zhang et al., 2022a, Algorithm 1). We initialize with $\mathrm{off}(S)$, retain the signed largest-magnitude eigenpair, and stop when the relative diagonal change is at most 10^{-8}. Unresolved runs are continued up to 10^5 iterations; all 680 fits in the displayed comparison meet this criterion.

Estimated-whitening PCA follows the noise-whitening approach of Leeb & Romanov (2021), who also discuss estimated noise covariances in Section 4.3. We estimate $\widehat{\Gamma}_{ii} = \max(S_{ii}, 10^{-12})$ and apply PCA to $S_{\mathrm{W}} = \widehat{\Gamma}^{-1/2} S \widehat{\Gamma}^{-1/2}$. If $\boldsymbol{v}_{\mathrm{W}}$ is its leading eigenvector, alignment is evaluated using $\widehat{\Gamma}^{1/2}\boldsymbol{v}_{\mathrm{W}}$ in the original coordinates. This baseline uses whitening and unwhitening without additional spectral shrinkage.

Dynamic masking achieves higher mean alignment than diagonal-deletion PCA and HeteroPCA across both displayed sweeps, without explicit noise estimation or whitening. Neither the coordinate variances nor their distribution are supplied to the autoencoder: adaptation to heterogeneity emerges through the masked prediction objective (App. D.4). This provides a complementary perspective to covariance-aware inference. For example, Zhang & Mondelli (2024) construct spectral estimators using known noise covariances and signal strength, with optimality guarantees under their assumptions. Estimated-whitening PCA generally achieves higher alignment in our experiments and lies closer to the Bayes reference, using variances estimated from the same observations. The comparison therefore highlights the ability of masked prediction to learn informative structure without a separate procedure for estimating and correcting the noise.

The whitening-PCA threshold. The whitening threshold follows from the spiked-model analysis of Leeb & Romanov (2021). In our rank-one model with Gaussian latent priors, whitening by $(\Gamma^\star)^{-1/2}$ produces isotropic noise and a population signal eigenvalue $\ell^{\mathrm{W}} = \beta\|(\Gamma^\star)^{-1/2}\boldsymbol{u}^\star\|^2/d \to \beta\mathbb{E}_\gamma[\gamma^{-1}]$. Their recovery condition is $\ell^{\mathrm{W}} > \alpha^{-1/2}$, with the same onset after unwhitening, giving $\alpha_c^{\mathrm{W}} = [\beta^2\mathbb{E}_\gamma[\gamma^{-1}]^2]^{-1}$.

Leeb & Romanov (2021) also justify replacing the noise covariance by an estimate consistent in operator norm. For diagonal noise and a delocalized signal direction, they obtain such an estimate from the sample coordinate variances, as in the baseline above. Our bounded variance law and normalized Gaussian spike satisfy these conditions, so the same threshold applies to estimated whitening in that setting. For the shifted-lognormal law, α_c^{W} is the ideal-whitening reference; their

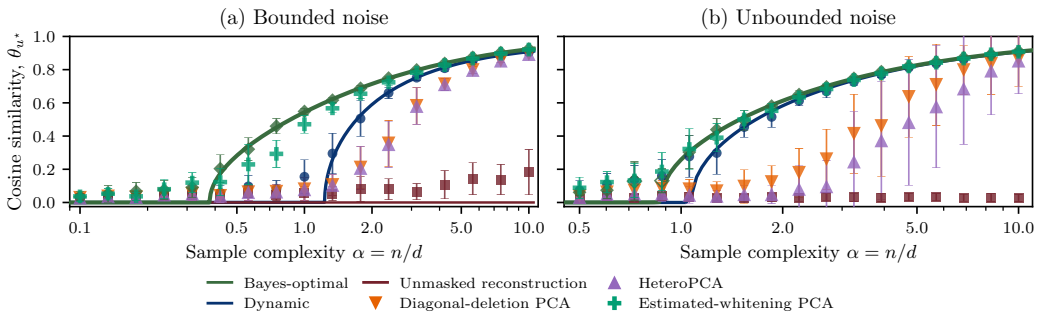


Figure 8: **Noise-aware spectral baselines for heteroskedastic noise.** Cosine similarity versus sample complexity under **(a)** bounded and **(b)** unbounded heterogeneous noise. The comparison includes dynamic masking, unmasked reconstruction (ordinary PCA), diagonal-deletion PCA Florescu & Perkins (2016), HeteroPCA (Zhang et al., 2022a), estimated-whitening PCA (Leeb & Romanov, 2021), and the Bayes-optimal baseline. Lines show theoretical references; markers show numerical estimates. Green curves give the Bayes-optimal state-evolution reference; green diamonds show AMP results (Apps. F.5 and H.14). Bars show ± 1 SD; hollow AMP markers identify groups containing iteration-limited runs.

bounded-spectrum assumptions do not directly cover the estimated-whitening algorithm, whose performance is shown empirically.

