
Learning Curves for Noisy Heterogeneous Feature-Subsampled Ridge Ensembles

Benjamin S. Ruben¹, Cengiz Pehlevan^{2,3,4}

¹Biophysics Graduate Program

²Center for Brain Science, ³John A. Paulson School of Engineering and Applied Sciences, ⁴Kempner Institute for the Study of Natural and Artificial Intelligence,

Harvard University

Cambridge, MA 02138

benruben@g.harvard.edu, cpehlevan@seas.harvard.edu

Abstract

Feature bagging is a well-established ensembling method which aims to reduce prediction variance by combining predictions of many estimators trained on subsets or projections of features. Here, we develop a theory of feature-bagging in noisy least-squares ridge ensembles and simplify the resulting learning curves in the special case of equicorrelated data. Using analytical learning curves, we demonstrate that subsampling shifts the double-descent peak of a linear predictor. This leads us to introduce heterogeneous feature ensembling, with estimators built on varying numbers of feature dimensions, as a computationally efficient method to mitigate double-descent. Then, we compare the performance of a feature-subsampling ensemble to a single linear predictor, describing a trade-off between noise amplification due to subsampling and noise reduction due to ensembling. Our qualitative insights carry over to linear classifiers applied to image classification tasks with realistic datasets constructed using a state-of-the-art deep learning feature map.

1 Introduction

Ensembling methods are ubiquitous in machine learning practice [1]. A class of ensembling methods (known as attribute bagging [2] or the random subspace method [3]) is based on feature subsampling [2–6], where predictors are independently trained on subsets of the features, and their predictions are combined to achieve a stronger prediction. The random forest method is a popular example [3, 7].

While commonly used in practice, a theoretical understanding of ensembling via feature subsampling is not well developed. Here, we provide an analysis of this technique in the linear ridge regression setting. Using methods from statistical physics [8–12], we obtain analytical expressions for typical-case generalization error in linear ridge ensembles (proposition 1), and simplify these expressions in the special case of equicorrelated data with isotropic feature noise (proposition 2). The result provides a powerful tool to quickly probe the generalization error of ensembled regression under a rich set of conditions. In section 3, we study the behavior of a single feature-subsampling regression model. We observe that subsampling shifts the location of a predictor’s sample-wise double-descent peak [13–15]. This motivates section 4, where we study ensembles built on predictors which are heterogeneous in the number of features they access, as a method to mitigate double-descent. We demonstrate this method’s efficacy in a realistic image classification task. In section 5 we apply our theory to the trade-off between ensembling and subsampling in resource-constrained settings. We

characterize how a variety of factors influence the optimal ensembling strategy, finding a particular significance to the level of noise in the predictions made by ensemble members.

In summary, we make the following contributions:

- Using the replica trick from statistical physics [8, 11], we derive the generalization error of ensembled least-squares ridge regression in a general setting, and simplify the resulting expressions in the tractable special case where features are equicorrelated.
- We demonstrate benefits of heterogeneous ensembling as a robust and computationally efficient regularizer for mitigating double-descent with analytical theory and in a realistic image classification task.
- We describe the ensembling-subsampling trade-off in resource-constrained settings, and characterize the effect of label noise, feature noise, readout noise, regularization, sample size and task structure on the optimal ensembling strategy.

Related works: A substantial body of work has elucidated the behavior of linear predictors for a variety of feature maps [14, 16–30]. Several recent works have extended this research to characterize the behavior of ensembled regression using solvable models [24, 31–33]. Additional recent works study the performance of ridge ensembles with example-wise subsampling [34, 35] and simultaneous subsampling of features and examples [32], finding that subsampling behaves as an implicit regularization. Methods from statistical physics have long been used for machine learning theory [10–12, 26, 27, 30, 36, 37]. Relevant work in this domain include [38] which studied ensembling by data-subsampling in linear regression.

2 Learning Curves for Ensembled Ridge Regression

We consider noisy ensembled ridge regression in the setting where ensemble members are trained independently on masked versions of the available features. We derive our main analytical formula for generalization error of ensembled linear regression, as well as analytical expressions for generalization error in the special case of equicorrelated features with isotropic noise.

2.1 Problem Setup

Consider a training set $\mathcal{D} = \{\bar{\psi}^\mu, y^\mu\}_{\mu=1}^P$ of size P . The training examples $\bar{\psi}^\mu \in \mathbb{R}^M$ are drawn from a Gaussian distribution with Gaussian feature noise: $\bar{\psi}^\mu = \psi^\mu + \sigma^\mu$, where $\psi^\mu \sim \mathcal{N}(0, \Sigma_s)$ and $\sigma^\mu \sim \mathcal{N}(0, \Sigma_0)$. Data and noise are drawn i.i.d. so that $\mathbb{E}[\psi^\mu \psi^{\nu\top}] = \delta_{\mu\nu} \Sigma_s$ and $\mathbb{E}[\sigma^\mu \sigma^{\nu\top}] = \delta_{\mu\nu} \Sigma_0$. Labels are generated from a noisy teacher function $y^\mu = \frac{1}{\sqrt{M}} \mathbf{w}^{*\top} \psi^\mu + \epsilon^\mu$ where $\epsilon^\mu \sim \mathcal{N}(0, \zeta^2)$. Label noises are drawn i.i.d. so that $\mathbb{E}[\epsilon^\mu \epsilon^{\nu}] = \delta_{\mu\nu} \zeta^2$.

We seek to analyze the quality of predictions which are averaged over an ensemble of ridge regression models, each with access to a subset of the features. We consider k linear predictors with weights $\hat{\mathbf{w}}_r \in \mathbb{R}^{N_r}$, $r = 1, \dots, k$. Critically, we allow $N_r \neq N_{r'}$ for $r \neq r'$, which allows us to introduce *structural* heterogeneity into the ensemble of predictors. A forward pass of the model is given as:

$$f(\psi) = \frac{1}{k} \sum_{r=1}^k f_r(\psi), \quad f_r(\psi) = \frac{1}{\sqrt{N_r}} \hat{\mathbf{w}}_r^\top \mathbf{A}_r(\psi + \sigma) + \xi_r. \quad (1)$$

The model’s prediction $f(\psi)$ is an average over k linear predictors. The “measurement matrices” $\mathbf{A}_r \in \mathbb{R}^{N_r \times M}$ act as linear masks restricting the information about the features available to each member of the ensemble. Subsampling may be implemented by choosing the rows of each \mathbf{A}_r to coincide with the rows of the identity matrix – the row indices corresponding to indices of the sampled features. The feature noise $\sigma \sim \mathcal{N}(0, \Sigma_0)$ and the readout noises $\xi_r \sim \mathcal{N}(0, \eta_r^2)$, are drawn independently at the execution of each forward pass of the model. Note that while the feature noise is shared across the ensemble, readout noise is drawn independently for each readout: $\mathbb{E}[\xi_r \xi_{r'}] = \delta_{rr'} \eta_r^2$.

The weight vectors are trained separately in order to minimize an ordinary least-squares loss function with ridge regularization:

$$\hat{\mathbf{w}}_r = \arg \min_{\mathbf{w}_r \in \mathbb{R}^{N_r}} \left[\sum_{\mu=1}^P \left(\frac{1}{\sqrt{N_r}} \mathbf{w}_r^\top \mathbf{A}_r \bar{\boldsymbol{\psi}}^\mu + \xi_r^\mu - y^\mu \right)^2 + \lambda_r |\mathbf{w}_r|^2 \right] \quad (2)$$

Here $\{\xi_r^\mu\}$ represents the readout noise which is present during training, and independently drawn: $\xi_r^\mu \sim \mathcal{N}(0, \eta_r^2)$, $\mathbb{E}[\xi_r^\mu \xi_r^\nu] = \eta_r^2 \delta_{\mu\nu}$. As a measure of model performance, we consider the generalization error, given by the mean-squared-error (MSE) on ensemble-averaged prediction:

$$E_g(\mathcal{D}) = \mathbb{E}_{\boldsymbol{\psi}, \sigma, \{\xi_r\}} \left[\left(f(\boldsymbol{\psi}) - \frac{1}{\sqrt{M}} \mathbf{w}^{*\top} \boldsymbol{\psi} \right)^2 \right] \quad (3)$$

Here, the expectation is over the data distribution and noise: $\boldsymbol{\psi} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$, $\sigma \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$, $\xi_r \sim \mathcal{N}(0, \eta_r^2)$. The generalization error depends on the particular realization of the dataset \mathcal{D} through the learned weights $\{\hat{\mathbf{w}}^*\}$. We may decompose the generalization error as follows:

$$E_g(\mathcal{D}) = \frac{1}{k^2} \sum_{r, r'=1}^k E_{rr'}(\mathcal{D}) \quad (4)$$

$$E_{rr'}(\mathcal{D}) \equiv \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r^\top \hat{\mathbf{w}}_r - \mathbf{w}^* \right)^\top \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{r'r'}}} \mathbf{A}_{r'}^\top \hat{\mathbf{w}}_{r'} - \mathbf{w}^* \right) + \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \hat{\mathbf{w}}_r^\top \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_{r'}^\top \hat{\mathbf{w}}_{r'} + M \delta_{rr'} \eta_r^2 \right] \quad (5)$$

Computing the generalization error of the model is then a matter of calculating $E_{rr'}$ in the cases where $r = r'$ and $r \neq r'$. In the asymptotic limit we consider, we expect that the generalization error concentrates over randomly drawn datasets \mathcal{D} .

2.2 Main Result

We calculate the generalization error using the replica trick from statistical physics, and present the calculation in Appendix F. The result of our calculation is stated in proposition 1.

Proposition 1. *Consider the ensembled ridge regression problem described in Section 2.1. Consider the asymptotic limit where $M, P, \{N_r\} \rightarrow \infty$ while the ratios $\alpha = \frac{P}{M}$ and $\nu_{rr} = \frac{N_r}{M}$, $r = 1, \dots, k$ remain fixed. Define the following quantities:*

$$\tilde{\boldsymbol{\Sigma}}_{rr'} \equiv \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \mathbf{A}_r [\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_0] \mathbf{A}_{r'}^\top \quad (6)$$

$$\mathbf{G}_r \equiv \mathbf{I}_{N_r} + \hat{q}_r \tilde{\boldsymbol{\Sigma}}_{rr} \quad (7)$$

$$\gamma_{rr'} \equiv \frac{\alpha}{M(\lambda_r + q_r)(\lambda_{r'} + q_{r'})} \text{tr} \left[\mathbf{G}_r^{-1} \tilde{\boldsymbol{\Sigma}}_{rr'} \mathbf{G}_{r'}^{-1} \tilde{\boldsymbol{\Sigma}}_{r'r} \right] \quad (8)$$

Then the terms of the average generalization error (eq. 5) may be written as:

$$\begin{aligned} \langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}} &= \frac{\gamma_{rr'} \zeta^2 + \delta_{rr'} \eta_r^2}{1 - \gamma_{rr'}} + \frac{1}{1 - \gamma_{rr'}} \left(\frac{1}{M} \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \mathbf{w}^* \right) \\ &\quad - \frac{1}{M(1 - \gamma_{rr'})} \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \left[\frac{1}{\nu_{rr}} \hat{q}_r \mathbf{A}_r^\top \mathbf{G}_r^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{q}_{r'} \mathbf{A}_{r'}^\top \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \right] \boldsymbol{\Sigma}_s \mathbf{w}^* \quad (9) \\ &\quad + \frac{\hat{q}_r \hat{q}_{r'}}{M(1 - \gamma_{rr'})} \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \tilde{\boldsymbol{\Sigma}}_{rr'} \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \boldsymbol{\Sigma}_s \mathbf{w}^* \end{aligned}$$

where the pairs of order parameters $\{q_r, \hat{q}_r\}$ for $r = 1, \dots, K$, satisfy the following self-consistent saddle-point equations

$$\hat{q}_r = \frac{\alpha}{\lambda_r + q_r}, \quad q_r = \frac{1}{M} \text{tr} \left[\mathbf{G}_r^{-1} \tilde{\boldsymbol{\Sigma}}_{rr} \right]. \quad (10)$$

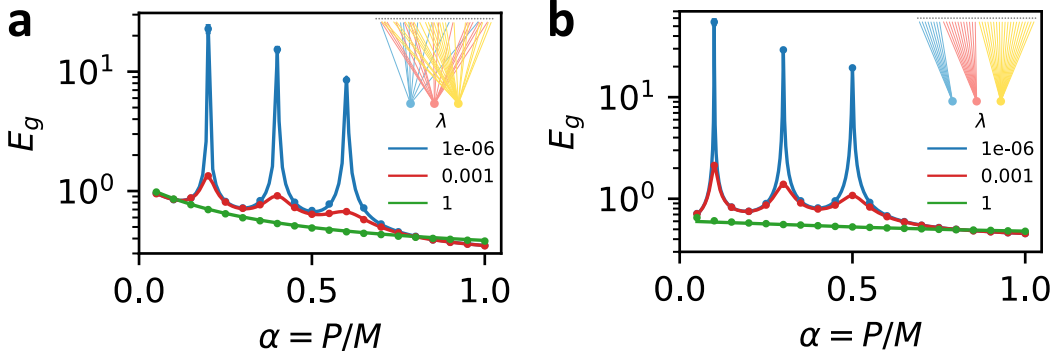


Figure 1: Comparison of numerical and theoretical learning curves for ensemble linear regression. Circles represent numerical results averaged over 100 trials; lines indicate theoretical predictions. Error bars represent the standard error of the mean but are often smaller than the markers. (a) Testing of proposition 1 with $M = 2000$, $[\Sigma_s]_{ij} = .8^{|i-j|}$, $[\Sigma_0]_{ij} = \frac{1}{10}(0.3)^{|i-j|}$, $\zeta = 0.1$, and all $\eta_r = 0.2$ and $\lambda_r = \lambda$ (see legend). $k = 3$ linear predictors access fixed, randomly selected (with replacement) subsets of the features with fractional sizes $\nu_{rr} = 0.2, 0.4, 0.6$. Fixed ground-truth weights \mathbf{w}^* are drawn from an isotropic Gaussian distribution. (b) Testing of proposition 2 with $M = 5000$, $s = 1$, $c = 0.6$, $\omega^2 = 0.1$, $\zeta = 0.1$, all $\eta_r = 0.1$, and all $\lambda_r = \lambda$ (see legend). Ground truth weights sampled as in eq. 11 with $\rho = 0.3$. Feature subsets accessed by each readout are mutually exclusive (inset) with fractional sizes $\nu_{rr} = 0.1, 0.3, 0.5$.

Proof. We calculate the terms in the generalization error using the replica trick, a standard but non-rigorous method from the statistical physics of disordered systems. The full derivation may be found in the Appendix F. When the matrices $\mathbf{A}_r (\Sigma_s + \Sigma_0) \mathbf{A}_r^\top$, $r = 1, \dots, k$ have bounded spectra, this result may be obtained by extending the results of [31] to include readout noise, as shown in Appendix G. \square

We make the following remarks:

Remark 1. Implicit in this theorem is the assumption that the relevant matrix contractions and traces which appear in the generalization error (eq. 9) and the surrounding definitions tend to a well-defined limit which remains $\mathcal{O}(1)$ as $M \rightarrow \infty$.

Remark 2. This result applies for any (well-behaved) linear masks $\{\mathbf{A}_r\}$. We will focus on the case where each \mathbf{A}_r implements subsampling of an extensive fraction ν_{rr} of the features.

Remark 3. When $k = 1$, our result reduces to the generalization error of a single ridge regression model, as studied in refs. [36, 39].

Remark 4. We include “readout noise” which independently corrupts the predictions of each ensemble member. This models sources of variation between ensemble members not otherwise accounted for. For example, ensembles of deep networks will vary due to random initialization of parameters [24, 31, 36]. Readout noise is more directly present in physical neural networks, such as an analog neural networks [40] or biological neural circuits[41] due to their inherent stochasticity.

In Figure 1a, we confirm the result of the general calculation by comparing with numerical experiments using a synthetic dataset with $M = 2000$ highly structured features (see caption for details). $k = 3$ readouts see random, fixed subsets of features. Theory curves are calculated by solving the fixed-point equations 10 numerically for the chosen Σ_s , Σ_0 and $\{\mathbf{A}_r\}_{r=1}^k$ then evaluating eq. 9.

2.3 Equicorrelated Data

Our general result allows the freedom to tune many important parameters of the learning problem: the correlation structure of the dataset, the number of ensemble members, the scales of noise, etc. However, the derived expressions are rather opaque. In order to better understand the phenomena captured by these expressions, we examine the following special case:

Proposition 2. *In the setting of section 2.1 and proposition 1, consider the following special case:*

$$\mathbf{w}^* = \sqrt{1 - \rho^2} \mathbb{P}_\perp \mathbf{w}_0^* + \rho \mathbf{1}_M \quad (11)$$

$$\mathbf{w}_0^* \sim \mathcal{N}(0, \mathbf{I}_M) \quad (12)$$

$$\boldsymbol{\Sigma}_s = s [(1 - c) \mathbf{I}_M + c \mathbf{1}_M \mathbf{1}_M^\top] \quad (13)$$

$$\boldsymbol{\Sigma}_0 = \omega^2 \mathbf{I}_M \quad (14)$$

with $c \in [0, 1]$, $\rho \in [-1, 1]$. Label and readout noises $\zeta, \eta_r \geq 0$ are permitted. Here $\mathbb{P}_\perp = \mathbf{I}_M - \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top$ is a projection matrix which removes the component of \mathbf{w}_0^* which is parallel to $\mathbf{1}_M$. The matrices $\{\mathbf{A}_r\}_{r=1}^k$ have rows consisting of distinct one-hot vectors so that each of the k readouts has access to a subset of $N_r = \nu_{rr} M$ features. For $r \neq r'$, denote by $n_{rr'}$ the number of neurons sampled by both \mathbf{A}_r and $\mathbf{A}_{r'}$ and let $\nu_{rr'} \equiv n_{rr'} / M$ remain fixed as $M \rightarrow \infty$.