Relation to the masking and Bayes thresholds. Our replica analysis connects this established whitening threshold to dynamic masking. Combining the Bayes expression (F.67) with the lower bound in (D.48), specialized to dynamic linear masking in (D.133), gives

$$\alpha_c^{\mathrm{BO}} = \frac{1}{\beta^2 \mathbb{E}_\gamma[\gamma^{-2}]} \le \alpha_c^{\mathrm{W}} = \frac{1}{\beta^2 \mathbb{E}_\gamma[\gamma^{-1}]^2} \le \alpha_{c,\mathrm{RS}}^{\mathrm{dyn}}. \tag{H.1}$$

The inverse moments are finite for both displayed laws. The first inequality is strict for a nonconstant variance law by Cauchy–Schwarz. It orders the recovery onsets; the relative alignments above threshold are assessed by the simulations. Thus masking can outperform diagonal correction while remaining less effective than whitening at the onset of recovery.

H.5 Dynamic masking and nonlinear autoencoders

Fig. 9 compares dynamic masking with the no-masking reconstruction, Eq. (5), for linear, ReLU, ELU, and tanh activations. We use the same setting as Fig. 3.

Across the four activations, dynamic masking develops substantially higher alignment, whereas full reconstruction retains low mean alignment throughout the displayed range. Thus, changing the activation of the unmasked autoencoder does not help to recover the spike. On the other hand, for dynamic masking, linear, ELU, and tanh dynamic curves are close, while ReLU requires larger sample complexity to attain comparable alignment.

H.6 Reconstruction errors across mask diversity

Fig. 10 evaluates the centered training and test errors defined in App. G.1, using linear activation, $\alpha = 8$, $\beta = 1$, $\lambda_R = 10^{-4}$, and the bounded variance law of Fig. 2(b). Again, $d = 2000$, 30 seeds. Each loss subtracts the zero-decoder loss on the same samples and masks, excludes the ridge penalty, and is reported per sample without division by ρ. Test errors use independent samples and fresh masks; dynamic training errors average fresh masks on the training samples.

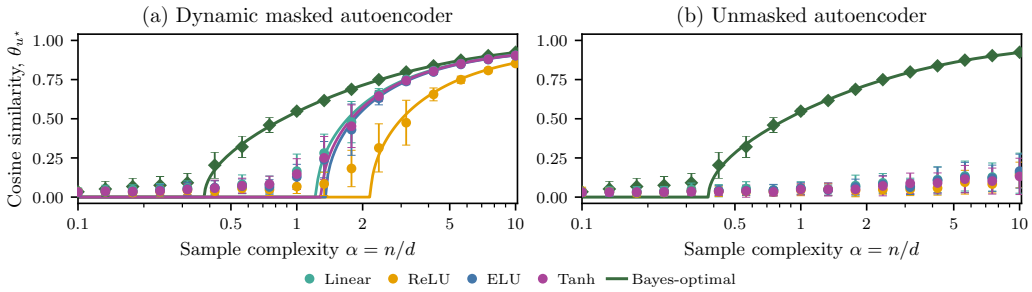


Figure 9: **Masking improves recovery across activation functions.** Cosine similarity for **(a)** dynamic masking and **(b)** no masking in the same setting as Fig. 3. Colored curves are replica predictions, colored markers in both panels are ERM simulations at finite $d = 2000$, averaged across 30 seeds ± 1 SD. In both panels, the green curve is the Bayes-optimal prediction with the AMP.

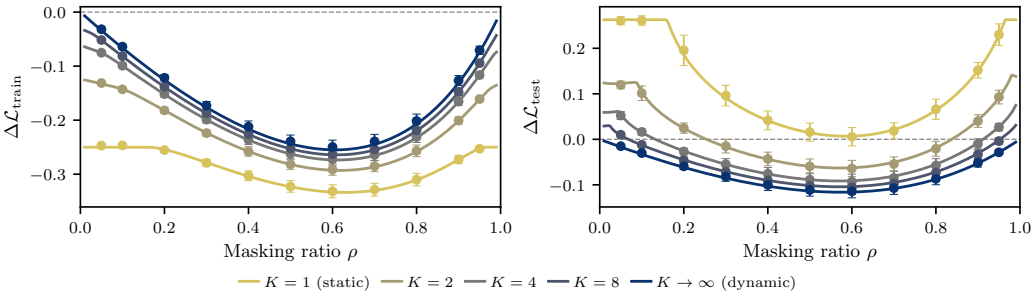


Figure 10: **Mask resampling worsens training error while improving test loss.** Centered reconstruction errors versus masking ratio at $\alpha = 8$ under bounded heterogeneous noise. Left: training error. Right: independent test error. Curves are the replica predictions in (G.3) and (G.4); markers and bars show mean ± 1 SD over 30 simulations. Lower values are better; zero denotes the zero-decoder baseline. Colors distinguish the fixed mask collections and dynamic masking.

H.7 Number of downstream labels

Fig. 11 varies the number of independently labeled examples used to fit the frozen scalar logistic probe. We reuse the linear dynamic-masking replica solution from Fig. 3, with bounded noise, $\beta = 1$, $\rho = 0.75$, and $\lambda_R = 10^{-4}$, and evaluate (G.15) for $n_{\text{ds}} \in [1, 128]$. The pretraining overlaps remain fixed for each α: additional labels only change the probability that the probe selects the correct orientation of the representation. We evaluate (G.17) with 10^5 Monte Carlo amplitude samples, except at $n_{\text{ds}} = 1$, where $p_1(A) = \frac{1}{2} + \pi^{-1} \arctan \sqrt{A}$ is exact.

Accuracy improves with the label count and approaches the population-orientation limit rapidly away from the recovery transition. Increasing n_{ds} therefore improves the finite-label readout without changing the pretraining recovery threshold.

H.8 Early-stopping without masking

Figs. 12 and 13 evaluate frozen linear probes at intermediate unmasked-pretraining checkpoints for the architectures and datasets of Fig. 1. In both architectures, training and validation reconstruction MSE continue to improve after downstream accuracy has peaked. The deterioration in the representation is therefore not accompanied by a rise in held-out reconstruction error.

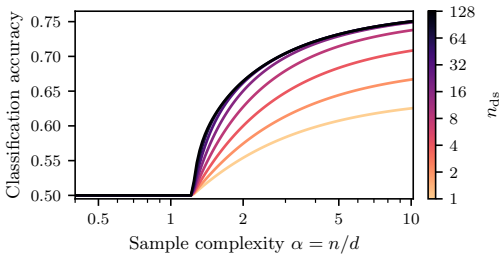


Figure 11: **Additional labels resolve the orientation of a frozen feature.** Predicted downstream classification accuracy for linear dynamic masking versus sample complexity. The colorbar gives n_{ds} on a logarithmic scale. All curves use the same pretraining solution at each α and differ only in the size of the independent labeled dataset used by the scalar logistic probe.

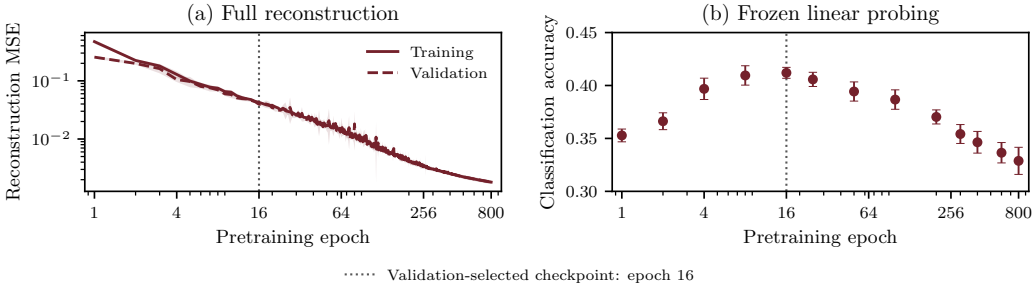


Figure 12: **Better reconstruction need not improve CNN representations. (a)** Training and validation reconstruction MSE during unmasked CNN pretraining. **(b)** Linear-probe test accuracy at different pretraining epochs. Shading and bars show ± 1 SD over five seeds. The dotted line marks epoch 16, selected by mean probe validation accuracy for Fig. 1.

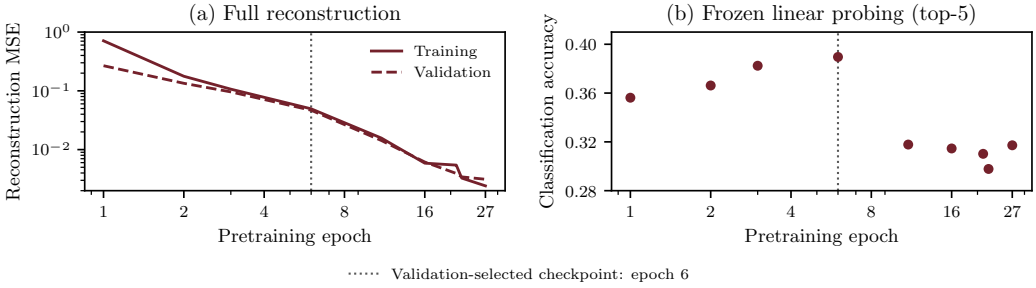


Figure 13: **Reconstruction and downstream performance also diverge for ViT. (a)** Training and validation reconstruction MSE during unmasked ViT pretraining. **(b)** Linear-probe top-5 accuracy on the validation set at different pretraining epochs. The dotted line marks six completed pretraining epochs, selected by probe validation accuracy for Fig. 1. Results use one pretraining seed.

For CNN, epoch 16 maximizes mean probe validation accuracy across seeds and supplies the unmasked baseline in Figs. 1 and 4; the plotted curve reports test accuracy. Reconstruction losses are measured in the standardized image coordinates.