Define the following quantities:

$$a \equiv s(1 - c) + \omega^2 \quad S_r \equiv \frac{\hat{q}_r}{\nu_{rr} + a \hat{q}_r}, \quad \gamma_{rr'} \equiv \frac{a^2 \nu_{rr'} S_r S_{r'}}{\alpha} \quad (15)$$

The terms of the decomposed generalization error may then be written:

$$\langle E_{rr'} \rangle_{\mathcal{D}, \mathbf{w}_0^*} = \frac{1}{1 - \gamma_{rr'}} \left((1 - \rho^2) I_{rr'}^0 + \rho^2 I_{rr'}^1 \right) + \frac{\gamma_{rr'} \zeta^2 + \delta_{rr'} \eta_r^2}{1 - \gamma_{rr'}} \quad (16)$$

where we have defined

$$I_{rr'}^0 \equiv s(1 - c) (1 - s(1 - c) \nu_{rr} S_r - s(1 - c) \nu_{r'r'} S_{r'} + a s(1 - c) \nu_{rr'} S_r S_{r'}) \quad (17)$$

$$I_{rr'}^1 \equiv \begin{cases} \frac{s(1-c)(\nu_{rr'} - \nu_{rr} \nu_{r'r'}) + \omega^2 \nu_{rr'}}{\nu_{rr} \nu_{r'r'}} & \text{if } 0 < c \leq 1 \\ I_{rr'}^0 & \text{if } c = 0 \end{cases} \quad (18)$$

and where $\{q_r, \hat{q}_r\}$ may be obtained analytically as the solution (with $q_r > 0$) to:

$$q_r = \frac{a \nu_{rr}}{\nu_{rr} + a \hat{q}_r}, \quad \hat{q}_r = \frac{\alpha}{\lambda_r + q_r} \quad (19)$$

In the ‘‘ridgeless’’ limit where all $\lambda_r \rightarrow 0$, we may make the following simplifications:

$$S_r \rightarrow \frac{2\alpha}{a(\alpha + \nu_{rr} + |\alpha - \nu_{rr}|)}, \quad \gamma_{rr'} \rightarrow \frac{4\alpha \nu_{rr'}}{(\alpha + \nu_{rr} + |\alpha - \nu_{rr}|)(\alpha + \nu_{r'r'} + |\alpha - \nu_{r'r'}|)} \quad (20)$$

Proof. Simplifying the fixed-point equations and generalization error formulas in this special case is an exercise in linear algebra. The main tools used are the Sherman-Morrison formula [42] and the fact that the data distribution is isotropic in the features so that the form of $\tilde{\boldsymbol{\Sigma}}_{rr}$ and $\tilde{\boldsymbol{\Sigma}}_{r'r'}$ depend only on the subsampling and overlap fractions $\nu_{rr}, \nu_{r'r'}, \nu_{rr'}$. To aid in computing the necessary matrix contractions we developed a custom Mathematica package which handles block matrices of symbolic dimension, with blocks containing matrices of the form $\mathbf{M} = c_1 \mathbf{I} + c_2 \mathbf{1}\mathbf{1}^\top$. This package and the Mathematica notebook used to derive these results are available online (see Appendix B) \square

In this tractable special case, $c \in [0, 1]$ is a parameter which tunes the strength of correlations between features of the data. When $c = 0$, the features are independent, and when $c = 1$ the features are always equivalent. s sets the overall scale of the features and the ‘‘Data-Task alignment’’ ρ tunes the alignment of the ground truth weights with the special direction in the covariance matrix (analogous to ‘‘task-model’’ alignment [14, 27]). A table of parameters is provided in Appendix A. In Figure 1b, we test these results by comparing the theoretical expressions for generalization error with the results of numerical experiments, finding perfect agreement.

With an analytical formula for the generalization error, we can compute the optimal regularization parameters λ_r which minimize the generalization error. These may, in general, depend on both r and the sample size α . Rather than minimizing the error of the ensemble, we may minimize the generalization error of predictions made by the ensemble members independently. We find that this ‘‘locally optimal’’ regularization, denoted λ^* , is independent of α , generalizing results from [14, 43] to correlated data distributions (see Appendix H.3).

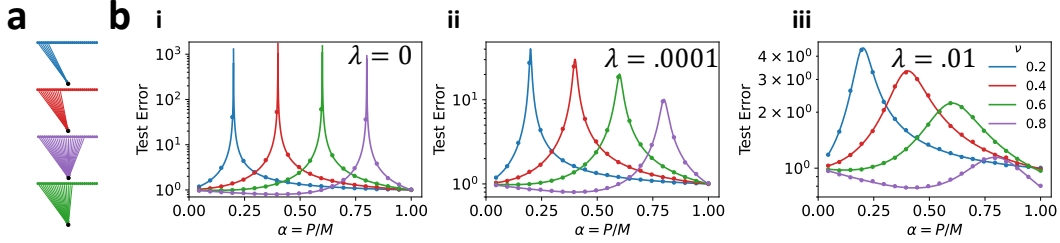


Figure 2: Subsampling alters the location of the double-descent peak of a linear predictor. (a) Illustrations of subsampled linear predictors with varying subsampling fraction ν . (b) Comparison between experiment and theory for subsampling linear regression on equicorrelated datasets. We choose task parameters as in proposition 2 with $c = \omega = \zeta = \eta = 0$, $s = 1$, and (i) $\lambda = 0$, (ii) $\lambda = 10^{-4}$, (iii) $\lambda = 10^{-2}$. All learning curves are for a single linear predictor $k = 1$ with subsampling fraction ν shown in legend. Circles show results of numerical experiment. Lines are analytical prediction.

3 Subsampling shifts the double-descent peak of a linear predictor

Consider a single linear regressor ($k = 1$) which connects to a subset of $N = \nu M$ features in the equicorrelated data setting of proposition 2. Also setting $c = 0$, $s = 1$, and $\eta_r = \omega = 0$ and taking the limit $\lambda \rightarrow 0$ the generalization error reads:

$$\langle E_g \rangle_{\mathcal{D}, \mathbf{w}^*} = \left\{ \begin{array}{ll} \frac{\nu}{\nu - \alpha} \left[(1 - \nu) + \frac{1}{\nu} (\alpha - \nu)^2 \right] + \frac{\alpha}{\nu - \alpha} \zeta^2, & \text{if } \alpha < \nu \\ \frac{\alpha}{\alpha - \nu} [1 - \nu] + \frac{\nu}{\alpha - \nu} \zeta^2, & \text{if } \alpha > \nu \end{array} \right\} \quad (21)$$

We thus see that double descent can arise from two possible sources of variance: explicit label noise (if $\zeta > 0$) or implicit label noise induced by feature subsampling ($\nu < 1$). As $E_g \sim (\alpha - \nu)^{-1}$, generalization error diverges when sample size is equal to the number of sampled features. Intuitively, this occurs because subsampling changes the number of parameters of the regression model, and thus its interpolation threshold. To demonstrate this, we plot the learning curves for subsampled linear regression on equicorrelated data in Figure 2. At small finite ridge the test error no longer diverges when $\alpha = \nu$, but still displays a distinctive peak.

4 Heterogeneous connectivity mitigates double-descent

Double-descent – over-fitting to noise in the training set near a model’s interpolation threshold – poses a serious risk in practical machine-learning applications [22]. Cross-validating the regularization strength against the training set is the canonical approach to avoiding double-descent [17, 43], but in practice requires a computationally expensive parameter sweep and prior knowledge of the task. In situations where computational resources are limited or hyperparameters are fixed prior to specification of the task, it is natural to seek an alternative solution. Considering again the plots in Figure 2(b), we observe that at any value of α , the double-descent peak can be avoided with an acceptable choice of the subsampling fraction ν . This suggests another strategy to mitigate double descent: heterogeneous ensembling. Ensembling over predictors with a heterogeneous distribution of interpolation thresholds, we may expect that when one predictor fails due to over-fitting, the other members of the ensemble compensate with accurate predictions.

In Figure 3, we show that heterogeneous ensembling can guard against double-descent. We define two ensembling strategies: in homogeneous ensembling, each of k readouts connects a fraction $\nu_{rr} = 1/k$ features. In heterogeneous ensembling, the number of features connected by each of the k readouts are drawn from a Gamma distribution $\Gamma_{k, \sigma}(\nu)$ with mean $1/k$ and standard deviation σ (see Fig. 3b) then re-scaled to sum to 1 (see Appendix C for details). All feature subsets are mutually exclusive ($\nu_{rr'} = 0$ for $r \neq r'$). Homogeneous and heterogeneous ensembling are illustrated for $k = 10$ in Figs. 3 a.i and 3 a.ii respectively. We test this hypothesis using eq. 16 in 3c. At small regularization

($\lambda = .001$), we find that heterogeneity of the distribution of subsampling fractions ($\sigma > 0$) lowers the double-descent peak of an ensemble of linear predictors, while at larger regularization ($\lambda = 0.1$), there is little difference between homogeneous and heterogeneous learning curves. The asymptotic ($\alpha \rightarrow \infty$) error is unaffected by the presence of heterogeneity in the degrees of connectivity, which can be seen as the coincidence of the triangular markers in Fig. 3c, as well as from the $\alpha \rightarrow \infty$ limit of eq. 16 (see Appendix H.5). Fig. 3c also shows the learning curve of a single linear predictor with no feature subsampling and optimal regularization. We see that the feature-subsampling ensemble appreciably outperforms the fully-connected model when $c = 0.8$ and $\eta = 0.5$, suggesting the important roles of data correlations and readout noise in determining the optimal readout strategy. These roles are further explored in section 5 and fig 4.

We also test the effect of heterogeneous ensembling in the a realistic classification task. Specifically, we train ensembles of linear classifiers to predict the labels of imagenet [44] images corresponding to 10 different dog breeds (the ‘‘Imagewoof’’ task [45]) from their top-hidden-layer representations in a pre-trained ResNext deep network [46] (see Appendix E for details). We characterize the statistics of the resulting $M = 2048$ -dimensional feature set in Fig. S1. This ‘‘ResNext-Imagewoof’’ classification task has multiple features which make it amenable to learning with a feature-subsampling ensemble. First, the ResNext features have a high degree of redundancy [47], allowing classification to be performed accurately using only a fraction of the available features (see Fig. 3d and S1c). Second, when classifications of multiple predictors are combined by a majority vote, there is a natural upper bound on the influence of a single erring ensemble member (unlike in regression where predictions can diverge). Calculating learning curves for the imagewoof classification task using homogeneous ensembles, we see sharp double-descent peaks in an ensemble of size k when $P = M/k$ (Fig. 3e.i). Using a heterogeneous ensemble mitigates this catastrophic over-fitting, leading to monotonically decreasing error without regularization (Fig. 3e.ii). A single linear predictor with a tuned regularization of $\lambda = 0.1$ performs only marginally better than the heterogeneous feature-subsampling ensemble with $k = 16$ or $k = 32$. This suggests heterogeneous ensembling can be an effective alternative to regularization in real-world classification tasks using pre-trained deep learning feature maps.

Note that training a feature-subsampling ensemble also benefits from improved computational complexity. Training an estimator of dimension N_r involves, in the worst case, inverting an $N_r \times N_r$ matrix, which requires $\mathcal{O}(N_r^3)$ operations. Setting $N_r = M/k$, we see that the number of operations required to train an ensemble of k predictors scales as $\mathcal{O}(k^{-2})$.

5 Correlations, Noise, and Task Structure Dictate the Ensembling-Subsampling Trade-off

In resource-constrained settings, one must decide between training a single large predictor or an ensemble of smaller predictors. When the number of weights is constrained, ensembling may benefit generalization by averaging over multiple predictions, but at the expense of each prediction incorporating fewer features. Intuitively, the presence of correlations between features limits the penalty incurred by subsampling, as measurements from a subset of features will also confer information about the unsampled features. The equicorrelated data model of proposition 2 permits a solvable toy model for these competing effects. We consider the special case of ensembling over k readouts, each connecting the same fraction $\nu_{rr} = \nu = 1/k$ of all features. For simplicity, we set $\nu_{rr'} = 0$ for $r \neq r'$. We asses the learning curves of this toy model in both the ridgeless limit $\lambda \rightarrow 0$ where double-descent has a large effect on test error, and at ‘locally optimal’ regularization $\lambda = \lambda^*$ for which double-descent is eliminated. In these special cases, one can write the generalization error in the following forms (see Appendix H.4 for derivation):

$$E_g(k, s, c, \eta, \omega, \zeta, \rho, \alpha, \lambda = 0) = s(1 - c)\mathcal{E}(k, \rho, \alpha, H, W, Z) \quad (22)$$

$$E_g(k, s, c, \eta, \omega, \zeta, \rho, \alpha, \lambda = \lambda^*) = s(1 - c)\mathcal{E}^*(k, \rho, \alpha, H, W, Z) \quad (23)$$

where we have defined the effective noise-to-signal ratios:

$$H \equiv \frac{\eta^2}{s(1 - c)}, \quad W = \frac{\omega^2}{s(1 - c)}, \quad Z = \frac{\zeta^2}{s(1 - c)} \quad (24)$$

Therefore, given fixed parameters s, c, ρ, α , the value k^* which minimizes error depends on the noise scales, s , and c only through the ratios H, W and Z :

$$k_{\lambda=0}^*(H, W, Z, \rho, \alpha) \equiv \arg \min_{k \in \mathbb{N}} E_g(k) = \arg \min_{k \in \mathbb{N}} \mathcal{E}(k, \rho, \alpha, H, W, Z) \quad (25)$$

$$k_{\lambda=\lambda^*}^*(H, W, Z, \rho, \alpha) \equiv \arg \min_{k \in \mathbb{N}} E_g(k) = \arg \min_{k \in \mathbb{N}} \mathcal{E}^*(k, \rho, \alpha, H, W, Z) \quad (26)$$

In Fig. 4a, we plot these reduced errors curves $\mathcal{E}, \mathcal{E}^*$ as a function of α for varying ensemble sizes k and reduced readout noise scales H . At zero regularization learning curves diverge at their interpolation threshold. At locally optimal regularization $\lambda = \lambda^*$, learning curves decrease monotonically with sample size. Increasing readout noise H raises generalization error more sharply for smaller k . In Fig. 4b we plot the optimal k^* in various two-dimensional slices of parameter space in which ρ is fixed and $W = Z = 0$ while α and H vary. The resulting phase diagrams may be divided into three regions. In the signal-dominated phase a single fully-connected readout is optimal ($k^* = 1$). In an intermediate phase, $1 < k^* < \infty$ minimizes error. And in a noise-dominated phase $k^* = \infty$. At zero regularization, we have determined an analytical expression for the boundary between the intermediate and noise-dominated phases (see Appendix H.4.1 and dotted lines in Figs 4.b,c,d). The signal-dominated, intermediate, and noise-dominated phases persist when $\lambda = \lambda^*$, removing the effects of double descent. In all panels, an increase in H causes an increase in k^* . This can occur because of a decrease in the signal-to-readout noise ratio s/η^2 , or through an increase in the correlation strength c . An increase in ρ also leads to an increase in k^* , indicating that ensembling is more effective for easier tasks. Figs 4c,d show analogous phase diagrams where W or Z are varied. Signal-dominated, intermediate, and noise-dominated regimes are visible in the resulting phase diagrams at zero regularization. However, when optimal regularization is used, $k^* = 1$ is always optimal. The presence of regions where $k^* > 1$ can thus be attributed to double-descent at sub-optimal regularization or to the presence of readout noise which is independent across predictors. We chart the parameter-space of the reduced errors and optimal ensemble size k^* extensively in Appendix I. We plot learning curves for the ‘‘ResNext-Imagewoof’’ ensembled linear classification task with varying strength of readout noise in Fig. 4e, and phase diagrams of optimal ensemble size k in Fig. 4f, finding similar behavior to the toy model. See Figs. S3, S4, S5 and Appendix E.4.3 for further discussion.

6 Conclusion

In this paper, we provided a theory of feature-subsampled linear ridge regression. We identified the special case in which features of the data are ‘‘equicorrelated’’ as a minimal toy model to explore the combined effects of subsampling, ensembling, and different types of noise on generalization error. The resulting learning curves displayed two potentially useful phenomena.

First, we demonstrated that heterogeneous ensembling can mitigate over-fitting, reducing or eliminating the double-descent peak of an under-regularized model. In most machine learning applications, the size of the dataset is known at the outset and suitable regularization may be determined to mitigate double descent, either by selecting a highly over-parameterized model [22] or by cross-validation techniques (see for example [17]). However, in contexts where a single network architecture is designed for an unknown task or a variety of tasks with varying structure and noise levels, heterogeneous ensembling may be used to smooth out the perils of double-descent.

Next, we described a trade-off between noise reduction due to ensembling and noise amplification due to subsampling in a resource-constrained setting where the total number of weights is fixed. Our analysis suggests that ensembling is particularly useful in neural networks with an inherent noise. Physical neural networks, such as analog neural networks[40] and biological neural circuits [41] present such a resource-constrained environments where intrinsic noise is a significant issue.

Much work remains to achieve a full understanding of the interactions between data correlations, readout noise, and ensembling. In this work, we have given a thorough treatment of the convenient special case where features are equicorrelated. Future work should analyze subsampling and ensembling for codes with realistic correlation structure, such as the power-law spectra (see Fig. S1) [27, 30, 48, 49] and sparse activation patterns [50].

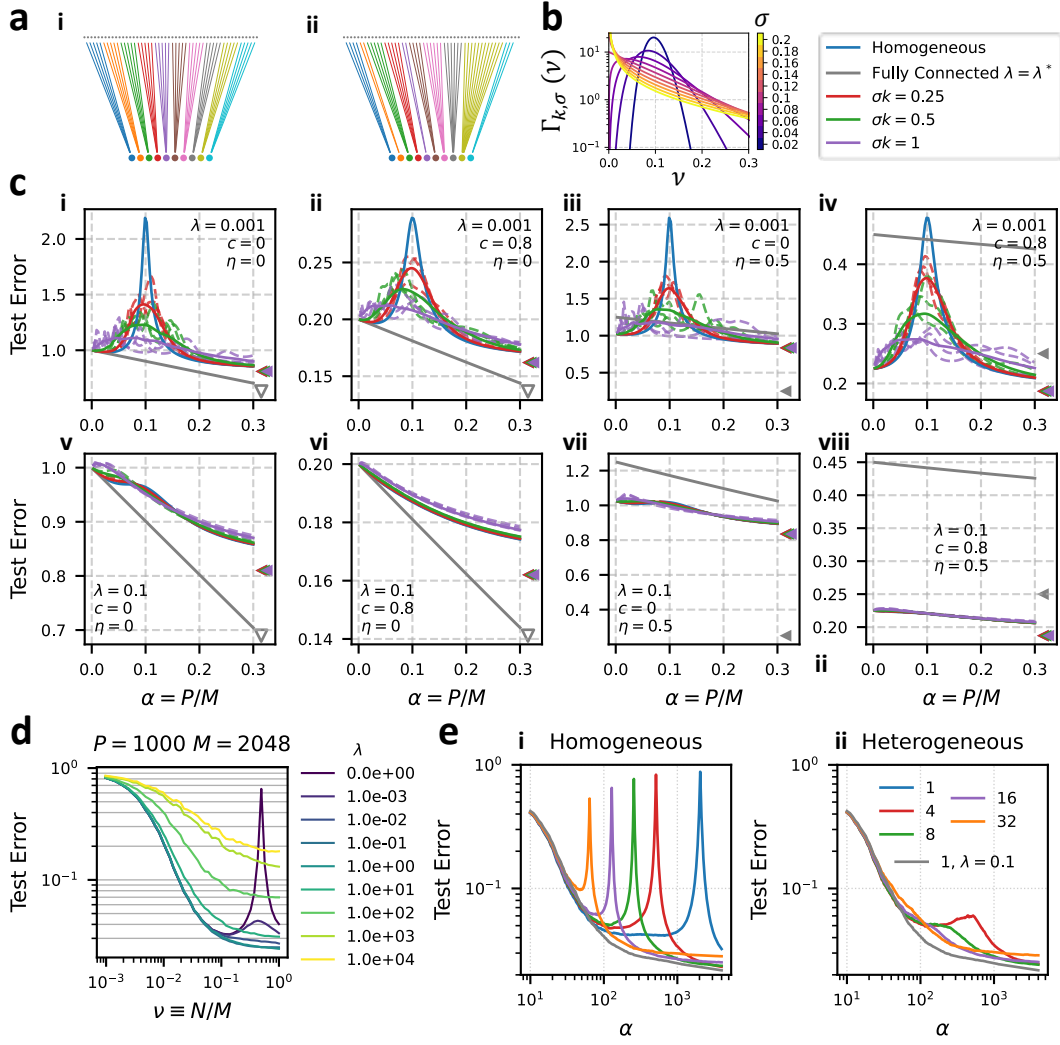


Figure 3: Heterogeneous ensembling mitigates double-descent. (a) We compare (i) homogeneous ensembling, in which k readouts connect to the same fraction $\nu = 1/k$ of features, and (ii) heterogeneous ensembling (b) In heterogeneous ensembling subsampling fractions are drawn i.i.d. from $\Gamma_{k,\sigma}(\nu)$, shown here for $k = 10$, then re-scaled to sum to 1. (c) Generalization Error Curves for Homogeneous and Heterogeneous ensembling with $k = 10$, $\zeta = 0$, $\rho = 0.3$ and indicated values of λ , c , and η . Blue: homogeneous subsampling. Red, green, and purple show heterogeneous subsampling with $\sigma = 0.25/k, 0.5/k, 1/k$ respectively. Dashed lines show learning curves for 3 particular realizations of $\{\nu_{11}, \dots, \nu_{kk}\}$. Solid curves show the average over 100 realizations. Gray shows the learning curve for a single linear readout with $\nu = 1$ and optimal regularization (eq. 193). Triangular marks show the asymptotic generalization error ($\alpha \rightarrow \infty$), with downward-pointing gray triangles indicating an asymptotic error of zero. (d,e) Generalization error of linear classifiers applied to the imagewoof dataset with ResNext features averaged over 100 trials. (d) $P = 100$, $k = 1$ varying subsampling fraction ν and regularization λ (legend). (e) Generalization error of (i) homogeneous and (ii) heterogeneous (with $\sigma = 0.75/k$) ensembles of classifiers. Legend indicates k values. $\lambda = 0$ except for gray curves, where $\lambda = 0.1$

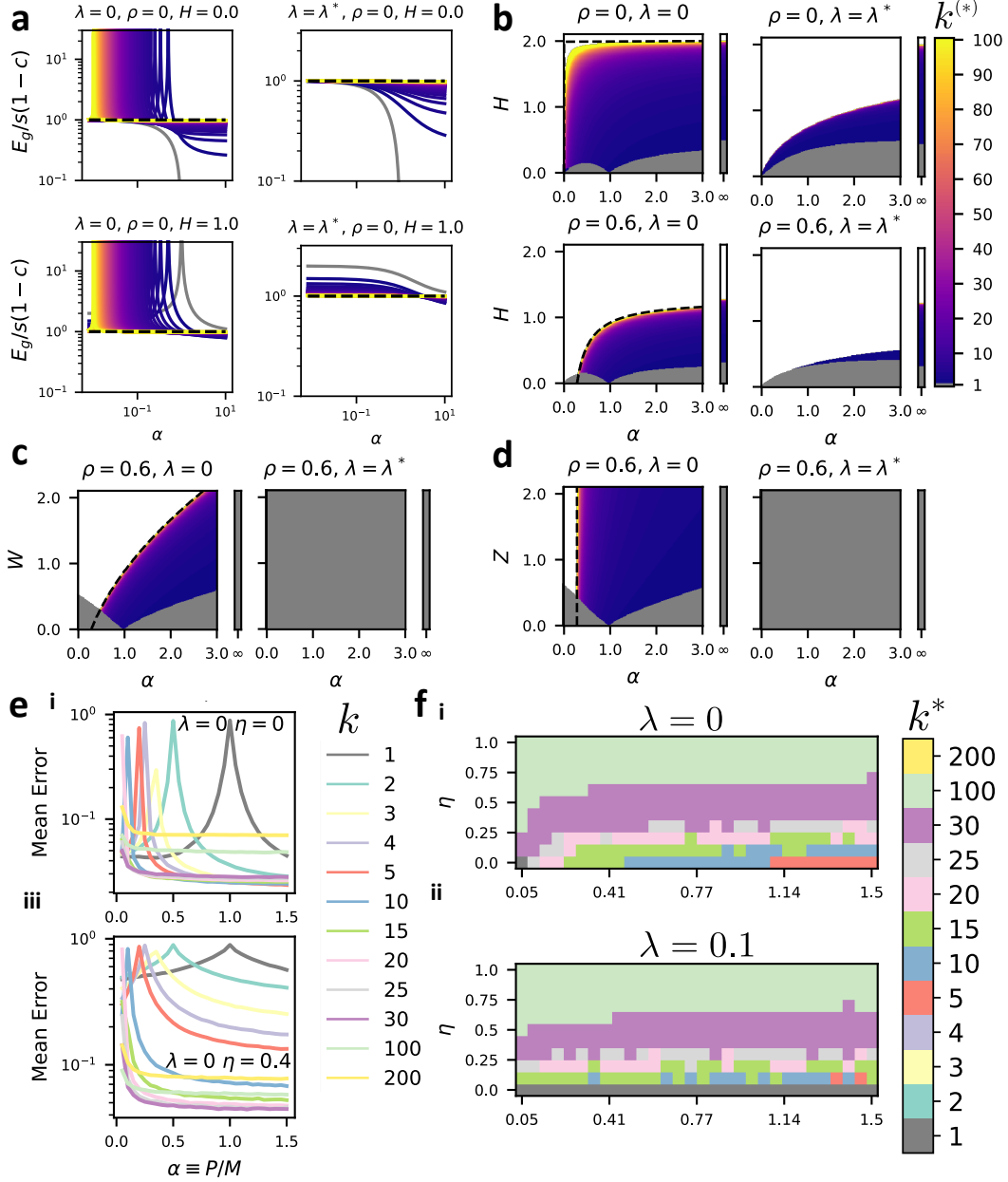


Figure 4: Task parameters dictate the ensembling-subsampling trade-off: (a-d) In the setting of proposition 2 in the special case where all $\nu_{rr'} = \frac{1}{k} \delta_{rr'}$, so that feature subsets are mutually exclusive and the total number of weights is conserved. (a) We plot the reduced generalization errors \mathcal{E} (for $\lambda = 0$, using eq. 22) and \mathcal{E}^* (for $\lambda = \lambda^*$ using eq. 23) of linear ridge ensembles of varying size k with $\rho = 0$ and $H = 0, 1$ (values indicated above plots). Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar. (b) We plot optimal k value k^* (eqs. 25, 26) in the parameter space of sample size α and reduced readout noise scale H setting $W = Z = 0$. Grey indicates $k^* = 1$ and white indicates $k^* = \infty$, with intermediate values given by the colorbar. Appended vertical bars show $\alpha \rightarrow \infty$. Dotted black lines show the analytical boundary between the intermediate and noise-dominated phases given by eq. 214. (c) optimal readout k^* phase diagrams as in (b) but showing W -dependence with $H = Z = 0$. (d) optimal readout k^* phase diagrams as in (b) but showing Z -dependence with $H = W = 0$. (e) Learning curves for feature-subsampling ensembles of linear classifiers combined using a majority vote rule on the imagewoof classification task (see Appendix E). As in (a-d) we set $\nu_{rr'} = \frac{1}{k} \delta_{rr'}$. Error is calculated as the probability of incorrectly classifying a test example. λ and η values are indicated in each panel. (f) Numerical phase diagrams showing the value of k which minimizes test error in the parameter space of sample size P and readout noise scale η , with regularization (i) $\lambda = 0$ (pseudoinverse rule) (ii) $\lambda = 0.1$.

7 Acknowledgements

CP is supported by NSF Award DMS-2134157, NSF CAREER Award IIS-2239780, and a Sloan Research Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. BSR was also supported by the National Institutes of Health Molecular Biophysics Training Grant NIH/ NIGMS T32 GM008313. We thank Jacob Zavatone-Veth and Blake Bordelon for thoughtful discussion and comments on this manuscript.

References

- [1] Gautam Kunapuli. *Ensemble Methods for Machine Learning*. Simon and Schuster, 2023.
- [2] Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291–1302, 2003.
- [3] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.
- [4] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.
- [5] Gilles Louppe and Pierre Geurts. Ensembles on random patches. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I 23*, pages 346–361. Springer, 2012.
- [6] Marina Skurichina and Robert P. W. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, June 2002.
- [7] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- [8] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. World Scientific Publishing Company, 1987.
- [9] M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, November 1986.
- [10] Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.
- [11] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.
- [12] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11:501–528, 2020.
- [13] Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.
- [14] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.
- [15] Fatih Furkan Yilmaz and Reinhard Heckel. Regularization-wise double descent: Why it occurs and how to eliminate it. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 426–431. IEEE, 2022.

- [16] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.
- [17] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [18] Jason W. Rocks and Pankaj Mehta. Bias-variance decomposition of overparameterized regression with random linear features. *Physical Review E*, 106:025304, Aug 2022.
- [19] Jason W. Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4:013201, Mar 2022.
- [20] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 2019.
- [21] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- [22] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [23] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.
- [24] Stéphane D’Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2280–2290. PMLR, 13–18 Jul 2020.
- [25] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 11022–11032, 2020.
- [26] Jacob A. Zavatone-Veth, William L. Tong, and Cengiz Pehlevan. Contrasting random and learned features in deep bayesian linear regression, 2022.
- [27] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.
- [28] James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The Eigen-learning framework: A conservation law perspective on kernel regression and wide neural networks. *arXiv*, 2022.
- [29] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.
- [30] Jacob A. Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured gaussian feature models, 2023.
- [31] Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14283–14314. PMLR, 17–23 Jul 2022.

- [32] Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3525–3535. PMLR, 26–28 Aug 2020.
- [33] Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning, 2023.
- [34] Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation, 2023.
- [35] Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. Bagging in overparameterized learning: Risk characterization and risk monotonicity, 2022.
- [36] Alexander Atanasov, Blake Bordelon, Sabarish Sainathan, and Cengiz Pehlevan. The onset of variance-limited behavior for networks in the lazy and rich regimes. In *The Eleventh International Conference on Learning Representations*, 2023.
- [37] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12600–12612. Curran Associates, Inc., 2021.
- [38] Peter Sollich and Anders Krogh. Learning with ensembles: How overfitting can be useful. *Advances in neural information processing systems*, 8, 1995.
- [39] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Learning curves of generic features maps for realistic datasets with a teacher-student model. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [40] Devon Janke and David V. Anderson. Analyzing the effects of noise and variation on the accuracy of analog neural networks. In *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, August 2020.
- [41] A. Aldo Faisal, Luc P. J. Selen, and Daniel M. Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, April 2008.
- [42] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, March 1950.
- [43] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.
- [44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [45] J Howard. Imagenette. Accessed: 25-10-2025.
- [46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.
- [47] Diego Doimo, Aldo Glielmo, Sebastian Goldt, and Alessandro Laio. Redundant representations help generalization in wide neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [48] Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws, 2022.
- [49] Blake Bordelon and Cengiz Pehlevan. Population codes enable learning from few examples by shaping inductive bias. *eLife*, 11, December 2022.

- [50] B OLSHAUSEN and D FIELD. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, August 2004.
- [51] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer New York, 1986.
- [52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32, pages 8024–8035. Curran Associates, Inc., 2019.
- [53] John R. Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, November 2000.
- [54] Wolfram Research, Inc. Mathematica, Version 13.3. Champaign, IL, 2023.

A Table of Parameters from Proposition 2 and Figures 2,3,4

Data Scale	s
Correlation Strength	c
Data-Task Alignment	ρ
Readout Noise Scale	η
Feature noise scale	ω
Label Noise Scale	ζ
Ensemble Size	k
Subsampling Fractions	$\nu_{rr} \quad r = 1, \dots, K$
Subsampling Overlap Fractions	$\nu_{rr'} \quad r \neq r'$
Sample Size	$P \rightarrow \infty$
Data Dimensionality	$M \rightarrow \infty$
Sample Complexity	$\alpha \equiv \frac{P}{M} \quad (\text{finite})$
Effective Noise-To-Signal Ratios	$H \equiv \frac{\eta^2}{s(1-c)}, W \equiv \frac{\omega^2}{s(1-c)}, Z \equiv \frac{\zeta^2}{s(1-c)}$

B Code Availability and Compute

All Code used in this paper has been made available online (see <https://github.com/benruben87/Learning-Curves-for-Heterogeneous-Feature-Subsampled-Ridge-Ensembles.git>). This includes code used to perform numerical experiments, calculate theoretical learning curves, and produce plots as well as the custom Mathematica libraries used to simplify the generalization error in the special case of equicorrelated data. The compute time required to do all of the calculations in this paper is approximately 3 GPU days.

C Homogeneous and Heterogeneous Subsampling

In this section, we describe the homogeneous and heterogeneous subsampling strategies, as used in this work, in detail. This sampling method was used in calculating the theoretical loss curves for heterogeneous subsampling experiments seen in main text Fig. 3, and in the heterogeneous subsampling experiments applied to the ResNext-features-based image classification task in Figs. 3e, S2.

In homogeneous ensembling, the subsampling fractions $\nu_{rr} = N_r/M$ are chosen as $\nu_{rr} = 1/k$ for all $r = 1, \dots, k$. In heterogeneous ensembling, the subsampling fractions $\{\nu_{11}, \dots, \nu_{kk}\}$ are generated according to the following statistical process:

1. Each fraction ν_{rr} is drawn independently as $\nu_{rr} \sim \Gamma_{k,\sigma}$, where a $\Gamma_{k,\sigma}$ represents a Gamma distribution with mean $\frac{1}{k}$ and variance σ^2 .
2. The fractions are re-scaled in order to sum to 1: $\nu_{rr} \rightarrow \nu_{rr}/(\nu_{11} + \dots + \nu_{kk})$

Equivalently, the subsampling fractions are drawn from a Dirichlet distribution parameterized by the ensemble size k and a chosen variance σ as $(\nu_{11}, \dots, \nu_{kk}) \sim \text{Dir}((\sigma k)^{-2}, \dots, (\sigma k)^{-2})$ [51].

In main text Fig. 3c, we combine this sampling strategy with theoretical learning curves for the equicorrelated data model in a quasi-numerical experiment. At each trial of the experiment, we draw a particular realization of the subsampling fractions $\{\nu_{11}, \dots, \nu_{kk}\}$, then use the analytical expression (eq. 16) to calculate the resulting learning curve. Dotted lines show the loss curves for 3 single trials, corresponding to three particular realizations of the subsampling fractions. The solid lines show the average over 100 trials.

Note that we have defined our own convention for the parameterization of the Γ distribution in which the inverse of the mean and the standard deviation are specified. In terms of the standard ‘‘shape’’ and ‘‘scale’’ parameters, we have:

$$\Gamma_{k,\sigma} \equiv \Gamma(\text{shape} = (k\sigma)^{-2}, \text{scale} = k\sigma^2) \quad (27)$$

D Numerical Linear Regression with Synthetic Datasets

Numerical experiments were performed using the PyTorch library [52]. The code used to perform numerical experiments and generate plots has been made publicly available (see section B).

In numerical regression experiments, synthetic datasets with label noise are constructed as described in section 2.1, drawing data randomly from multivariate Gaussian distributions and adding label noise (see “DatasetMaker.py” in available code). Representing the training set in terms of a data matrix $\Psi \in \mathbb{R}^{M \times P}$ in which column μ consist of the training point ψ_μ , and the labels with a column vector \mathbf{y} such that $y_\mu = y_\mu$, the learned weights are calculated as:

$$\hat{\mathbf{w}} = \Psi (\Psi^\top \Psi + \lambda \mathbf{I}_p)^{-1} \mathbf{y} \quad (28)$$

In the ridgeless case, a pseudoinverse is used:

$$\hat{\mathbf{w}} = \Psi^\dagger \mathbf{y} \quad (29)$$

E Ensembled Linear Classification of Imagenet Images

In this section, we provide the details of numerical experiments which demonstrate that qualitative insights gained from our analysis of the linear regression task with Gaussian data carries over to a practical machine learning task. In particular, we apply ensembles of linear classifiers to datasets constructed using a pre-trained ResNext [46] a specific type of Convolutional Neural Network (CNN).

E.1 Dataset Construction

To construct the dataset, we start with a set of n images $\{\mathbf{x}^\mu\}_{\mu=1}^n$ from a subset of $C = 10$ classes of the imagenet dataset. For each image \mathbf{x}^μ , we obtain a corresponding feature vector $\psi^\mu \in \mathbb{R}^M$ as the last-hidden-layer activation of the ResNext [46], which has been pre-trained on the imagenet classification task [44]. The architecture we use produces $M = 2048$ features per image. These features will serve as the data input to the downstream linear classifier. The corresponding labels $\mathbf{y}^\mu \in \mathbb{R}^C$ are one-hot vectors.

We construct two datasets using this method, using images from the “Imagenette” and “Imagewoof” datasets [45]. For the “Imagenette” task, we a construct a training set of size $n_{tr} = 9469$ and a test set of size $n_{test} = 3925$ containing features corresponding to images from 10 unrelated classes (tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, parachute). For the “Imagewoof” task, we a construct a training set of size $n_{tr} = 9025$ and a test set of size $n_{test} = 3929$ containing features corresponding to images of 10 different dog breeds (Australian terrier, Border terrier, Samoyed, Beagle, Shih-Tzu, English foxhound, Rhodesian ridgeback, Dingo, Golden retriever, Old English sheepdog). The imagewoof classification task is naturally more difficult. The statistics of the resulting datasets are described in Fig. S1, where we plot the data-data covariance matrix, feature-feature covariance matrix, and the eigenvalue spectrum for both the “Imagenette” and “Imagewoof” tasks.