For ViT, unmasked pretraining uses the architecture, deterministic preprocessing and optimizer settings of App. H.1, retaining all patches and evaluating reconstruction MSE over all patches. Targets remain standardized within each patch, and the learning-rate schedule retains its 100-epoch horizon. The final probe's validation top-5 accuracy selects six pretraining epochs (38.96%). Fig. 13(a) shows normalized-patch reconstruction MSE, excluding the detached online classifier's loss; validation

MSE is evaluated on all 5,000 validation images. Neither the CNN nor the ViT checkpoint is selected by minimum reconstruction validation loss since it is misleading for downstream performance.

Masked CNNs use validation-selected reconstruction checkpoints, whereas the main **masked** ViT comparisons use the final pretraining checkpoint. This selection favors the unmasked reference, whose reported score is a selected validation result rather than an independent test estimate.

H.9 Mask renewal and image augmentation in CNNs

Fig. 14 complements Fig. 5 using the CNN autoencoder on CIFAR-10 from Fig. 1(b). We compare static and dynamic masking with and without stochastic cropping and flipping, averaging over five seeds. Within each seed, the four conditions share the data split and initial autoencoder weights. The architecture, optimizer settings, validation-based checkpoint selection and stopping rule, and frozen linear-probing protocol follow App. H.1.

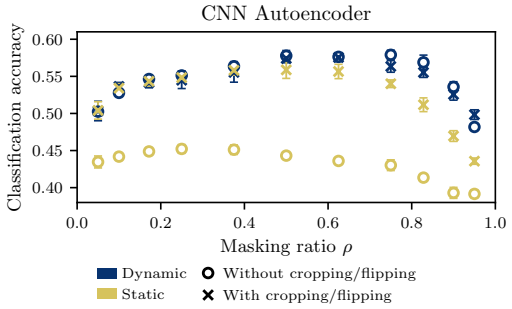


Figure 14: **Mask renewal and image augmentation in CNNs.** Frozen linear-probe test top-1 accuracy on CIFAR-10 versus masking ratio ρ. Blue/yellow indicate dynamic/static masking; circles/crosses denote pretraining without/with stochastic cropping and flipping. Each marker reports the mean over five seeds, with data splits and autoencoder initializations shared across conditions within each seed. Linear probes use no image augmentation.

Cropping and flipping substantially improve static masking across the sweep, bringing its mean accuracy close to dynamic masking at lower masking ratios. At larger masking ratios, dynamic masking retains an advantage even with augmentation. The CNN comparison therefore supports the same interpretation as the ViT experiment: geometric transformations can partly compensate for mask reuse by changing the content to be predicted. The remaining gap shows that the extent of this compensation depends on the masking ratio and experimental setting.

H.10 CNN training budgets and validation selection

We examine the CNN mask-diversity comparison of Fig. 4(b) at $\rho = 0.625$, with $K \in \{1, 2, 4, 8, 16, 32\}$ and dynamic masking. Both experiments use the CIFAR-10 architecture, AdamW settings, and the frozen linear-probing protocol of App. H.1, without cropping or flipping. Each condition has five seeds (different pretraining/validation splits). Encoders are selected by validation reconstruction MSE followed by downstream accuracy measured on a separate test set.

Fig. 15 shows the original training and validation trajectories with the *plateau learning-rate* (LR) scheduler. Finite-mask runs continue reducing training error after validation error levels off or rises (*overfitting*), motivating validation-based checkpoint selection. Since the original scheduler operates in epochs whose update counts depend on K, we also test whether the results persist under a common schedule measured in optimizer updates.

To assess the role of the scheduler, we repeat the experiments with the same data splits, random seeds, and initial weights, using a fixed learning rate (LR) of 10^{-3} and a longer training number of optimizer updates (Fig. 16). All conditions now use 2,500 updates between validation evaluations, a stopping patience of 125,000 updates without improvement, and a total maximum of 1,150,000 updates.