E.2 Model Training

At training time a dataset of P examples is constructed: $\mathcal{D} = \{\psi^\mu, \mathbf{y}^\mu\}_{\mu=1}^P$. We represent the training set with a “design matrix” $\Phi \in \mathbb{R}^{P \times M}$ and a label matrix $\mathbf{Y} \in \{0, 1\}^{P \times C}$. The loss function for each ensemble member is generalized to multi-class regression. With ridge regularization, the objective for each ensemble member r becomes:

$$\hat{\mathbf{W}}_r = \arg \min_{\mathbf{W}_r \in \mathbb{R}^{N_r \times C}} \left[\sum_{\mu=1}^P \left\| \frac{1}{\sqrt{N_r}} \mathbf{A}_r \bar{\psi}^\mu \mathbf{W}_r + \xi_r^\mu - \mathbf{y}^\mu \right\|_2^2 + \lambda_r \|\mathbf{W}_r\|_F^2 \right],$$

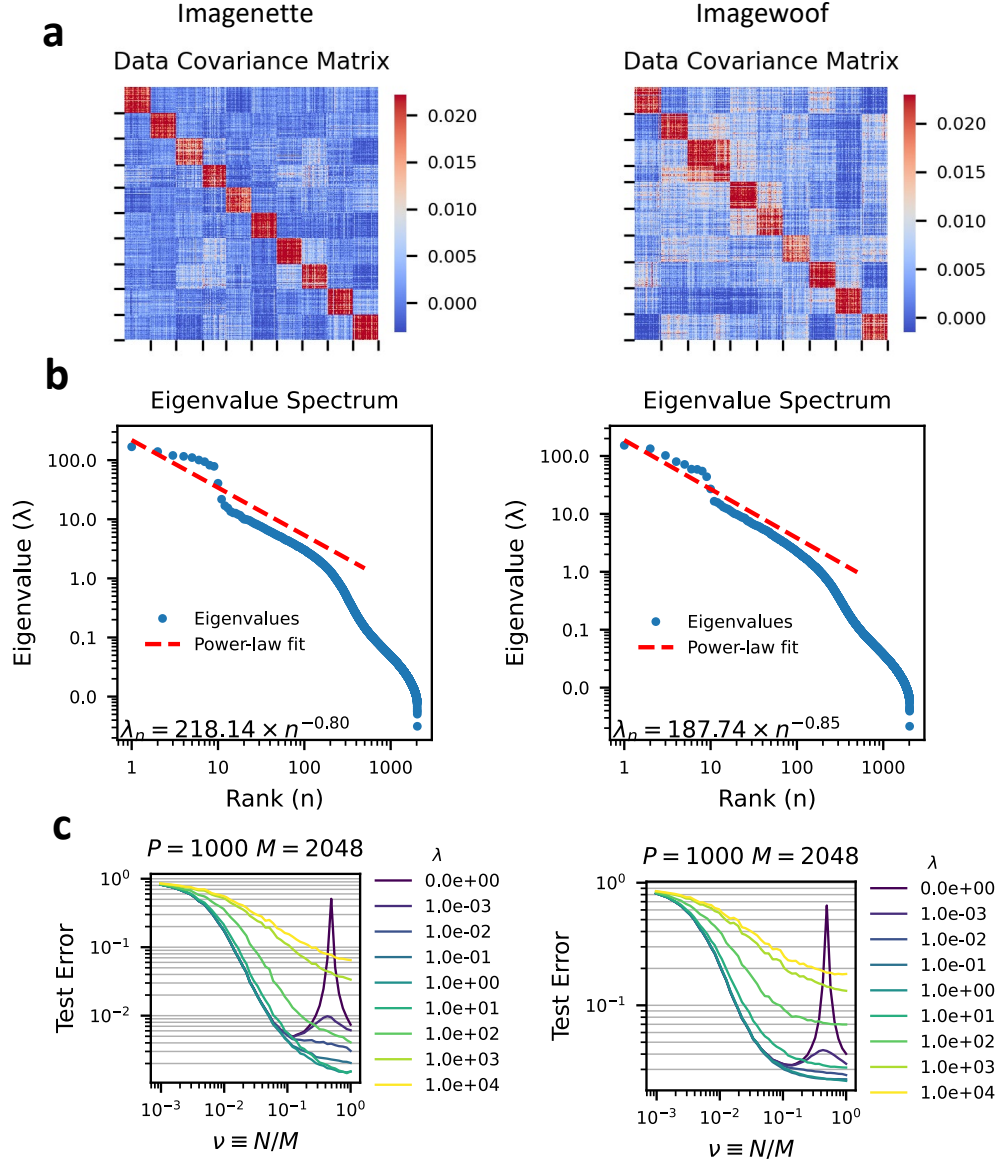


Figure S1: In numerical experiments, we train linear classifiers to predict labels of imagenet images based on their last-hidden-layer representations in a pre-trained ResNext deep learning architecture [46]. Here, we show the structure of the datasets constructed using the ResNext feature map for the Imagenette task (left), which consists of categorizing images from 10 unrelated categories, and the Imgewoof task (right), which consists of categorizing images from 10 different dog breeds. (a) Gram matrix of the centered ResNext features defined as $\frac{1}{P} (\Phi - \bar{\Phi})^\top (\Phi - \bar{\Phi})$ for data matrix $\Phi \in \mathbb{R}^{P \times M}$ where P is the total size of the dataset. Dataset is sorted by label and tick marks show the boundaries between classes. (b) The covariance eigenspectrum of the ResNext features is well described by a power law decay. (c) Generalization error of Linear classification with a single linear predictor with access to a fraction $\nu = N/M$ of the ResNext features averaged over 100 trials (see discussion in section E.4.1)

where $\|\cdot\|_F$ denotes the Frobenius norm and where the readout noise vector $\xi_r^\mu \sim \mathcal{N}(0, \eta^2 \mathbf{I}_C)$ is a C -dimensional readout noise which corrupts the model’s prediction. The weights $\hat{\mathbf{W}}_r$ can be determined in closed form as follows:

$$\hat{\mathbf{W}}_r = \frac{1}{\sqrt{N_r}} \left(\frac{1}{N_r} \mathbf{A}_r \Phi^\top \Phi \mathbf{A}_r^\top + \lambda_r \mathbf{I} \right)^{-1} (\mathbf{A}_r \Phi^\top (\mathbf{Y} - \Xi_r)).$$

where we have defined $[\Xi_r]_{\mu c} = [\xi_r^\mu]_c$. When $\lambda = 0$ we instead use a pseudoinverse rule. In all experiments presented here, we use measurement matrices \mathbf{A}_r which implement an ordinary subsampling of the features, so that the rows and columns of \mathbf{A}_r consist of one-hot vectors.

E.3 Model Prediction

Once trained, the learned weights may be used to predict the label of a new example ψ as follows. For $r = 1, \dots, k$ we calculate the prediction of each ensemble member by first assigning each class a “score”. The scores of predictor r are stored in a vector $\mathbf{f}_r \in \mathbb{R}^C$:

$$\mathbf{f}_r(\psi) = \frac{1}{\sqrt{N_r}} \mathbf{A}_r \psi \hat{\mathbf{W}}_r + \xi_r$$

Where $\xi_r \sim \mathcal{N}(0, \eta^2 \mathbf{I}_C)$ is drawn randomly at model evaluation. Each ensemble member’s “vote” corresponds to the class with the largest score. The prediction of the ensemble is then calculated as a majority vote of the ensemble members. Generalization error is then calculated as the probability of misclassifying an example from the test set.

E.4 Linear Classification Experiments

We apply the described majority-vote linear classifier ensembles to the ResNext-Imagenette and ResNext-Imagewoof tasks in three different experiments.

E.4.1 Reduncancy of ResNext Features

In the first experiment, we investigate the performance of a single linear classifier ($k = 1$) as the fraction of features ν which it has access to varies. We set $P = 1000$ and vary ν over 50 values on a logarithmic scale from 10^{-3} to 1. We also vary the regularization strength over 0 and a logarithmic scale from 10^{-3} to 10^4 . We average over 100 trials. At each trial the particular subset of $P = 1000$ training examples and the particular subset of $\nu * M$ features is randomly re-sampled. We find that ResNext features are highly redundant – classification accuracy is very robust to subsampling of the features. For example, in the Imagewoof classification task with the best regularization tested ($\lambda = 0.1$), test error increases from about 1% to about 2% as the subsampling fraction decreases from $\nu = 1$ to $\nu = 0.1$ (meaning 90% of the features are ignored) (see Fig. 3d). Similarly, for the imagenette task, test error increases from about 0.1% to about 0.2% as the subsampling fraction decreases from $\nu = 1$ to $\nu = 0.1$ (see Fig. S1c). Code used to run these experiments may be found in the folder “DeepNet_Subsamp” in the GitHub repository.

E.4.2 Heterogeneous Ensembling Mitigates Double-Descent

In the second experiment, we compare learning curves for homogeneous ensembling and heterogeneous ensembling applied to the ResNext-Imagewoof classification task. In each trial, we train ensembles of $k = \{1, 4, 8, 16, 32\}$ linear predictors whose subsampling fractions $\{\nu_{rr}, \dots, \nu_{kk}\}$ are assigned either with Homogeneous ensembling or with Heterogeneous Ensembling with $\sigma = 0.75/k$. After subsampling fraction are assigned, the training set is randomly shuffled. We iterate over 50 sample sizes P logarithmically distributed from 400 to 4000, and then add the values M/k for each k to the list of P values. We repeat for 100 trials for both $\lambda_r = 0$ (pseudoinverse rule) and $\lambda = 0.1$,

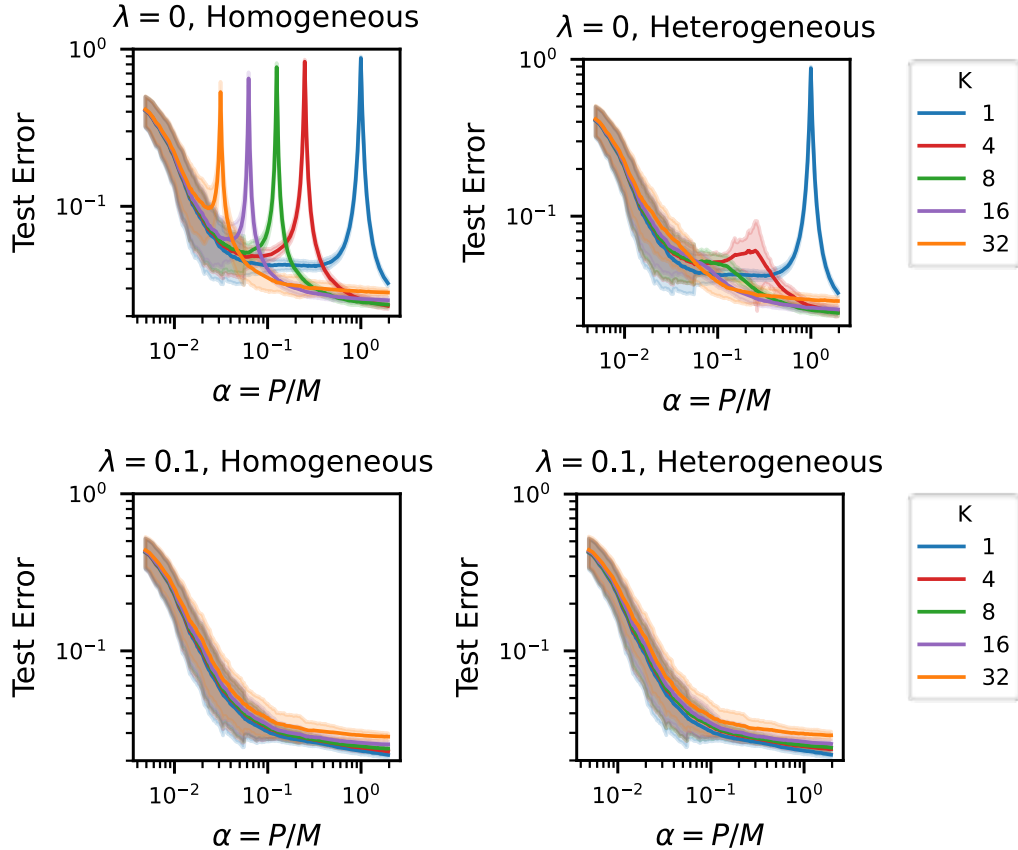


Figure S2: Results of the experiment described in section E.4.2. Learning curves for ResNext-Imagenetwoof classification task in linear classifier ensembles. In Homogeneous ensembling (left) all $\nu_{rr} = 1/k$ with k indicated by the legend. In heterogeneous ensembling ν_{rr} are drawn from a Dirichlet distribution as described in section C, with $\sigma = 3/(4k)$. We use regularization $\lambda = 0$ (top) and $\lambda = 0.1$ (bottom). $\lambda = 0, 0.1$. Lines represent an average over 100 trials, shaded regions show standard deviation. We set $\nu_{rr'} = 0$ for $r \neq r'$.

which was found to mitigate double-descent by the parameter sweep in Fig. S1c. In main text Fig. 3e, we plot the mean learning curves over 100 trials. In Fig. S2 we show standard deviation over the 100 trials as shaded error bars. When $\lambda = 0$, heterogeneous ensembling mitigates double-descent, leading to a monotonically decreasing learning curve for sufficiently high k . When $\lambda = 0.1$, homogeneous and heterogeneous ensembles of size k perform similarly. Code used to run these experiments may be found in the folder “DeepNet_HomVHet” in the GitHub repository.

E.4.3 Readout Noise Encourages Ensembling

In the third experiment, we test the effect of a readout noise which is independent across the members of the ensemble on generalization error. We do this by sweeping over the readout noise scale η as defined in sections E.2, E.3. For $\lambda = 0, 0.1$ and $\nu = 0, .1, \dots, 1$ we compute the learning curves of linear predictors with $k = 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 100, 200$ and all $\nu_{rr} = 1/k$, averaged over 50 trials. These learning curves are shown in Fig. S3 for both the ResNext-Imagenette and ResNext-Imagenetwoof task. In Figs. S4a and S5a, we plot the value k^* which minimizes error as phase diagrams in the parameter space of α and η , analogous to the phase diagrams in Fig. 4b. We see that the qualitative shape of these phase diagrams is similar to the equicorrelated model. The differences may be attributed to the nonlinear nature of the classification task. Furthermore, we find

that, in general the optimal k^* tends to be higher with the Imagenette dataset than with the Imagewoof dataset, in agreement with our finding in the equicorrelated regression model that as ρ increases (making the classification task easier), optimal ensemble size tends to increase (Fig. 4b and S6, S7, S8). In Fig S3 we see that there are often a number of k values for which test error is at or near to its lowest. To quantify this, we also plot diagrams of the minimum and maximum values of k that are within a small margin ϵ of the minimum measured error. For the ResNext-Imagenette task, we use $\epsilon = 0.001$ and for ResNext-Imagewoof $\epsilon = 0.01$. We see that there is a wide array of k values which bring error near-to-minimum in practice (Figs S4b,c, S5b,c). We also plot the minimum achieved error, and the difference between minimum errors at $\lambda = 0.1, 0$ (Figs. S4d, S5d). Code used to run these experiments may be found in the folder “DeepNet_PD” in the GitHub repository.

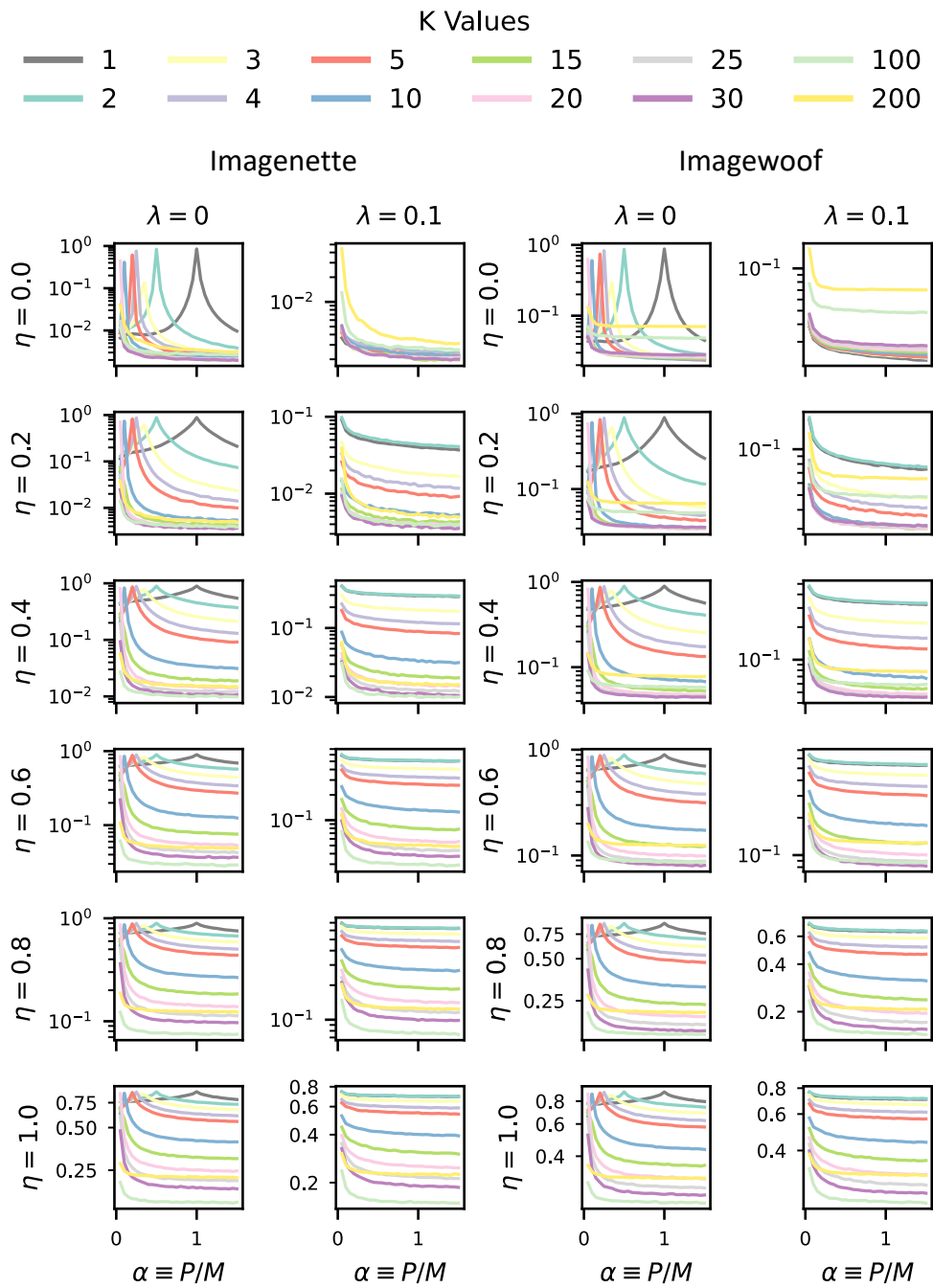


Figure S3: Learning curves for ensembles of linear classifiers with homogeneous subsampling for $\lambda = 0, 0.1$ and readout noise η values indicated in the figure. Results are averaged over 50 trials. Experiments are described in section E.4.3

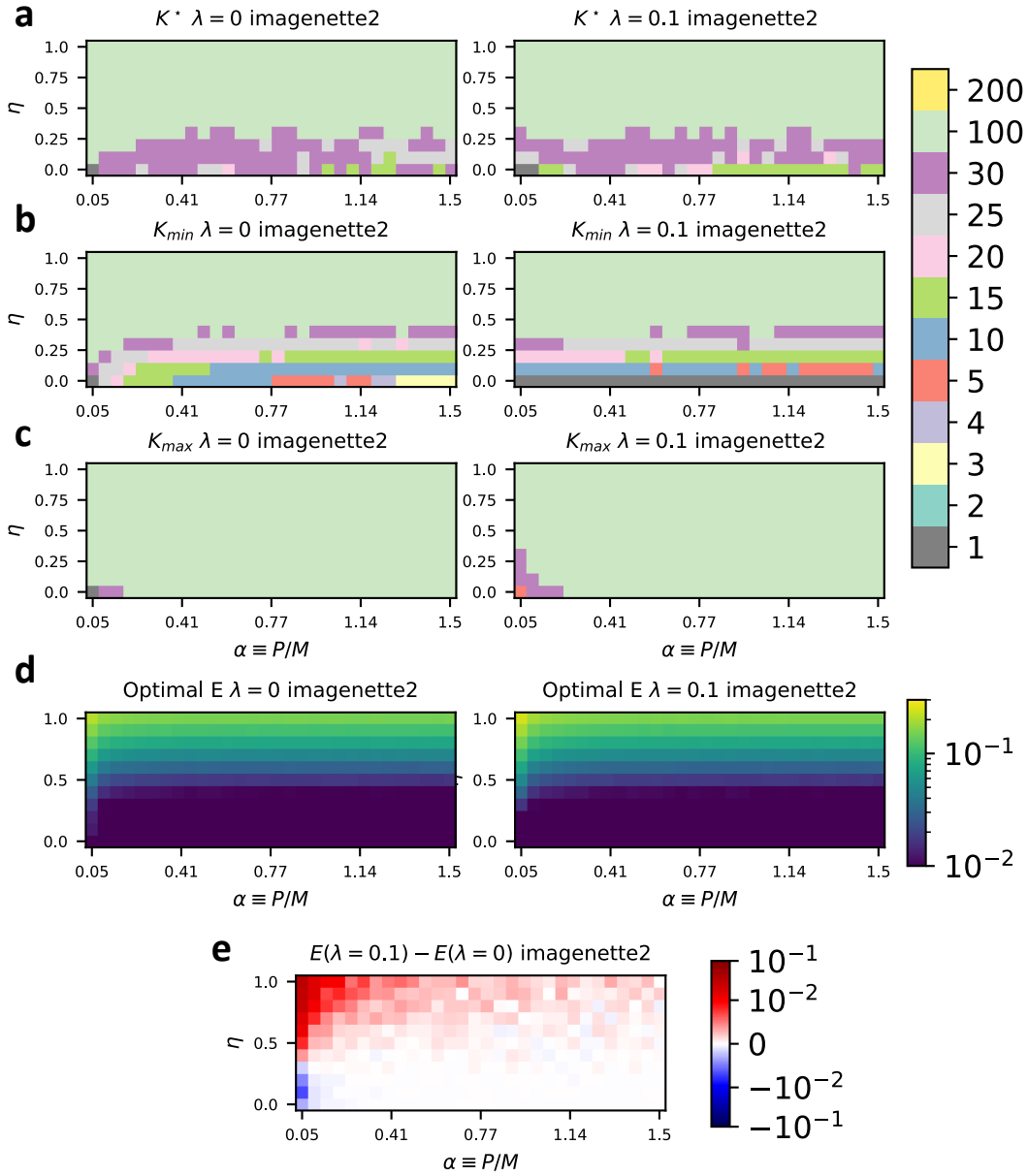


Figure S4: Diagrams described in section E.4.3 for the ResNext-Imagenette experiment. Using learning curves in Fig. S3 (and for additional values of η not shown there), we plot (a) Optimal k^* in the parameter space of α and η , (b) Minimum value of k for which error is within a tolerance $\epsilon = .001$ of its value for k^* , (c) Maximum value of k for which error is within a tolerance $\epsilon = .001$ of its value for k^* , (d) the value of the minimum error $E(k^*)$, and (e) the difference between this optimal error for $\lambda = 0.1$ and $\lambda = 0$.

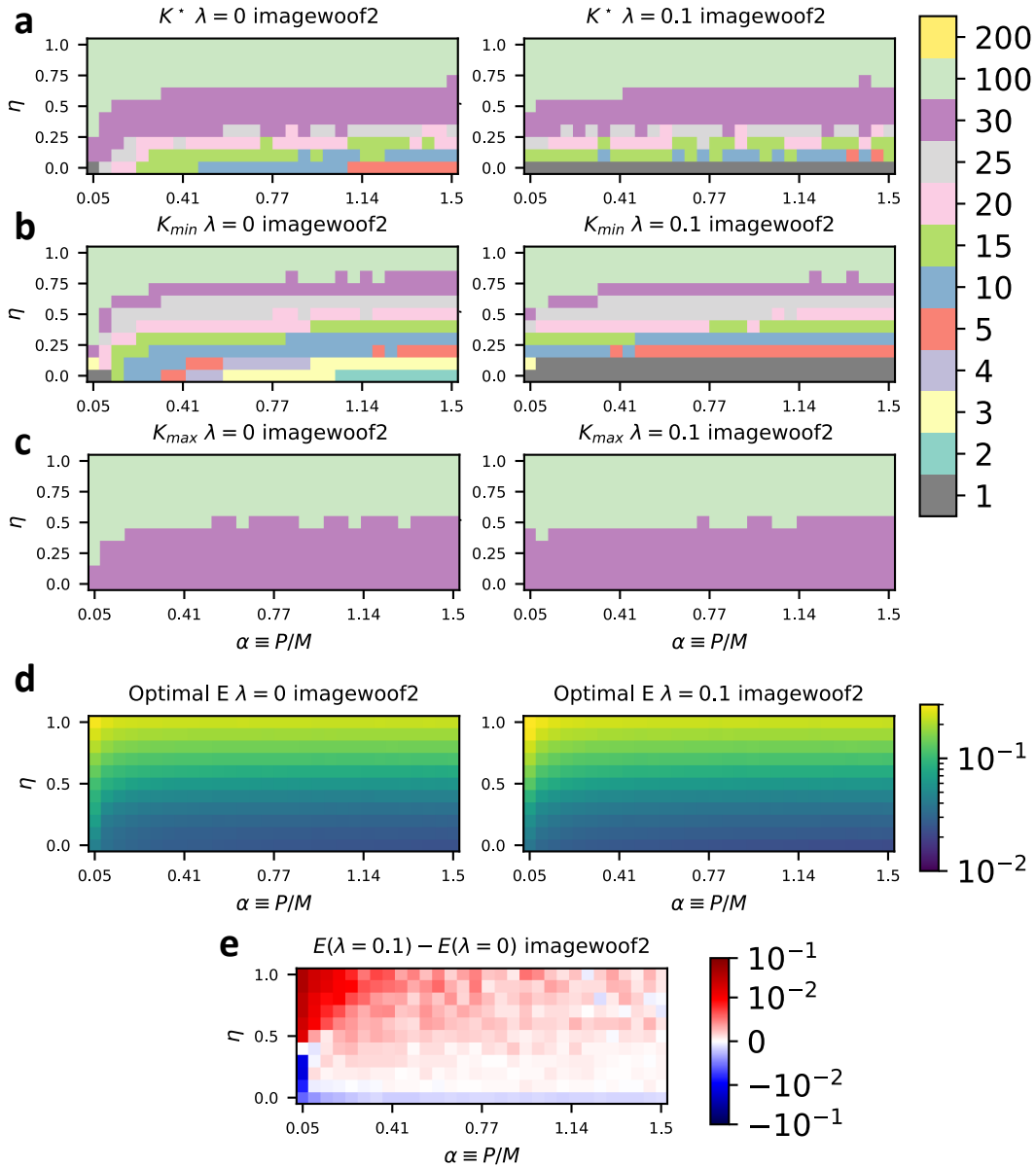


Figure S5: Diagrams described in section E.4.3 for the ResNext-Imagewoof experiment. Using learning curves in Fig. S3 (and for additional values of η not shown there), we plot (a) Optimal k^* in the parameter space of α and η , (b) Minimum value of k for which error is within a tolerance $\epsilon = .001$ of its value for k^* , (c) Maximum value of k for which error is within a tolerance $\epsilon = .001$ of its value for k^* , (d) the value of the minimum error $E(k^*)$, and (e) the difference between this optimal error for $\lambda = 0.1$ and $\lambda = 0$.

F Generalization error of ensembled linear regression from the replica trick

Here we use the Replica Trick from statistical physics to derive analytical expressions for $E_{rr'}$. We treat the cases where $r = r'$ and $r \neq r'$ separately. Following a statistical mechanics approach, we calculate the average generalization error over a Gibbs measure with inverse temperature β ;

$$Z = \int \prod_r d\mathbf{w} \exp \left(-\frac{\beta}{2} \sum_r E_t^r - \frac{M\beta}{2} \sum_{r,r'} J_{rr'} E_{rr'}(\mathbf{w}_r, \mathbf{w}_{r'}) \right) \quad (30)$$

$$E_t^r = \sum_{\mu=1}^P \left(\frac{1}{\sqrt{N_r}} \mathbf{w}_r^\top \mathbf{A}_r \bar{\psi}_\mu + \xi_r - y_\mu \right)^2 + \lambda |\mathbf{w}_r^2| \quad (31)$$

In the limit where $\beta \rightarrow \infty$ the gibbs measure will concentrate around the weight vector $\hat{\mathbf{w}}_r$ which minimizes the regularized loss function. The replica trick relies on the following identity:

$$\langle \log(Z[\mathcal{D}]) \rangle_{\mathcal{D}} = \lim_{n \rightarrow 0} \frac{1}{n} \log(\langle Z^n \rangle_{\mathcal{D}}) \quad (32)$$

where $\langle \cdot \rangle_{\mathcal{D}}$ represents an average over all quenched disorder in the system. In this case, quenched disorder – the disorder which is fixed prior to and throughout training of the weights – consists of the selected training examples along with their feature noise and label noise: $\mathcal{D} = \{\psi_\mu, \sigma^\mu, \epsilon^\mu\}_{\mu=1}^P$. The calculation proceeds by first computing the average of the replicated partition function assuming n is a positive integer. Then, in a non-rigorous but standard step, we analytically extend the result to $n \rightarrow 0$.

F.1 Diagonal Terms

We start by calculating E_{rr} for some fixed choice of r . This derivation partially follows section D.3 from [36], with the addition of readout noise and label noise. Noting that the diagonal terms of the generalization error E_{rr} only depend on the learned weights \mathbf{w}_r , and the loss function separates over the readouts, we may consider the Gibbs measure over only these weights:

$$Z = \int d\mathbf{w}_r \exp \left(-\frac{\beta}{2\lambda} E_r^t - \frac{JM\beta}{2} E_{rr}(\mathbf{w}_r) \right) \quad (33)$$

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} &= \int \prod_a d\mathbf{w}_r^a \mathbb{E}_{\{\psi_\mu, \sigma^\mu, \epsilon^\mu\}} \\ &\exp \left(-\frac{\beta M}{2\lambda} \sum_{\mu,a} \frac{1}{M} \left[\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r (\psi_\mu + \sigma^\mu) - \mathbf{w}^{*\top} \psi_\mu - \sqrt{M}(\epsilon^\mu - \xi_r^\mu) \right]^2 \right. \\ &\quad \left. - \frac{\beta}{2} \sum_a |\mathbf{w}_r^a|^2 - \frac{JM\beta}{2} \sum_a E_{rr}(\mathbf{w}^a) \right) \end{aligned} \quad (34)$$

Next we must perform the averages over quenched disorder. We first integrate over $\{\psi_\mu, \sigma^\mu, \xi_r^\mu, \epsilon^\mu\}_{\mu=1}^P$. Noting that the scalars

$$h_\mu^{ra} \equiv \frac{1}{\sqrt{M}} \left[\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r (\psi_\mu + \sigma^\mu) - \mathbf{w}^{*\top} \psi_\mu - \sqrt{M}(\epsilon^\mu - \xi_r^\mu) \right]$$

are Gaussian random variables (when conditioned on A_r) with mean zero and covariance:

$$\langle h_\mu^{ra} h_\nu^{rb} \rangle = \delta_{\mu\nu} Q_{ab}^{rr} \quad (35)$$

$$Q_{ab}^{rr} = \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r^\top \mathbf{w}_r^b - \mathbf{w}^* \right) + \frac{1}{\nu_{rr}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_r^\top \mathbf{w}_r^b + M(\zeta^2 + \eta_r^2) \right] \quad (36)$$

To perform this integral we re-write in terms of $\{\mathbf{H}_\mu^r\}_{\mu=1}^P$, where

$$\mathbf{H}_\mu^r = \begin{bmatrix} h_\mu^{r1} \\ h_\mu^{r2} \\ \vdots \\ h_\mu^{rn} \end{bmatrix} \in \mathbb{R}^n \quad (37)$$

$$\langle Z^n \rangle_{\mathcal{D}} = \int \prod_a d\mathbf{w}_r^a \mathbb{E}_{\{\psi_\mu, \sigma^\mu, \epsilon^\mu\}} \exp \left(-\frac{\beta}{2\lambda} \sum_\mu \mathbf{H}_\mu^{r\top} \mathbf{H}_\mu^r - \frac{\beta}{2} \sum_a |\mathbf{w}_r^a|^2 - \frac{JM\beta}{2} \sum_a E_{rr}(\mathbf{w}_r^a) \right) \quad (38)$$

Integrating over the \mathbf{H}_μ^r we get:

$$\langle Z^n \rangle_{\mathcal{D}} = \int \prod_a d\mathbf{w}_r^a \exp \left(-\frac{P}{2} \log \det \left(\mathbf{I}_n + \frac{\beta}{\lambda} \mathbf{Q}^{rr} \right) - \frac{\beta}{2} \sum_a |\mathbf{w}_r^a|^2 - \frac{JM\beta}{2} \sum_a E_{rr}(\mathbf{w}_r^a) \right) \quad (39)$$

Next we integrate over \mathbf{Q}^r and add constraints. We use the following identity:

$$1 = \prod_{ab'} \int dQ_{ab}^{rr} \delta \left(Q_{ab}^{rr} - \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r^\top \mathbf{w}_r^b - \mathbf{w}^* \right) + \frac{1}{\nu_{rr}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_r^\top \mathbf{w}_r^b + M(\zeta^2 + \eta_r^2) \right] \right) \quad (40)$$

Using the Fourier representation of the delta function, we get:

$$1 = \prod_{ab} \int \frac{1}{4\pi i/M} dQ_{ab}^{rr} d\hat{Q}_{ab}^{rr} \exp \left(\frac{M}{2} \hat{Q}_{ab}^{rr} \left(Q_{ab}^{rr} - \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r^\top \mathbf{w}_r^b - \mathbf{w}^* \right) + \frac{1}{\nu_{rr}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_r^\top \mathbf{w}_r^b + M(\zeta^2 + \eta_r^2) \right] \right) \right) \quad (41)$$

Inserting this identity into the replicated partition function and substituting $E_{rr}(\mathbf{w}_r^a) = Q_{aa}^{rr} - \zeta^2$ we find:

$$\begin{aligned}
\langle Z^n \rangle_{\mathcal{D}} &\propto \\
&\int \prod_{ab} dQ_{ab}^{rr} d\hat{Q}_{ab}^{rr} \exp \left(-\frac{P}{2} \log \det \left(\mathbf{I}_n + \frac{\beta}{\lambda} \mathbf{Q}^{rr} \right) + \frac{1}{2} \sum_{ab} M \hat{Q}_{ab}^{rr} Q_{ab}^{rr} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr} - \zeta^2) \right) \\
&\int \prod_a d\mathbf{w}_r^a \exp \left(-\frac{\beta}{2} \sum_a |\mathbf{w}_r^a|^2 - \frac{1}{2} \sum_{ab} \hat{Q}_{ab}^{rr} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r^\top \mathbf{w}_r^b - \mathbf{w}^* \right) \right. \right. \\
&\quad \left. \left. + \frac{1}{\nu_{rr}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_r^\top \mathbf{w}_r^b + M(\zeta^2 + \eta_r^2) \right] \right)
\end{aligned} \tag{42}$$

In order to perform the Gaussian integral over the $\{\mathbf{w}_r^a\}$, we unfold over the replica index a . We first define the following:

$$\mathbf{w}_r \equiv \begin{bmatrix} \mathbf{w}_r^1 \\ \vdots \\ \mathbf{w}_r^n \end{bmatrix} \tag{43}$$

$$T^r \equiv \beta \mathbf{I}_n \otimes \mathbf{I}_{N_r} + \hat{\mathbf{Q}}^{rr} \otimes \left(\frac{1}{\nu_{rr}} \mathbf{A}_r (\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_0) \mathbf{A}_r^\top \right) \tag{44}$$

$$V^r \equiv (\hat{\mathbf{Q}}^{rr} \otimes \mathbf{I}_{N_r}) (\mathbf{1}_n \otimes \frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r \boldsymbol{\Sigma}_s \mathbf{w}^*) \tag{45}$$

We then have for the integral over w

$$\int d\mathbf{w}_r \exp \left(-\frac{1}{2} \mathbf{w}_r^\top T^r \mathbf{w}_r + V^{r\top} \mathbf{w}_r \right) \tag{46}$$

$$= \exp \left(\frac{1}{2} V^{r\top} (T^r)^{-1} V^r - \frac{1}{2} \log \det(T^r) \right) \tag{47}$$

We can finally write the replicated partition function as:

$$\begin{aligned}
\langle Z^n \rangle_{\mathcal{D}} &\propto \\
&\int \prod_{ab} dQ_{ab}^{rr} d\hat{Q}_{ab}^{rr} \exp \left(-\frac{P}{2} \log \det \left(\mathbf{I}_n + \frac{\beta}{\lambda} \mathbf{Q}^{rr} \right) + \frac{1}{2} \sum_{ab} M \hat{Q}_{ab}^{rr} Q_{ab}^{rr} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr} - \zeta^2) \right) \\
&\exp \left(\frac{1}{2} V^{r\top} (T^r)^{-1} V^r - \frac{1}{2} \log \det(T^r) - \frac{1}{2} \sum_{ab} \hat{Q}_{ab}^{rr} (M(\zeta^2 + \eta_r^2) + \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \mathbf{w}^*) \right)
\end{aligned} \tag{48}$$

We now make the following replica-symmetric ansatz:

$$Q_{ab}^{rr} = \beta^{-1} q \delta_{ab} + q_0 \tag{49}$$

$$\hat{Q}_{ab}^{rr} = \beta \hat{q} \delta_{ab} + \beta^2 \hat{q}_0 \tag{50}$$

which is well-motivated because the loss function is convex. We will verify that the chosen scalings are self-consistent in the zero-temperature limit where $\beta \rightarrow \infty$. We may then rewrite the partition function as follows:

$$\langle Z^n \rangle_{\mathcal{D}} = \exp \left(-\frac{nM}{2} \mathbf{g} [q, q_0, \hat{q}, \hat{q}_0] \right) \quad (51)$$

Where the effective action is written:

$$\begin{aligned} \mathbf{g} [q, q_0, \hat{q}, \hat{q}_0] &= \alpha \left[\log \left(1 + \frac{q}{\lambda} \right) + \frac{\beta q_0}{\lambda + q} \right] - (q\hat{q} + \beta q\hat{q}_0 + \beta q_0\hat{q}) + \beta J [(\beta^{-1}q + q_0) - \zeta^2] \\ &- \frac{\beta}{\nu_{rr}M} \hat{q}^2 \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s \mathbf{w}^* + \frac{1}{M} \left[\log \det(\mathbf{G}) + \beta \hat{q}_0 \text{tr}[\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}}] \right] + \beta \hat{q} \left(\zeta^2 + \eta_r^2 + \frac{1}{M} \mathbf{w}^{*\top} \boldsymbol{\Sigma}_s \mathbf{w}^* \right) \end{aligned} \quad (52)$$

Where $\mathbf{G} \equiv \mathbf{I}_{N_r} + \hat{q} \tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}} \equiv \frac{1}{\nu_{rr}} \mathbf{A}_r (\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_0) \mathbf{A}_r^\top$

To determine the values of the order parameters we set the derivatives of $\mathbf{g} [q, q_0, \hat{q}, \hat{q}_0]$, evaluated at zero source ($J = 0$), to zero:

$$\frac{\partial \mathbf{g}}{\partial q_0} = 0 = \frac{\alpha \beta}{\lambda + q} - \beta \hat{q} \quad \Rightarrow \hat{q} = \frac{\alpha}{\lambda + q} \quad (53)$$

$$\frac{\partial \mathbf{g}}{\partial \hat{q}_0} = 0 = -\beta q + \frac{\beta}{M} \text{tr} [\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}}] \quad \Rightarrow q = \frac{1}{M} \text{tr} [\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}}] \quad (54)$$

$$\frac{\partial \mathbf{g}}{\partial q} = 0 = \frac{\alpha}{\lambda + q} - \frac{\alpha q_0 \beta}{(\lambda + \beta q)^2} - \hat{q} - \beta \hat{q}_0 \quad \Rightarrow \hat{q}_0 = -\frac{\alpha q_0}{(\lambda + \beta q)^2} \quad (55)$$

$$\begin{aligned} \frac{\partial \mathbf{g}}{\partial \hat{q}} = 0 &= -q - \beta q_0 + \beta \zeta^2 + \beta \eta_r^2 + \frac{1}{M} \text{tr} [\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}}] - \frac{\beta \hat{q}_0}{M} \text{tr} \left[(\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}})^2 \right] \\ &+ \frac{\beta}{M} \mathbf{w}^{*\top} \left[\boldsymbol{\Sigma}_s - \frac{2}{\nu_{rr}} \hat{q} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s + \frac{1}{\nu_{rr}} \hat{q}^2 \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}} \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s \right] \mathbf{w}^* \\ \Rightarrow q_0 &= \frac{1}{1 - \gamma} \left(\zeta^2 + \eta_r^2 + \frac{1}{M} \mathbf{w}^{*\top} \left[\boldsymbol{\Sigma}_s - \frac{2}{\nu_{rr}} \hat{q} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s + \frac{1}{\nu_{rr}} \hat{q}^2 \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}} \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s \right] \mathbf{w}^* \right) \end{aligned} \quad (56)$$

$$(57)$$

where $\gamma \equiv \frac{\alpha}{M\kappa^2} \text{tr} [(\mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}})^2]$ and $\kappa \equiv \lambda + q$. We retroactively confirm that the chosen β -scalings are correct by noticing that the saddle-point equations permit a solution where all order parameters remain $\mathcal{O}(1)$ as $\beta \rightarrow \infty$. The error is given as:

$$E_{rr} = \lim_{\beta \rightarrow \infty} \frac{1}{\beta} \frac{\partial}{\partial J} \mathbf{g} [q, q_0, \hat{q}, \hat{q}_0] = \lim_{\beta \rightarrow \infty} \beta^{-1} q + q_0 - \zeta^2 = q_0 - \zeta^2 \quad (58)$$