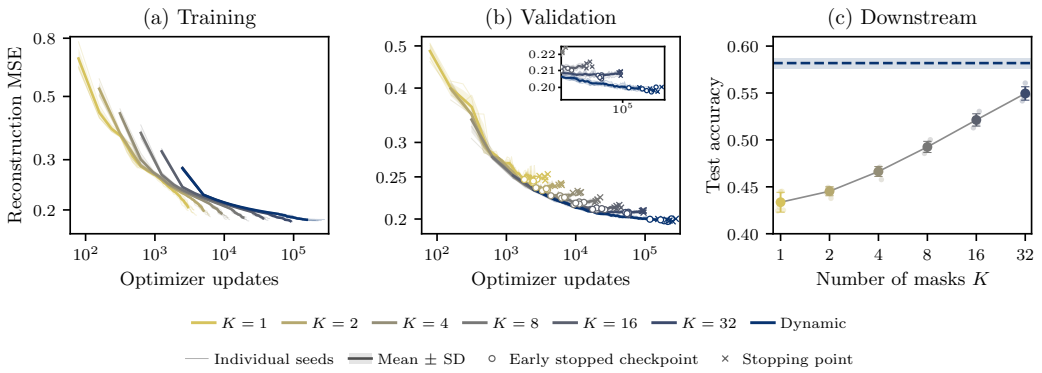


Figure 15: **CNN training under the original protocol. (a,b)** Training and validation reconstruction MSE versus optimizer updates. Thin traces show individual seeds; thick curves and shading show mean ± 1 SD while all five runs are observed. Circles and crosses mark early stopped and final checkpoints, respectively. **(c)** Frozen linear-probe test accuracy of early-stopped encoders, with ± 1 SD. The horizontal dashed line indicates dynamic masking.

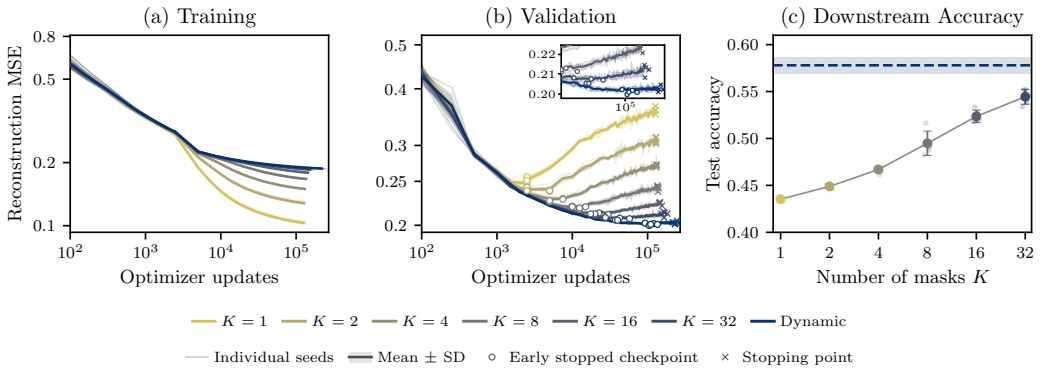


Figure 16: **CNN control with a fixed learning rate and extended training.** Panels and uncertainty conventions follow Fig. 15.

Fig. 17(a) compares validation-selected (early-stopped checkpoints) encoders under the two protocols. Their mean test accuracies remain close across masking conditions, and the downstream improvement with mask diversity persists under the fixed-learning-rate control. The extended runs also exhibit a reconstruction-generalization gap.

Fig. 17(b) instead selects the encoder checkpoint saved at step 125,000 from each fixed learning rate run (Fig. 16) and evaluates it using the same frozen linear-probing protocol. This changes the checkpoint used for downstream evaluation, not the training or stopping procedure. Thus, the downstream mask-diversity trend persists both when selecting the encoder by validation reconstruction MSE and when selecting its checkpoint at a common training step.

H.11 Numerical assessment of finite-mask threshold ratios

How well does the weak-signal mask-reuse law describe recovery at finite signal strength? Fig. 18 compares it with thresholds computed from the full linear equations at zero ridge. We solve the uninformative closures in (D.12)–(D.16) and locate the gain crossing in (D.23); App. H.12 gives the numerical procedure.

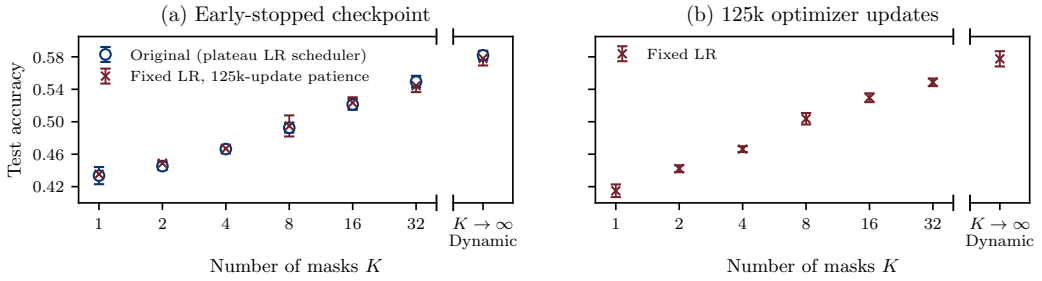


Figure 17: **The downstream mask-diversity trend persists across training protocols.** Test accuracies from **(a)** are taken from linear probing the early-stopped checkpoints from Figs. 15 and 16. **(b)** corresponds to linear probing the checkpoints after 125k optimizer updates in the setting of Fig. 16. Markers and bars show mean ± 1 SD over five seeds.

Variance laws and normalization. The two rows of Fig. 18 use homogeneous noise, $\gamma = 1$, and the bounded law

$$\gamma = 1 + c\left(\frac{9}{4}B - \frac{3}{4}\right), \qquad B \sim \mathrm{Beta}(1,2), \qquad c \simeq 0.5743390142. \tag{H.2}$$

This bounded law differs from the example in Fig. 2(b).