Where $q, q_0, \hat{q}, \hat{q}_0$ are the solutions to the saddle-point equations 53, 54, 55, 57. Substituting eq. 57 for q_0 , we obtain:

$$E_{rr} = \frac{1}{1 - \gamma} \frac{1}{M} \mathbf{w}^{*\top} \left[\boldsymbol{\Sigma}_s - \frac{2}{\nu_{rr}} \hat{q} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s + \frac{1}{\nu_{rr}} \hat{q}^2 \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \tilde{\boldsymbol{\Sigma}} \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s \right] \mathbf{w}^* + \frac{\gamma \zeta^2 + \eta_r^2}{1 - \gamma} \quad (59)$$

$$= \frac{1}{1 - \gamma} \frac{1}{M} \mathbf{w}^{*\top} \left[\boldsymbol{\Sigma}_s - \frac{1}{\nu_{rr}} \hat{q} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-1} \mathbf{A}_r \boldsymbol{\Sigma}_s - \frac{1}{\nu_{rr}} \hat{q} \boldsymbol{\Sigma}_s \mathbf{A}_r^\top \mathbf{G}^{-2} \mathbf{A}_r \boldsymbol{\Sigma}_s \right] \mathbf{w}^* + \frac{\gamma \zeta^2 + \eta_r^2}{1 - \gamma} \quad (60)$$

F.2 Off-Diagonal Terms

We now calculate $E_{rr'}$ for $r \neq r'$. We now must consider the joint Gibbs Measure over \mathbf{w}_r and $\mathbf{w}_{r'}$:

$$Z = \int d\mathbf{w}_r d\mathbf{w}_{r'} \exp \left(-\frac{\beta}{2\lambda} (E_r^t + E_{r'}^t) - \frac{JM\beta}{2} E_{rr'}(\mathbf{w}_r, \mathbf{w}_{r'}) \right) \quad (61)$$

$$(62)$$

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} &= \int \prod_a d\mathbf{w}_r^a d\mathbf{w}_{r'}^a \mathbb{E}_{\{\psi_\mu, \sigma^\mu, \epsilon^\mu\}} \\ &\exp \left(-\frac{\beta M}{2\lambda} \sum_{\mu, a} \frac{1}{M} \left[(h_\mu^{ra})^2 + (h_\mu^{r'a})^2 \right] - \frac{\beta}{2} \sum_a [|\mathbf{w}_r^a|^2 + |\mathbf{w}_{r'}^a|^2] - \frac{JM\beta}{2} \sum_a E_{rr'}(\mathbf{w}_r^a, \mathbf{w}_{r'}^a) \right) \end{aligned} \quad (63)$$

Where the h_μ^{ra} are defined as before. Next we must perform the averages over quenched disorder. To do so, we note that the h_μ^{ra} are Gaussian random variables with covariance structure:

$$\langle h_\mu^{ra} h_\nu^{r'b} \rangle = \delta_{\mu\nu} Q_{ab}^{rr'} \quad (64)$$

$$\begin{aligned} Q_{ab}^{rr'} &= \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{r'r'}}} \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b - \mathbf{w}^* \right) \right. \\ &\quad \left. + \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b + M\zeta^2 \right] \end{aligned} \quad (65)$$

To perform this integral we re-write in terms of $\{\mathbf{H}_\mu\}_{\mu=1}^P$, where

$$\mathbf{H}_\mu = \begin{bmatrix} \mathbf{H}_\mu^r \\ \mathbf{H}_\mu^{r'} \end{bmatrix} \in \mathbb{R}^{2n} \quad (66)$$

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} &= \int \prod_a d\mathbf{w}_r^a d\mathbf{w}_{r'}^a \mathbb{E}_{\{\psi_\mu, \sigma^\mu, \epsilon^\mu\}} \\ &\exp \left(-\frac{\beta}{2\lambda} \sum_\mu \mathbf{H}_\mu^\top \mathbf{H}_\mu - \frac{\beta}{2} \sum_a [|\mathbf{w}_r^a|^2 + |\mathbf{w}_{r'}^a|^2] - \frac{JM\beta}{2} \sum_a E_{rr'}(\mathbf{w}_r^a, \mathbf{w}_{r'}^a) \right) \end{aligned} \quad (67)$$

Integrating over \mathbf{H}_μ we get:

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} &= \int \prod_a d\mathbf{w}_r^a d\mathbf{w}_{r'}^a \\ &\exp \left(-\frac{P}{2} \log \det \left(\mathbf{I}_{2n} + \frac{\beta}{\lambda} \mathbf{Q} \right) - \frac{\beta}{2} \sum_a [|\mathbf{w}_r^a|^2 + |\mathbf{w}_{r'}^a|^2] - \frac{JM\beta}{2} \sum_a E_{rr'}(\mathbf{w}_r^a, \mathbf{w}_{r'}^a) \right) \end{aligned} \quad (68)$$

Where we have defined the matrix \mathbf{Q} so that:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{rr} & \mathbf{Q}^{rr'} \\ \mathbf{Q}^{r'r'} & \mathbf{Q}^{r'r} \end{bmatrix} \quad (69)$$

Next we integrate over \mathbf{Q} and add constraints. We use the following identity:

$$1 = \prod_{ab} \int dQ_{ab}^{rr'} \delta \left(Q_{ab}^{rr'} - \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{r'r'}}} \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b - \mathbf{w}^* \right) + \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b + M\zeta^2 \right] \right) \quad (70)$$

Using the Fourier representation of the delta function, we get:

$$1 = \prod_{ab} \int \frac{1}{4\pi i/M} dQ_{ab}^{rr'} d\hat{Q}_{ab}^{rr'} \exp \left(\frac{M}{2} \hat{Q}_{ab}^{rr'} \left(Q_{ab}^{rr'} - \frac{1}{M} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \mathbf{w}_r^{a\top} \mathbf{A}_r - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{r'r'}}} \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b - \mathbf{w}^* \right) + \frac{1}{\nu_{rr}} \mathbf{w}_r^{a\top} \mathbf{A}_r \boldsymbol{\Sigma}_0 \mathbf{A}_{r'}^\top \mathbf{w}_{r'}^b + M\zeta^2 \right] \right) \right) \quad (71)$$

Inserting this identity and the corresponding statements for Q_{ab}^{rr} and $Q_{ab}^{r'r'}$ into the replicated partition function and substituting $E_{rr'}(\mathbf{w}^a) = Q_{aa}^{rr'} - \zeta^2$ we find:

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} &\propto \int \prod_{abr_1r_2} dQ_{ab}^{r_1r_2} d\hat{Q}_{ab}^{r_1r_2} \\ &\exp \left(-\frac{P}{2} \log \det \left(\mathbf{I}_{2n} + \frac{\beta}{\lambda} \mathbf{Q} \right) + \frac{1}{2} \sum_{abr_1r_2} M \hat{Q}_{ab}^{r_1r_2} Q_{ab}^{r_1r_2} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr'} - \zeta^2) \right) \\ &\int \prod_a d\mathbf{w}_r^a d\mathbf{w}_{r'}^a \exp \left(-\frac{\beta}{2} \sum_a [|\mathbf{w}_r^a|^2 + |\mathbf{w}_{r'}^a|^2] - \frac{1}{2} \sum_{abr_1r_2} \hat{Q}_{ab}^{r_1r_2} \left[\left(\frac{1}{\sqrt{\nu_{r_1}}} \mathbf{w}_{r_1}^{a\top} \mathbf{A}_{r_1} - \mathbf{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{r_2}}} \mathbf{A}_{r_2}^\top \mathbf{w}_{r_2}^b - \mathbf{w}^* \right) + \frac{1}{\sqrt{\nu_{r_1}\nu_{r_2}}} \mathbf{w}_{r_1}^{a\top} \mathbf{A}_{r_1} \boldsymbol{\Sigma}_0 \mathbf{A}_{r_2}^\top \mathbf{w}_{r_2}^b + M\zeta^2 \right] \right) \end{aligned} \quad (72)$$

Where sums over r_1 and r_2 run over $\{r, r'\}$.

In order to perform the Gaussian integral over the $\{\mathbf{w}_r^a\}$, we unfold in two steps. We first define the following:

$$\mathbf{w}_r \equiv \begin{bmatrix} \mathbf{w}_r^1 \\ \vdots \\ \mathbf{w}_r^n \end{bmatrix} \quad (73)$$

$$[\hat{Q}^{rr'}]_{ab} \equiv \hat{Q}_{ab}^{rr'} \quad (74)$$

$$\tilde{\boldsymbol{\Sigma}}_{rr'} \equiv \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}} \mathbf{A}_r [\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_0] \mathbf{A}_{r'}^\top \quad (75)$$

$$T^{rr'} \equiv \beta \delta_{rr'} \mathbf{I}_n \otimes \mathbf{I}_{N_r} + \hat{Q}^{rr'} \otimes \tilde{\boldsymbol{\Sigma}}_{rr'} \quad (76)$$

Unfolding over the replica indices, we then get:

$$\begin{aligned}
\langle Z^n \rangle_{\mathcal{D}} &\propto \int \prod_{abr_1r_2} dQ_{ab}^{r_1r_2} d\hat{Q}_{ab}^{r_1r_2} \\
&\exp\left(-\frac{P}{2} \log \det \left(\mathbf{I}_{2n} + \frac{\beta}{\lambda} \mathbf{Q} \right) + \frac{1}{2} \sum_{abr_1r_2} M \hat{Q}_{ab}^{r_1r_2} Q_{ab}^{r_1r_2} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr'} - \zeta^2)\right) \\
&\exp\left(-\frac{1}{2} \sum_{abr_1r_2} \hat{Q}_{ab}^{r_1r_2} (\mathbf{w}^{*\top} \Sigma_s \mathbf{w}^* + M\zeta^2)\right) \\
&\int d\mathbf{w}_r d\mathbf{w}_{r'} \exp\left(-\frac{1}{2} \sum_{r_1r_2} \mathbf{w}_{r_1}^\top T^{r_1r_2} \mathbf{w}_{r_2} + \sum_{r_1r_2} \left[(\hat{\mathbf{Q}}^{r_1r_2} \otimes \mathbf{I}_{N_{r_1}}) \left(\mathbf{1}_n \otimes \frac{1}{\sqrt{\nu_{r_1}}} \mathbf{A}_{r_1} \Sigma_s \mathbf{w}^* \right) \right]^\top \mathbf{w}_{r_1} \right)
\end{aligned} \tag{77}$$

Note that the dimensionality of $T^{r_1r_2}$ varies for different choices of r_1 and r_2 . Next, we unfold over the two readouts:

$$\mathbf{w} \equiv \begin{bmatrix} \mathbf{w}_r \\ \mathbf{w}_{r'} \end{bmatrix} \tag{78}$$

$$T \equiv \begin{bmatrix} T^{rr} & T^{rr'} \\ T^{r'r} & T^{r'r'} \end{bmatrix} \tag{79}$$

$$V \equiv \begin{bmatrix} \left((\hat{\mathbf{Q}}^{rr} + \hat{\mathbf{Q}}^{rr'}) \otimes \mathbf{I}_{N_r} \right) \left(\mathbf{1}_n \otimes \frac{1}{\sqrt{\nu_{rr}}} \mathbf{A}_r \Sigma_s \mathbf{w}^* \right) \\ \left((\hat{\mathbf{Q}}^{r'r'} + \hat{\mathbf{Q}}^{r'r}) \otimes \mathbf{I}_{N_{r'}} \right) \left(\mathbf{1}_n \otimes \frac{1}{\sqrt{\nu_{r'r'}}} \mathbf{A}_{r'} \Sigma_s \mathbf{w}^* \right) \end{bmatrix} \tag{80}$$

The integral over \mathbf{w} then becomes:

$$\int d\mathbf{w} \exp\left(-\frac{1}{2} \mathbf{w}^\top T \mathbf{w} + V^\top \mathbf{w}\right) \propto \exp\left(\frac{1}{2} V^\top T^{-1} V - \frac{1}{2} \log \det T\right) \tag{81}$$

We are now ready to make a replica-symmetric ansatz. The order parameter that we wish to constrain is $Q_{ab}^{rr'}$. Overlaps go between the weights from different replicas of the system as well as different readouts. The scale of the overlap between two measurements depends on their overlap with each other and with the principal components of the data distribution. An ansatz which is replica-symmetric but makes no assumptions about the overlaps between different measurements is as follows:

$$Q_{ab}^{r_1r_2} = \beta^{-1} q^{r_1r_2} \delta_{ab} + Q^{r_1r_2} \tag{82}$$

$$\hat{Q}_{ab}^{r_1r_2} = \beta \hat{q}^{r_1r_2} \delta_{ab} + \beta^2 \hat{Q}^{r_1r_2} \tag{83}$$

Next step is to plug the RS ansatz into the free energy and simplify. To make calculations more transparent, we re-label the parameters in the RS ansatz as follows:

$$\mathbf{Q}^{rr} = \beta^{-1} q \mathbf{I} + Q \mathbf{11}^\top \tag{84}$$

$$\mathbf{Q}^{r'r'} = \beta^{-1} r \mathbf{I} + R \mathbf{11}^\top \tag{85}$$

$$\mathbf{Q}^{rr'} = \beta^{-1} c \mathbf{I} + C \mathbf{11}^\top \tag{86}$$

$$\hat{\mathbf{Q}}^{rr} = \beta \hat{q} \mathbf{I} + \beta^2 \hat{Q} \mathbf{11}^\top \tag{87}$$

$$\hat{\mathbf{Q}}^{r'r'} = \beta \hat{r} \mathbf{I} + \beta^2 \hat{R} \mathbf{11}^\top \tag{88}$$

$$\hat{\mathbf{Q}}^{rr'} = \beta \hat{c} \mathbf{I} + \beta^2 \hat{C} \mathbf{11}^\top \tag{89}$$

In order to simplify $\log \det (\lambda \mathbf{I}_{2n} + \beta \mathbf{Q})$, we note that this is a symmetric 2-by-2-block matrix, where each block commutes with all other blocks. We may then use [53]'s result to simplify.

$$\log \det (\lambda \mathbf{I}_{2n} + \beta \mathbf{Q}) = n \left[\log ((\lambda + q)(\lambda + r) - c^2) + \beta \frac{(\lambda + q)R + (\lambda + r)Q - 2cC}{(\lambda + q)(\lambda + r) - c^2} \right] + \mathcal{O}(n^2) \quad (90)$$

$$\sum_{abr_1r_2} \hat{\mathbf{Q}}_{ab}^{r_1r_2} \mathbf{Q}_{ab}^{r_1r_2} = n \left[(q\hat{q} + \beta\hat{q}Q + \beta q\hat{Q}) + (r\hat{r} + \beta\hat{r}R + \beta r\hat{R}) + 2(c\hat{c} + \beta\hat{c}C + \beta c\hat{C}) \right] + \mathcal{O}(n^2) \quad (91)$$

$$\sum_a (\mathbf{Q}_{aa}^{rr'} - \zeta^2) = n \left[\frac{1}{\beta}c + C - \zeta^2 \right] + \mathcal{O}(n^2) \quad (92)$$

$$\sum_{abr_1r_2} \hat{\mathbf{Q}}_{ab}^{r_1r_2} = \beta n [\hat{q} + \hat{r} + 2\hat{c}] + \mathcal{O}(n^2) \quad (93)$$

$$\log \det(T) = n \left[\log(\beta) + \log \det \begin{bmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rr'} \\ \mathbf{G}_{r'r} & \mathbf{G}_{r'r'} \end{bmatrix} + \beta \operatorname{tr} \left(\begin{bmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rr'} \\ \mathbf{G}_{r'r} & \mathbf{G}_{r'r'} \end{bmatrix}^{-1} \begin{bmatrix} \hat{Q}\tilde{\Sigma}_{rr} & \hat{C}\tilde{\Sigma}_{rr'} \\ \hat{C}\tilde{\Sigma}_{r'r} & \hat{R}\tilde{\Sigma}_{r'r'} \end{bmatrix} \right) \right] + \mathcal{O}(n^2) \quad (94)$$

$$\text{where } \mathbf{G}_{rr} = \mathbf{I}_{N_r} + \hat{q}\tilde{\Sigma}_{rr} \quad \mathbf{G}_{r'r'} = \mathbf{I}_{N_{r'}} + \hat{r}\tilde{\Sigma}_{r'r'} \quad \mathbf{G}_{rr'} = \hat{c}\tilde{\Sigma}_{rr'} \quad \mathbf{G}_{r'r} = \hat{c}\tilde{\Sigma}_{r'r} \quad (95)$$

$$V^\top T^{-1} V = n\beta \mathbf{w}^{*\top} \begin{bmatrix} \frac{1}{\sqrt{\nu_{rr}}}(\hat{q} + \hat{c})\mathbf{A}_r \Sigma_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r} + \hat{c})\mathbf{A}_{r'} \Sigma_s \end{bmatrix}^\top \begin{bmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rr'} \\ \mathbf{G}_{r'r} & \mathbf{G}_{r'r'} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{\nu_{rr}}}(\hat{q} + \hat{c})\mathbf{A}_r \Sigma_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r} + \hat{c})\mathbf{A}_{r'} \Sigma_s \end{bmatrix} \mathbf{w}^* + \mathcal{O}(n^2) \quad (96)$$

Collecting these terms, we may write the replicated partition function as follows:

$$\langle Z^n \rangle_{\mathcal{D}} = \exp \left(-\frac{nM}{2} \mathfrak{g} [q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C}] \right) \quad (97)$$