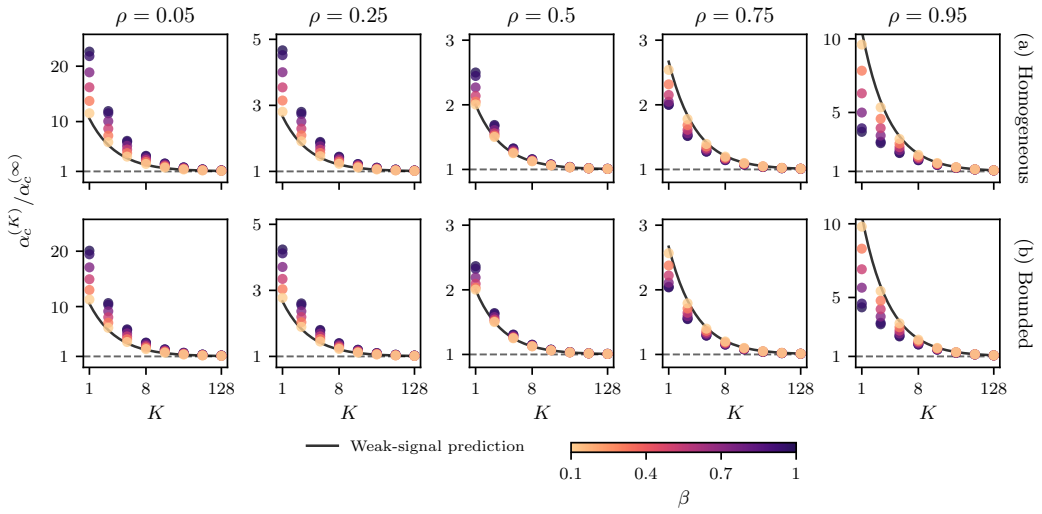


Figure 18: **Finite-mask threshold ratios under homogeneous and bounded noise.** Rows show **(a)** homogeneous variance distribution $\gamma = 1$ and **(b)** the bounded law (H.2); columns vary the masking ratio ρ. Colors indicate signal strength β. Markers are numerical teacher-overlap instability thresholds divided by their corresponding dynamic threshold ($K \to \infty$). The black curve is the weak-signal ratio prediction $1 + K^{-1}([2\rho(1-\rho)]^{-1} - 1)$; the dashed line marks unity.

Across the displayed parameters, increasing mask diversity lowers the threshold toward the dynamic reference. The finite-signal deviations from the weak-signal curve depend on ρ, while the diminishing cost of mask reuse remains visible.

H.12 Synthetic experiments and numerical predictions

Data and masking-ratio sweeps. The synthetic autoencoder simulations use $P_u = P_\lambda = \mathcal{N}(0,1)$ and $\beta = 1$. The masked results in Fig. 1(a) and Fig. 4(a) use linear activation, $d = 2000$, $\alpha = 8$, and $\gamma = \frac{1}{4} + \frac{9}{4}B$ with $B \sim \mathrm{Beta}(1,2)$, as in Fig. 2(b). Each condition uses 30 seeds and 4000 full-batch

Adam updates. Fig. 10 uses the same experiments. Bars show ± 1 SD. The red PCA references in Fig. 1(a) and Fig. 4(a) show the asymptotic chance accuracy $1/2$ for this law (Proposition E.4 and App. E.1.7).

Sample-complexity sweeps. Fig. 2 uses the three variance laws stated in its caption, $d = 2000$, $\sigma(x) = x$, $\rho = 0.75$, 20 seeds per point, and 4000 Adam updates for each autoencoder. The no masking points are found by extracting the first principal component of the pretraining set and then linear probing following the procedure detailed in Sec. 1.

In Fig. 3, the four activations use the bounded law $\gamma = \frac{1}{4} + \frac{9}{4}B$, $B \sim \mathrm{Beta}(1, 2)$, as in Fig. 2(b), with $d = 2000$, $\beta = 1$, and $\rho = 0.75$. Each autoencoder condition uses 30 seeds and 4000 full-batch Adam updates. Both rows report means with ± 1 SD across seeds. AMP dimensions and seeds are given in App. H.14.

Optimization and downstream evaluation. Adam uses learning rate 0.05, no schedule or weight decay, and one initialization with iid $\mathcal{N}(0, 1)$ coordinates, without warm starts across conditions. Each update averages the reconstruction loss over all hidden entries in the fixed mask collection, or over one fresh mask per sample for dynamic training. The masked autoencoders in Fig. 1(a) and Fig. 4(a), the bounded autoencoders in Fig. 2(b), and all masked autoencoders use $\lambda_R = 10^{-4}$. Sampled coordinate variances are not renormalized or truncated. Frozen scalar probes use $n_{\mathrm{ds}} = 100$ independently labeled examples and 10,000 independent test examples from the same quenched spike, with labels $\mathrm{sign}(\lambda)$. For every activation, probing uses the raw, unmasked preactivation $\boldsymbol{w}^\top \boldsymbol{x}/\sqrt{d}$. The scalar logistic coefficient has no intercept or regularization and is fitted with L-BFGS for at most 100 iterations.