Where the free energy is written:

$$\mathfrak{g} [q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C}] = \quad (98)$$

$$\alpha \left[\log ((\lambda + q)(\lambda + r) - c^2) + \beta \frac{(\lambda + q)R + (\lambda + r)Q - 2cC}{(\lambda + q)(\lambda + r) - c^2} \right] \quad (99)$$

$$- \left[(q\hat{q} + \beta\hat{q}Q + \beta q\hat{Q}) + (r\hat{r} + \beta\hat{r}R + \beta r\hat{R}) + 2(c\hat{c} + \beta\hat{c}C + \beta c\hat{C}) \right] \quad (100)$$

$$+ J(c + \beta C - \beta \zeta^2) \quad (101)$$

$$+ \beta [\hat{q} + \hat{r} + 2\hat{c}] \left(\frac{1}{M} \mathbf{w}^{*\top} \Sigma \mathbf{w}^* + \zeta^2 \right) \quad (102)$$

$$- \frac{1}{M} \beta \mathbf{w}^{*\top} \begin{bmatrix} \frac{1}{\sqrt{\nu_{rr}}}(\hat{q} + \hat{c})\mathbf{A}_r \Sigma_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r} + \hat{c})\mathbf{A}_{r'} \Sigma_s \end{bmatrix}^\top \mathbf{G}^{-1} \begin{bmatrix} \frac{1}{\sqrt{\nu_{rr}}}(\hat{q} + \hat{c})\mathbf{A}_r \Sigma_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r} + \hat{c})\mathbf{A}_{r'} \Sigma_s \end{bmatrix} \mathbf{w}^* \quad (103)$$

$$+ \frac{1}{M} \left[\log(\beta) + \log \det \mathbf{G} + \beta \operatorname{tr} \left(\mathbf{G}^{-1} \begin{bmatrix} \hat{Q}\tilde{\Sigma}_{rr} & \hat{C}\tilde{\Sigma}_{rr'} \\ \hat{C}\tilde{\Sigma}_{r'r} & \hat{R}\tilde{\Sigma}_{r'r'} \end{bmatrix} \right) \right] \quad (104)$$

where we have defined $\mathbf{G} \equiv \begin{bmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rr'} \\ \mathbf{G}_{r'r} & \mathbf{G}_{r'r'} \end{bmatrix}$

The saddle-point equations for the replica-diagonal order parameters are:

$$\frac{\partial \mathbf{g}}{\partial Q} = 0 = \beta \frac{\alpha(\lambda + r)}{(\lambda + q)(\lambda + r) - c^2} - \beta \hat{q} \quad (105)$$

$$\frac{\partial \mathbf{g}}{\partial \hat{Q}} = 0 = -\beta q + \beta \frac{1}{M} \text{tr} \left(\mathbf{G}^{-1} \begin{bmatrix} \tilde{\Sigma}_{rr} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \quad (106)$$

$$\frac{\partial \mathbf{g}}{\partial R} = 0 = \beta \frac{\alpha(\lambda + q)}{(\lambda + q)(\lambda + r) - c^2} - \beta \hat{r} \quad (107)$$

$$\frac{\partial \mathbf{g}}{\partial \hat{R}} = 0 = -\beta r + \beta \frac{1}{M} \text{tr} \left(\mathbf{G}^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\Sigma}_{r'r'} \end{bmatrix} \right) \quad (108)$$

$$\frac{\partial \mathbf{g}}{\partial C} = 0 = -\beta \frac{2\alpha c}{(\lambda + q)(\lambda + r) - c^2} - 2\beta \hat{c} + \beta J \quad (109)$$

$$\frac{\partial \mathbf{g}}{\partial \hat{C}} = 0 = -2\beta c + \beta \frac{1}{M} \text{tr} \left(\mathbf{G}^{-1} \begin{bmatrix} \mathbf{0} & \tilde{\Sigma}_{r'r'} \\ \tilde{\Sigma}_{r'r} & \mathbf{0} \end{bmatrix} \right) \quad (110)$$

Note that when $J = 0$, the saddle point equations 109, 110 are solved by setting $c = \hat{c} = 0$, and in this case the remaining saddle-point equations decouple over the readouts (as expected for independently trained ensemble members) giving: For readout r :

$$0 = \frac{\alpha}{(\lambda + q)} - \hat{q} \quad (111)$$

$$0 = -q + \frac{1}{M} \text{tr} \left(\mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr} \right) \quad (112)$$

and for readout r' :

$$0 = \frac{\alpha}{(\lambda + r)} - \hat{r} \quad (113)$$

$$0 = -r + \frac{1}{M} \text{tr} \left(\mathbf{G}_{r'r'}^{-1} \tilde{\Sigma}_{r'r'} \right) \quad (114)$$

These are equivalent to the saddle-point equations for a single readout given in equation 53, 54 as expected for independently trained readouts. It is physically sensible that $c = 0$ when $J = 0$, because at zero source, there is no term in the replicated system energy function which would distinguish the overlap between two readouts from the same replica from the overlap between two readouts in separate replicas (we expect that the total overlap between readouts is non-zero, as we may still have $C > 0$).

The saddle-point equations obtained by setting the derivatives $\frac{\partial \mathbf{g}}{\partial q} = \frac{\partial \mathbf{g}}{\partial \hat{q}} = \frac{\partial \mathbf{g}}{\partial r} = \frac{\partial \mathbf{g}}{\partial \hat{r}} = 0$ will similarly decouple to recover two copies of the diagonal case 57 55. We will not re-write the expressions here as they are not necessary to determine the off-diagonal error term $E_{rr'}$.

The remaining saddle-point equations are obtained by setting $\frac{\partial \mathbf{g}}{\partial c} = \frac{\partial \mathbf{g}}{\partial \hat{c}} = 0$

$$\left. \frac{\partial \mathbf{g}}{\partial c} \right|_{c=\hat{c}=J=0} = -\frac{2\alpha\beta C}{(\lambda + q)(\lambda + r)} - 2\beta \hat{C} \Rightarrow \hat{C} = -\frac{\alpha C}{(\lambda + q)(\lambda + r)} \quad (115)$$

$$\begin{aligned}
\left. \frac{\partial \mathbf{g}}{\partial \hat{c}} \right|_{c=\hat{c}=J=0} = 0 &= -2\beta C + 2\beta \left(\frac{1}{M} \mathbf{w}^{*\top} \Sigma_s \mathbf{w}^* + \zeta^2 \right) \\
&\quad - \frac{2\beta}{M} \mathbf{w}^{*\top} \Sigma_s \left[\frac{1}{\nu_{rr}} \hat{q} \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{r} \mathbf{A}_{r'}^\top \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \right] \Sigma_s \mathbf{w}^* \\
&\quad + \frac{2\beta \hat{q} \hat{r}}{M} \frac{1}{\sqrt{\nu_{rr} \nu_{r'r'}}} \mathbf{w}^{*\top} \Sigma_s \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \Sigma_s \mathbf{w}^* \\
&\quad - \frac{2\hat{C}\beta}{M} \text{tr} \left[\mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'r'}^{-1} \tilde{\Sigma}_{r'r} \right]
\end{aligned} \tag{116}$$

Solving equations 115 and 116 for C , we obtain:

$$\begin{aligned}
C &= \frac{1}{1-\gamma} \zeta^2 + \frac{1}{1-\gamma} \left(\frac{1}{M} \mathbf{w}^{*\top} \Sigma_s \mathbf{w}^* \right) \\
&\quad - \frac{1}{M(1-\gamma)} \mathbf{w}^{*\top} \Sigma_s \left[\frac{1}{\nu_{rr}} \hat{q} \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{r} \mathbf{A}_{r'}^\top \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \right] \Sigma_s \mathbf{w}^* \\
&\quad + \frac{1}{M(1-\gamma)} \hat{q} \hat{r} \frac{1}{\sqrt{\nu_{rr} \nu_{r'r'}}} \mathbf{w}^{*\top} \Sigma_s \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \Sigma_s \mathbf{w}^*
\end{aligned} \tag{117}$$

$$\text{where } \gamma \equiv \frac{\alpha}{(\lambda+q)(\lambda+r)} \text{tr} \left[\mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'r'}^{-1} \tilde{\Sigma}_{r'r} \right] \tag{118}$$

We can obtain the generalization error as:

$$E_{rr'} = \lim_{\beta \rightarrow \infty} \frac{1}{\beta} \frac{\partial}{\partial J} \mathbf{g} \left[q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C} \right] = \lim_{\beta \rightarrow \infty} \frac{1}{\beta} (c + \beta C - \beta \zeta^2) = C - \zeta^2 \tag{119}$$

We may then simplify the expression for the $E_{rr'}$ error as follows:

$$\begin{aligned}
E_{rr'} &= \frac{\gamma}{1-\gamma} \zeta^2 + \frac{1}{1-\gamma} \left(\frac{1}{M} \mathbf{w}^{*\top} \Sigma_s \mathbf{w}^* \right) \\
&\quad - \frac{1}{M(1-\gamma)} \mathbf{w}^{*\top} \Sigma_s \left[\frac{1}{\nu_{rr}} \hat{q} \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{r} \mathbf{A}_{r'}^\top \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \right] \Sigma_s \mathbf{w}^* \\
&\quad + \frac{1}{M(1-\gamma)} \hat{q} \hat{r} \frac{1}{\sqrt{\nu_{rr} \nu_{r'r'}}} \mathbf{w}^{*\top} \Sigma_s \mathbf{A}_r^\top \mathbf{G}_{rr}^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'r'}^{-1} \mathbf{A}_{r'} \Sigma_s \mathbf{w}^*
\end{aligned} \tag{120}$$

Re-labeling the order parameters: $\hat{q} \rightarrow \hat{q}_r$, $\hat{r} \rightarrow \hat{q}_{r'}$, $\gamma \rightarrow \gamma_{rr'}$ and $\mathbf{G}_{rr} \rightarrow \mathbf{G}_r$, we obtain the result given in the main text.

G Derivation of Proposition 1 from [31]

In the case where the data and noise covariance matrices Σ_s and Σ_0 have bounded spectra, our main result may be derived using Theorem 4.1 from Loureiro et. al. [31], with a few additional arguments to incorporate a readout noise which varies across ensemble members, and to allow for the presence of label noise in the training set but not at test time. Rather than reproducing their very lengthy statements here, we direct the reader to theorem 4.1 and corollary 4.2 in [31].

G.1 Altered Expectations at Evaluation

In [31], labels are generated at both training and test time through the same statistical process $y \sim P_y^0(y|\nu)$ where $\nu = \frac{\langle \mathbf{x} | \boldsymbol{\theta} \rangle}{\sqrt{d}}$. However, their results may be easily extended to the case where data are generated through a different statistical process for the training set and at evaluation. This will allow the application of their results to the case where label noise is present during training but not at evaluation as studied in this work. We therefore introduce separate distributions of the labels y at training and evaluation. During training, we still have $y \sim P_y^0(y|\nu)$. At evaluation, we put $y \sim P_y^g(y|\nu)$. This leads to the updated formula:

$$\mathbb{E}_{(y, \mathbf{x})} \left[\varphi \left(y, \frac{\langle \hat{\mathbf{W}} | \mathbf{U} \rangle}{\sqrt{p}} \right) \right] \xrightarrow{P} \mathbb{E}_{(\nu, \boldsymbol{\mu})} \left[\int dy P_y^g(y|\nu) \varphi(y, \boldsymbol{\mu}) \right] \quad (121)$$

when the expectation is over data-label pairs (y, \mathbf{x}) at *evaluation*.

G.2 Rigorous Proof of Proposition 1

We now restate the the problem setup of our main theorem using notation consistent with [31]. We study generalization error in an ensemble of estimators $\{\hat{\mathbf{w}}_k\}$, $k = 1, \dots, K$. We say $\hat{\mathbf{w}}_k \in \mathbb{R}^p$ for all $k = 1, \dots, K$. The weights are trained independently such that each minimizes a ridge loss function:

$$\begin{aligned} \hat{\mathbf{w}}_k &= \arg \min_{\mathbf{w}} \sum_{\mu=1}^n \left(y^\mu - \frac{1}{\sqrt{N_k}} \mathbf{u}_k^\top \mathbf{w} - \xi_k^\mu \right)^2 + \lambda_k |\mathbf{w}|^2 \quad k = 1, \dots, K \\ &= \arg \min_{\mathbf{w}} \sum_{\mu=1}^n \left(y^\mu - \frac{1}{\sqrt{p}} \mathbf{u}_k^\top \left(\frac{1}{\sqrt{\nu_{kk}}} \mathbf{w} \right) - \xi_k^\mu \right)^2 + \nu_{kk} \lambda_k \left| \frac{1}{\sqrt{\nu_{kk}}} \mathbf{w} \right|^2 \end{aligned} \quad (122)$$

So that our results will correspond to the results of [31] after re-scaling the regularizations $\lambda_k \rightarrow \nu_{kk} \lambda_k$. The training labels are drawn as $y^\mu \sim P_y^0(y|\frac{1}{\sqrt{p}} \boldsymbol{\theta}^\top \mathbf{x}^\mu)$ where $P_y^0(y|x) = \mathcal{N}(x, \zeta^2)$, $\xi_k^\mu \sim \mathcal{N}(0, \eta_r^2)$ and $\boldsymbol{\theta}$ represent the ground truth weights. For $k = 1, \dots, k$ we have a ‘‘measurement matrix’’ $\mathbf{A}_k \in \mathbb{R}^{N_k \times d}$, and we may set $d = p \geq \max_k N_k$ (so that $\gamma = \frac{d}{p} = 1$). The feature vectors $\mathbf{u}_k(\mathbf{x})$ are then drawn as

$$\mathbf{u}_k(\mathbf{x}) = \begin{bmatrix} \mathbf{A}_k(\mathbf{x} + \boldsymbol{\sigma}) \\ \mathbf{0}_{(p-N_k)} \end{bmatrix} = \bar{\mathbf{A}}_k(\mathbf{x} + \boldsymbol{\sigma})$$

where $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$ and $\boldsymbol{\sigma} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$. For convenience, we have defined the auxiliary matrices

$$\bar{\mathbf{A}}_k \equiv \begin{bmatrix} \mathbf{A}_k \\ \mathbf{0}_{(p-N_k) \times p} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

By constructing the feature vectors as p -dimensional vectors with only N_k non-zero components, we may apply the results of [31] while preserving structural heterogeneity (we may have $N_k \neq N_{k'}$ for $k \neq k'$) as is present in our main result. Because $[\mathbf{u}_k]_i = 0$ for all $i > N_k$, these auxiliary dimensions will not affect model predictions or generalization error. We then have $\mu_k = \frac{1}{\sqrt{p}} \mathbf{u}_k^\top \hat{\mathbf{w}}_k$, and labels are generated at evaluation according to :

$$y \sim P_y^g(y|\frac{1}{K} \sum_k \mu_k) \text{ where } P_y^g(y|x) = \mathcal{N}(x, \frac{1}{K^2} \sum_k \eta_k^2) \quad (123)$$

The generalization error may then be decomposed as $E_g = \frac{1}{K^2} \sum_{k, k'} E_{kk'}$ where $E_{kk'} = \mathbb{E}_{(y, \mathbf{x})} [(y - \mu_k)(y - \mu_{k'})]$. We will apply eq. 121 separately to calculate $E_{r, r'}$ in the cases where $r \neq r'$ and $r = r'$.

G.2.1 Off-Diagonal Terms

Eq. 121 cannot be used to calculate $E_{rr'}$ directly in the case where $r \neq r'$ due to the presence of noises ξ_k^μ which vary over elements of the ensemble (indexed by k) in the loss function (eq. 122). We may argue, however, that the presence of readout noise has no effect on the expected value of $E_{rr'}$ when $r \neq r'$, then apply eq. 121 with $\eta_k = \eta_{k'} = 0$. To see this, we examine the analytical form of the minimizer of the loss function (eq. 122):

$$\hat{\mathbf{w}}_k = \mathbf{U}_k^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} (\mathbf{y} - \boldsymbol{\xi}_k) \quad (124)$$

where we have defined the design matrices $[\mathbf{U}_k]_{i\mu} = \frac{1}{\sqrt{p}} [\mathbf{u}_k(\mathbf{x}^\mu)]_i$ and the vectors $[\mathbf{y}]_\mu = y^\mu$ and $[\boldsymbol{\xi}_k]_\mu = \xi_k^\mu$. We then have

$$\begin{aligned} E_{kk'} &= \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} \mathbf{U}_k \right) \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_{k'} \mathbf{U}_{k'}^\top + \lambda_{k'} \mathbf{I})^{-1} \mathbf{U}_{k'} \right) \\ &+ \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} \mathbf{U}_k \right) \left(\frac{1}{\sqrt{p}} \boldsymbol{\xi}_{k'}^\top (\mathbf{U}_{k'} \mathbf{U}_{k'}^\top + \lambda_{k'} \mathbf{I})^{-1} \mathbf{U}_{k'} \right) \\ &+ \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_{k'} \mathbf{U}_{k'}^\top + \lambda_{k'} \mathbf{I})^{-1} \mathbf{U}_{k'} \right) \left(\frac{1}{\sqrt{p}} \boldsymbol{\xi}_k^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} \mathbf{U}_k \right) \\ &+ \left(\frac{1}{\sqrt{p}} \boldsymbol{\xi}_k^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} \mathbf{U}_k \right) \left(\frac{1}{\sqrt{p}} \boldsymbol{\xi}_{k'}^\top (\mathbf{U}_{k'} \mathbf{U}_{k'}^\top + \lambda_{k'} \mathbf{I})^{-1} \mathbf{U}_{k'} \right) \end{aligned} \quad (125)$$

Taking the expectation value over the readout noise in the training set we get:

$$\mathbb{E}_{\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K\}} E_{kk'} = \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_k \mathbf{U}_k^\top + \lambda_k \mathbf{I})^{-1} \mathbf{U}_k \right) \left(y - \frac{1}{\sqrt{p}} \mathbf{y}^\top (\mathbf{U}_{k'} \mathbf{U}_{k'}^\top + \lambda_{k'} \mathbf{I})^{-1} \mathbf{U}_{k'} \right) \quad (k \neq k') \quad (126)$$

This is identical to $E_{kk'}$ when $\boldsymbol{\xi}_k = 0$ for all k . We may therefore calculate the off-diagonal error terms by setting $\boldsymbol{\xi}_k = 0$ in eq. 122, which gives a problem compatible with the theorem 4 of [31]. To calculate $E_{rr'}$, we appeal to theorem 4.1 and corollary 4.2 of [31]. The following objects defined in 121 are given the following definitions:

$$r(\{\hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_K\}) = \frac{1}{2} \sum_{k=1}^K \nu_{kk} \lambda_k |\hat{\mathbf{w}}_k|^2 \quad (127)$$

$$\Delta(y, \hat{y}(\mathbf{x})) = (y - \hat{y}(\mathbf{x}))^2 \quad (128)$$

$$\hat{\ell}(y, \boldsymbol{\mu}) = \frac{1}{2} |\boldsymbol{\mu} - y \mathbf{1}|^2 \quad (129)$$

$$\mathcal{Z}^0(y, \mu, \sigma) := \int \frac{P_y^0(y | x) dx}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma}} = \frac{1}{\sqrt{2\pi(\zeta^2 + \sigma)}} \exp\left(-\frac{(y-\mu)^2}{2(\zeta^2 + \sigma)}\right) \quad (130)$$

$$E_{kk'} = \mathbb{E}_{(y, \mathbf{x})} [(y - \mu_k)(y - \mu_{k'})] \rightarrow \mathbb{E}_{(\nu, \mathbf{x})} \left[\int dy P_y^g(y | \nu) (y - \mu_k)(y - \mu_{k'}) \right] \quad (131)$$

$$= \mathbb{E}_{(\nu, \mathbf{x})} [(\nu - \mu_k)(\nu - \mu_{k'})] + \frac{1}{K^2} \sum_{k=1}^K \eta_k^2 \quad (132)$$

$$= \rho - [\mathbf{m}]_k - [\mathbf{m}]_{k'} + [\mathbf{Q}]_{kk'} + \frac{1}{K^2} \sum_{k=1}^K \eta_k^2 \quad (133)$$

Where ρ , \mathbf{m} and \mathbf{Q} are defined as in [31]:

$$(\nu, \boldsymbol{\mu}) \sim \mathcal{N}\left(\mathbf{0}_{1+K}, \begin{pmatrix} \rho & \mathbf{m}^\top \\ \mathbf{m} & \mathbf{Q} \end{pmatrix}\right) \quad (134)$$

What remains is to determine the values of the order parameters ρ , \mathbf{m} and \mathbf{Q} . We next define the covariance matrices which characterize the feature maps. The feature-feature covariance is:

$$\mathbf{\Omega}_{kk'}^{ij} = \mathbb{E}_{\mathbf{x}} [(\mathbf{u}_k(\mathbf{x}))_i (\mathbf{u}_{k'}(\mathbf{x}))_j] = \mathbb{1}_{\{1 \leq i \leq N_k, 1 \leq j \leq N_{k'}\}} [\mathbf{A}_k (\mathbf{\Sigma}_s + \mathbf{\Sigma}_0) \mathbf{A}_{k'}^\top]_{ij} \quad (135)$$

$$= \sqrt{\nu_{kk} \nu_{k'k'}} \mathbb{1}_{\{1 \leq i \leq N_k, 1 \leq j \leq N_{k'}\}} \left[\tilde{\mathbf{\Sigma}}_{kk'} \right]_{ij} = [\bar{\mathbf{\Sigma}}_{kk'}]_{ij} \quad (136)$$

where we have defined $\nu_{kk} = \frac{N_k}{p}$, $\tilde{\mathbf{\Sigma}}_{kk'} = \frac{1}{\sqrt{\nu_{kk} \nu_{k'k'}}} \mathbf{A}_k (\mathbf{\Sigma}_s + \mathbf{\Sigma}_0) \mathbf{A}_{k'}^\top$, and $\bar{\mathbf{\Sigma}}_{kk'} = \sqrt{\nu_{kk} \nu_{k'k'}} \begin{bmatrix} \tilde{\mathbf{\Sigma}}_{kk'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{p \times p}$. Note that while in [31], all $\mathbf{\Omega}_{kk}$ must have strictly positive eigenvalues, their result can be easily extended to cover the case where some eigenvalues are zero by a continuity argument.