Numerical predictions. The RS learning curves use the small ridge $\lambda_R = 10^{-4}$, separately from the zero-ridge threshold calculation in Fig. 18. For the masking-ratio sweeps in Fig. 1(a), Fig. 4(a), and Fig. 10, we use 200-node Gauss–Legendre quadrature over the full bounded support $[1/4, 5/2]$ and analytic linear output averages.

For Fig. 3, variance integrals use 200-node Gauss–Legendre quadrature over the full bounded support $[1/4, 5/2]$; the Bayes reference uses 40 variance nodes. The RS curves use a 120-point grid over $\alpha \in [0.09, 11]$, with denser sampling near the transitions. Linear output averages are analytic. Antithetic scrambled Sobol field integration uses 2^{18} samples for tanh at $K = 1$ and ReLU at $K = 2, 4$, and 2^{14} for ELU at $K = 1, 2, 4$ and tanh at $K = 2, 4$. ReLU at $K = 1$ and dynamic masking uses Gauss–Hermite field quadrature of order 120. Dynamic tanh and ELU use order 64 for both field and mask quadrature. Finite-K ReLU uses its piecewise-quadratic proximal; tanh and ELU use numerical proximal minimization. Fixed-point updates use damping 0.3 and tolerance 10^{-6}, except tanh at $K = 1$ (10^{-7}) and dynamic tanh (10^{-8}). Downstream curves include finite-label orientation uncertainty through (G.15), evaluated with 10^4 Monte Carlo draws for the masking-ratio sweeps and 10^5 for the four-activation comparison.

Finite-mask threshold ratios. For Fig. 18, we use linear activation and $\lambda_R = 0$, evaluate the output channel analytically, and follow the uninformative saddle on a logarithmic grid in α. The first crossing of $\Lambda_K = 1$ in (D.24) is refined by logarithmic bisection. Each finite-K threshold is normalized by the dynamic threshold at the same β, ρ, and variance law. The bounded distribution is defined in (H.2) and differs from the law used in Fig. 2(b).

Matched-moment noise-adaptation examples. We evaluate the finite-support expectations $I_{j,\ell}(\tau)$ exactly as weighted sums. For a fixed β, parameterize the critical curve by τ using (D.131), then recover ρ from (D.134). Follow the regular continuation through $\tau = 0$ and solve for the specified masking ratio. Reconstructing the full order parameters verifies the susceptibility, variance, and teacher-response closures. For the three laws in Table 2, the homogeneous law, and $\frac{1}{2}\delta_{1/2} + \frac{1}{2}\delta_{3/2}$, these checks at $\rho \in \{0.8, 0.95, 0.99\}$ give absolute critical closure residuals below 2×10^{-13}. Solutions of (D.140) at sample ratios 1.001, 1.01, and 1.05 times each critical ratio have positive alignment and satisfy the full closures to absolute tolerance 10^{-7}. These checks support the continuous informative continuation used for the reported thresholds.

H.13 MASK DIVERSITY IN BERT

We compare one fixed mask per sequence ($K = 1$), a fixed collection of ten masks ($K = 10$), and dynamic mask resampling in a BERT-Medium pretraining pilot. Each condition uses three pretraining seeds and two fine-tuning seeds per pretraining seed and task. The corpus, architecture, and pretraining budget are shared across masking conditions.

Implementation and data. We train a BERT-style masked language model (Devlin et al., 2019) using our adaptation of DinkyTrain[3] (Wettig et al., 2023), built on fairseq (Ott et al., 2019) and its RoBERTa implementation (Liu et al., 2019). DinkyTrain's efficient pretraining recipe builds on Izsak et al. (2021). The pre-LayerNorm encoder has eight layers, hidden dimension 512, feed-forward dimension 2048, eight attention heads, and approximately 51.5 million pretraining parameters. We use the repository's preprocessed Wikipedia–BookCorpus data (Zhu et al., 2015), with GPT-2 byte-pair encoding and a 50,265-token model vocabulary. The pilot uses one fixed shard, retaining 3,756,032 contiguous sequences of length 128. Token blocks are constructed before mask views are assigned; the underlying text remains fixed across presentations.

Masking protocol. We select 15% of token positions for prediction. Every selected token is replaced by `[MASK]`; neither random-token substitution nor unchanged selected tokens are used. For finite K, the corruption depends on the mask seed, sequence identifier, and view identifier, and is fixed across presentations. Dynamic masking draws a fresh corruption at each presentation. Each sequence is presented 109 times in every condition: the single $K = 1$ mask is reused throughout, the ten $K = 10$ masks are reused approximately eleven times each, and dynamic masking draws 109 masks. Pretraining seeds are $0, 1, 2$, with the training and mask seed both set to the same value. Seeds control initialization, data order, dropout, and mask draws. Held-out validation uses 1,941 sequences.

Pretraining budget and optimization. Each run completes 99,953 optimizer updates with an effective batch size of 4,096. Adam uses a peak learning rate of 2×10^{-3}, 1,999 warmup updates followed by cosine decay, coefficients $(0.9, 0.98)$, $\epsilon = 10^{-6}$, weight decay 0.01, and no gradient clipping. Dropout and attention dropout are both 0.1. Pretraining checkpoints are selected using held-out masked-language-model validation loss.

Table 5: **BERT downstream results.** Development scores ($\times 100$, ↑), reported as mean ± SD across three pretraining seeds after averaging two fine-tuning runs per pretraining seed. MCC denotes Matthews correlation; STS-B uses Spearman correlation. MNLI-m and MNLI-mm are the two splits of the same task.

Task	Metric	$K = 1$	$K = 10$	Dynamic
MNLI-m	Accuracy	72.85 ± 0.33	76.27 ± 0.42	76.83 ± 0.18
MNLI-mm	Accuracy	74.47 ± 0.29	77.48 ± 0.13	77.65 ± 0.35
QNLI	Accuracy	82.55 ± 0.87	84.40 ± 0.61	85.45 ± 0.38
QQP	Accuracy	85.92 ± 0.10	87.02 ± 0.16	87.22 ± 0.16
SST-2	Accuracy	87.63 ± 0.60	88.08 ± 1.03	89.40 ± 0.46
STS-B	Spearman	79.76 ± 1.18	81.32 ± 1.04	82.46 ± 1.67
MRPC	MCC	46.27 ± 1.94	55.50 ± 2.63	55.55 ± 5.38
CoLA	MCC	37.23 ± 1.80	43.30 ± 2.82	47.17 ± 3.40
RTE	MCC	9.43 ± 3.29	10.42 ± 1.95	7.28 ± 4.27

Downstream optimization. We fine-tune the full model on eight GLUE tasks (Wang et al., 2018) for ten epochs and evaluate after the last epoch on the development split. The maximum input length is 512 tokens. Learning rate and batch size are 10^{-5} and 32 for MNLI, QNLI, QQP, and SST-2; 2×10^{-5} and 16 for STS-B; and 5×10^{-5} and 16 for MRPC, CoLA, and RTE.

For MRPC, CoLA, and RTE, a preliminary learning-rate sweep resolved the majority-class predictions observed with the initial fine-tuning settings. We selected 5×10^{-5} using the $K = 10$ model with pretraining seed 0 from $\{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 4 \times 10^{-4}\}$ and applied it across all masking conditions and seeds. This value is included in the released fine-tuning grid accompanying the

[3] https://github.com/princeton-nlp/DinkyTrain.

compact BERT models of Turc et al. (2019).[4] Adam uses coefficients $(0.9, 0.98)$, $\epsilon = 10^{-6}$, weight decay 0.1, and no gradient clipping. The learning rate warms up for 6% of the update budget and then decays linearly. Dropout and attention dropout are both 0.1. Classification tasks use cross-entropy; STS-B uses mean-squared-error regression.

Downstream performance. In Table 5, $K = 10$ improves the mean score over $K = 1$ on all eight tasks. Dynamic masking has the highest mean on seven of eight tasks, including both MNLI splits. The additional gain over $K = 10$ depends on the task: the MRPC means are nearly equal, while RTE remains weak and variable. The results support a benefit of mask diversity across independent pretraining runs.

H.14 Bayes curves and AMP estimates

The Bayes-optimal lines in Figs. 2, 3 and 8 show the state-evolution solution of (F.76).

Green diamonds use the alternating, Fisher-averaged AMP from App. F.5, specializing the bipartite Low-RAMP of Lesieur et al. (2017). Inference uses the known signal strength and Gaussian priors, with coordinate variances estimated by the empirical second moments in (F.9).

Initialization and convergence. The plotted runs use weak teacher-informed initialization

$$\widehat{\boldsymbol{u}}^{\,0} = \sqrt{p}\,\boldsymbol{u}^\star + \sqrt{1-p}\,\boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(0, I_d), \tag{H.3}$$

In Figs. 2, 3 and 8, AMP uses $d = 4000$, $p = 0.05$, and ten seeds per α value. Initial sample-amplitude estimates are Gaussian with variance 0.1, with damping factor $\eta = 0.75$.

Matched downstream comparison. The lower row of Fig. 3 applies the same unweighted raw-feature logistic probe to AMP and autoencoder estimates, using 100 labels and 10,000 test examples. Its green curve combines the Bayes pretraining solution with the raw-feature SNR (G.21) and the finite-label logistic-probe formulas in App. G.2.

[4] https://github.com/google-research/bert.