The feature-label covariance is given by:

$$\left[\hat{\Phi} \right]_k^i = \mathbb{E}_{\mathbf{x}} [\mathbf{u}_k(\mathbf{x}) \mathbf{x}^\top \boldsymbol{\theta}]_i = \mathbb{1}_{\{1 \leq i \leq N_k\}} [\mathbf{A}_k \mathbf{\Sigma}_s \boldsymbol{\theta}]_i = [\bar{\mathbf{A}}_k \mathbf{\Sigma}_s \boldsymbol{\theta}]_i \quad (137)$$

$$[\Theta]_{kk'}^{ij} = \left[\hat{\Phi} \right]_k^i \left[\hat{\Phi} \right]_{k'}^j = [\bar{\mathbf{A}}_k \mathbf{\Sigma}_s \boldsymbol{\theta}]_i [\bar{\mathbf{A}}_{k'} \mathbf{\Sigma}_s \boldsymbol{\theta}]_j \quad (138)$$

Recalling $\boldsymbol{\omega} := \mathbf{Q}^{1/2} \boldsymbol{\xi}$, we can now obtain an explicit form for the proximal \mathbf{h} :

$$\mathbf{h} := \underset{\mathbf{u}}{\operatorname{argmin}} \left[\frac{(\mathbf{u} - \boldsymbol{\omega})^\top \mathbf{V}^{-1} (\mathbf{u} - \boldsymbol{\omega})}{2} + \hat{\ell}(y, \mathbf{u}) \right] = \mathbf{V} (\mathbf{I} + \mathbf{V})^{-1} (\mathbf{V}^{-1} \mathbf{Q}^{1/2} \boldsymbol{\xi} + y \mathbf{1}) \quad (139)$$

The proximal \mathbf{G} should not arise in this special case.

Next, we will simplify the saddle-point equations. Simplifying where possible, we may write:

$$\mathbf{f} = \mathbf{V}^{-1} (\mathbf{h} - \mathbf{w}) = (\mathbf{I} + \mathbf{V})^{-1} (y \mathbf{1} - \boldsymbol{\omega}) \quad (140)$$

$$\partial_{\boldsymbol{\omega}} \mathbf{f} = -(\mathbf{I} + \mathbf{V})^{-1} \quad (141)$$

$$\rho = \mathbb{E}_{\mathbf{x}} \left[\left(\frac{1}{\sqrt{d}} \boldsymbol{\theta}^\top \mathbf{x} \right)^2 \right] = \frac{1}{d} \boldsymbol{\theta}^\top \mathbf{\Sigma}_s \boldsymbol{\theta} \quad (142)$$

$$\omega_0 \equiv \mathbf{m}^\top \mathbf{Q}^{-1/2} \boldsymbol{\xi} \quad (143)$$

$$\sigma_0 = \rho - \mathbf{m}^\top \mathbf{Q}^{-1} \mathbf{m} = \frac{1}{d} \boldsymbol{\theta}^\top \mathbf{\Sigma}_s \boldsymbol{\theta} - \mathbf{m}^\top \mathbf{Q}^{-1} \mathbf{m} \quad (144)$$

$$\hat{\mathbf{V}} = -\alpha \int dy \mathbb{E}_{\boldsymbol{\xi}} [Z^0(y, \omega_0, \sigma_0) (-(\mathbf{I} + \mathbf{V})^{-1})] = \alpha (\mathbf{I} + \mathbf{V})^{-1} \mathbb{E}_{\boldsymbol{\xi}} \left[\underbrace{\int dy Z^0(y, \omega_0, \sigma_0)}_1 \right] \quad (145)$$

$$\Rightarrow \hat{\mathbf{V}} = \alpha (\mathbf{I} + \mathbf{V})^{-1} \quad (146)$$

We can simplify the prior equation for $\hat{\mathbf{V}}$ to the following set of equations:

$$\mathbf{V}_{kk} = \frac{1}{p} \operatorname{tr} \left[\bar{\mathbf{\Sigma}}_{kk} \left[\nu_{kk} \lambda_k \mathbf{I}_p + \hat{\mathbf{V}}_{kk} \bar{\mathbf{\Sigma}}_{kk} \right]^{-1} \right] \quad k = 1, \dots, K \quad (147)$$

$$\mathbf{V}_{kk'} = \frac{1}{p} \operatorname{tr} \left[\bar{\mathbf{\Sigma}}_{kk'} \left[\hat{\mathbf{V}}_{k'k} \bar{\mathbf{\Sigma}}_{k'k} \right]^{-1} \right] \quad k' \neq k \quad (148)$$

Equations 146 and 148 are solved by setting $\hat{\mathbf{V}}_{kk'} = \mathbf{V}_{kk'} = 0$ for all $k \neq k'$ so that $\hat{\mathbf{V}}_{kk'} = \hat{V}_k \delta_{kk'}$ and $\mathbf{V}_{kk'} = V_k \delta_{kk'}$. The diagonal components then satisfy

$$\hat{V}_k = \frac{\alpha}{1 + V_k} \quad (149)$$

$$V_k = \frac{1}{p} \text{tr} \left[\bar{\boldsymbol{\Sigma}}_{kk} \left[\nu_{kk} \lambda_k \mathbf{I}_p + \hat{V}_k \bar{\boldsymbol{\Sigma}}_{kk} \right]^{-1} \right] \quad (150)$$

Which may be solved separately for each $k = 1, \dots, K$. Simplifying the remaining channel equations we have:

$$\hat{\mathbf{Q}} = \alpha(\mathbf{I} + \mathbf{V})^{-1} [(\zeta^2 + \rho) \mathbf{1}\mathbf{1}^\top - \mathbf{1}\mathbf{m}^\top - \mathbf{m}\mathbf{1}^\top + \mathbf{Q}] (\mathbf{I} + \mathbf{V})^{-1} \quad (151)$$

$$\Rightarrow \hat{Q}_{kk'} = \frac{1}{\alpha} \hat{V}_k \hat{V}_{k'} [\zeta^2 + \rho - m_k - m_{k'} + Q_{kk'}] \quad (152)$$

$$\hat{\mathbf{m}} = \alpha(\mathbf{I} + \mathbf{V})^{-1} \mathbf{1} \quad \Rightarrow \quad \hat{m}_k = \frac{\alpha}{(1 + V_k)} = \hat{V}_k \quad (153)$$

Simplifying the prior equations we obtain (through some tedious but straightforward algebra):

$$Q_{kk'} = \hat{Q}_{kk'} J_{kk'} + \hat{V}_k \hat{V}_{k'} \Lambda_{kk'} \quad (154)$$

$$m_k = \hat{V}_k R_k \quad (155)$$

Where we have defined:

$$\mathbf{G}_k \equiv \nu_{kk} \lambda_k \mathbf{I}_p + \hat{V}_k \bar{\boldsymbol{\Sigma}}_{kk} \quad (156)$$

$$J_{kk'} \equiv \frac{1}{p} \text{tr} [\bar{\boldsymbol{\Sigma}}_{kk'} \mathbf{G}_{k'}^{-1} \bar{\boldsymbol{\Sigma}}_{k'k} \mathbf{G}_k^{-1}] \quad (157)$$

$$\Lambda_{kk'} \equiv \frac{1}{p} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_s \bar{\mathbf{A}}_k^\top \mathbf{G}_k^{-1} \bar{\boldsymbol{\Sigma}}_{kk'} \mathbf{G}_{k'}^{-1} \bar{\mathbf{A}}_{k'} \boldsymbol{\Sigma}_s \boldsymbol{\theta} \quad (158)$$

$$R_k \equiv \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_s \bar{\mathbf{A}}_k^\top \mathbf{G}_k^{-1} \bar{\mathbf{A}}_k \boldsymbol{\Sigma}_s \boldsymbol{\theta} \quad (159)$$

Solving eq's 154,152 for \mathbf{Q} , we obtain:

$$Q_{kk'} = \frac{\gamma_{kk'}}{1 - \gamma_{kk'}} (\zeta^2 + \rho - m_k - m_{k'}) + \frac{\hat{V}_k \hat{V}_{k'}}{1 - \gamma_{kk'}} \Lambda_{kk'} \quad (k \neq k') \quad (160)$$

Combining these results, we arrive at a formula for $E_{kk'}$:

$$E_{kk'} = \frac{1}{1 - \gamma_{kk'}} \left(\gamma_{kk'} \zeta^2 + \rho - \hat{V}_k R_k - \hat{V}_{k'} R_{k'} + \hat{V}_k \hat{V}_{k'} \Lambda_{kk'} \right) + \frac{1}{K^2} \sum_k \eta_k^2 \quad (k \neq k') \quad (161)$$

$$\text{where } \gamma_{kk'} \equiv \frac{1}{\alpha} \hat{V}_k \hat{V}_{k'} J_{kk'} \quad (162)$$

Where the order parameters \hat{V}_k , $k = 1, \dots, K$ satisfy the fixed-point equations given by eq's 149, 150.

G.2.2 Diagonal Terms

To calculate E_{rr} , we cannot ignore the presence of readout noise in the loss function. However, as the loss function is separable over the readouts k , we may calculate E_{rr} using the results of [31] in

the special case where $K = 1$, incorporating the readout noise into the label noise. Concretely, we may calculate E_{rr} as the error of a single linear predictor under the same setup as the off-diagonal terms except that in the training set $y^\mu \sim P_y^0(y|\frac{1}{\sqrt{\rho}}\boldsymbol{\theta}^\top \mathbf{x}^\mu)$ where $P_y^0(y|x) = \mathcal{N}(x, \zeta^2 + \eta_k^2)$. This may be recovered from the calculation for the off-diagonal terms by setting $k = k'$ and re-scaling $\zeta^2 \rightarrow \zeta^2 + \eta_k^2$, giving:

$$E_{kk} = \frac{1}{1 - \gamma_{kk}} \left(\gamma_{kk}(\zeta^2 + \eta_r^2) + \rho - 2\hat{V}_k R_k + \hat{V}_k^2 \Lambda_{kk} \right) + \frac{1}{K^2} \sum_k \eta_k^2 \quad (k \neq k') \quad (163)$$

where the definitions of γ_{kk} , ρ , R_k , and Λ_{kk} can be inherited from the off-diagonal case, as well as the saddle-point equations 149, 150.

G.2.3 Full Error

The results obtained here are equivalent to the results of our main theorem, up to a reshuffling of additive constants η_k^2 among the error terms, and a trivial re-scaling of the order parameters as follows: $V_k \rightarrow \frac{1}{\lambda_k} V_k$, $\hat{V}_k \rightarrow \lambda_k \hat{V}_k$

H Equicorrelated Data Model

To gain an intuition for the joint effects of correlated data, subsampling, ensembling, feature noise, and readout noise, we simplify the formulas for the generalization error in the following special case:

$$\boldsymbol{\Sigma}_s = s \left[(1 - c) \mathbf{I}_M + c \mathbf{1}_M \mathbf{1}_M^\top \right] \quad (164)$$

$$\boldsymbol{\Sigma}_0 = \omega^2 \mathbf{I}_M \quad (165)$$

Here s is a parameter which sets the overall scale of the data and $c \in [0, 1]$ tunes the correlation structure in the data and ω^2 sets the scale of an isotropic feature noise. We consider an ensemble of k readouts, each of which sees a subset of the features. Due to the isotropic nature of the equicorrelated data model and the pairwise decomposition of the generalization error, we expect that the generalization error will depend on the partition of features among the readout neurons through only:

- The number of features sampled by each readout: $N_r \equiv \nu_{rr} M$, for $r = 1, \dots, k$
- The number of features jointly sampled by each pair of readouts $n_{rr'} \equiv \nu_{rr'} M$ for $r, r' \in \{1, \dots, k\}$

Here, we have introduced the subsampling fractions $\nu_{rr} = \frac{N_r}{M}$ and the overlap fractions $\nu_{rr'} = \frac{n_{rr'}}{M}$

We will average the generalization error over readout weights drawn randomly from the space perpendicular to $\mathbf{1}_M$, with an added spike along the direction of $\mathbf{1}_M$:

$$\mathbf{w}^* = \sqrt{1 - \rho^2} \mathbb{P}_\perp \mathbf{w}_0^* + \rho \mathbf{1}_M \quad (166)$$

$$\mathbf{w}_0^* \sim \mathcal{N}(0, \mathbf{I}_M) \quad (167)$$

The projection matrix may be written $\mathbb{P}_\perp = \mathbf{I}_M - \frac{1}{N} \mathbf{1}_M \mathbf{1}_M^\top$. The two components of the ground truth weights will yield independent contributions to the generalization error in the sense that

$$\langle E_{rr'} \rangle = (1 - \rho^2) E_{rr'}(\rho = 0) + \rho^2 E_{rr'}(\rho = 1) \quad (168)$$

Calculating E_{rr} and $E_{rr'}$ is an exercise in linear algebra which is straightforward but tedious. To assist with the tedious algebra, we wrote a Mathematica package which can handle multiplication, addition, and inversion of matrices of symbolic dimension of the specific form encountered in this problem. This form consists of block matrices, where the blocks may be written as $a \delta_{MN} \mathbf{I}_M + b \mathbf{1}_M \mathbf{1}_N^\top$, where a, b are scalars and δ_{MN} ensures that there is only a diagonal component for square blocks (when $M = N$). This package is included as supplemental material to this publication.

H.1 Diagonal Terms and Saddle-Point Equations

Here, we solve for the dominant values of q_r and \hat{q}_r and simplify the expressions for E_{rr} in the case of equicorrelated features described above. In this isotropic setting, E_{rr}, q_r, \hat{q}_r will depend on the subsampling only through $N_r = \nu_{rr}M$. We may then write, without loss of generality $\mathbf{A}_r = (\mathbf{I}_{N_r} \quad \mathbf{0}) \in \mathbb{R}^{N_r \times M}$ where $\mathbf{0}$ denotes a matrix of all zeros, of the appropriate dimensionality.

We start by simplifying the saddle-point equations 53,54. Expanding $\frac{1}{M} \text{tr} \left(\mathbf{G}_r^{-1} \tilde{\Sigma}_{rr} \right)$ and keeping only leading order terms, the saddle-point equations for q_r and \hat{q}_r reduce to:

$$q_r = \frac{\nu_{rr} (s(1-c) + \omega^2)}{\hat{q}_r (s(1-c) + \omega^2) + \nu_r} \quad (169)$$

$$\hat{q}_r = \frac{\alpha}{\lambda + q_r} \quad (170)$$

Defining $a \equiv s(1-c) + \omega^2$ and solving this system of equations, we find:

$$q_r = \frac{\sqrt{a^2 \alpha^2 + 2a\alpha(\lambda - a)\nu_r + (a + \lambda)^2 \nu_r^2} - a\alpha + (a - \lambda)\nu_r}{2\nu_r} \quad (171)$$

$$\hat{q}_r = \frac{\sqrt{a^2 \alpha^2 + 2a\alpha(\lambda - a)\nu_r + (a + \lambda)^2 \nu_r^2} + a\alpha - (a + \lambda)\nu_r}{2a\lambda} \quad (172)$$

We have selected the solution with $q_r > 0$ because self-overlaps must be at least as large as overlaps between different replicas. This solution to the saddle-point equations can be applied to each of the k readouts.

Next, we calculate E_{rr} . Expanding $\gamma_{rr} \equiv \frac{\alpha}{M\kappa^2} \text{tr} \left[(\mathbf{G}^{-1} \tilde{\Sigma})^2 \right]$ to leading order in M , we find:

$$\gamma_{rr} = \frac{a^2 \alpha \nu_r}{(\lambda + q_r)^2 (a\hat{q}_r + \nu_r)^2} \quad (173)$$

$$\begin{aligned} \langle E_{rr} \rangle_{\mathcal{D}, \mathbf{w}^*} (\rho = 0) &= \frac{1}{1 - \gamma_{rr}} \frac{1}{M} \text{tr} \left[\mathbb{P}_\perp \left(\Sigma_s - \frac{2}{\nu_{rr}} \hat{q}_r \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \mathbf{A}_r \Sigma_s + \frac{1}{\nu_{rr}} \hat{q}_r^2 \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \tilde{\Sigma} \mathbf{G}_r^{-1} \mathbf{A}_r \Sigma_s \right) \mathbb{P}_\perp \right] \\ &\quad + \frac{\gamma_{rr}}{1 - \gamma_{rr}} \zeta^2 + \eta_r^2, \end{aligned} \quad (174)$$

$$\begin{aligned} \langle E_{rr} \rangle_{\mathcal{D}, \mathbf{w}^*} (\rho = 1) &= \frac{1}{1 - \gamma_{rr}} \frac{1}{M} \mathbf{1}_M^\top \left[\Sigma_s - \frac{2}{\nu_{rr}} \hat{q}_r \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \mathbf{A}_r \Sigma_s + \frac{1}{\nu_{rr}} \hat{q}_r^2 \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \tilde{\Sigma} \mathbf{G}_r^{-1} \mathbf{A}_r \Sigma_s \right] \mathbf{1}_m \\ &\quad + \frac{\gamma_{rr}}{1 - \gamma_{rr}} \zeta^2 + \eta_r^2, \end{aligned} \quad (175)$$

With the aid of our custom Mathematica package, we calculate the traces and contractions in these expressions and expand them to leading order in M , finding:

$$\langle E_{rr} \rangle_{\mathcal{D}, \mathbf{w}^*} (\rho = 0) = \frac{1}{1 - \gamma_{rr}} \left(s(1-c) \left(1 - \frac{(1-c)s\hat{q}_r\nu_r (\hat{q}_r (s(1-c) + \omega^2) + 2\nu_r)}{(\hat{q}_r (s(1-c) + \omega^2) + \nu_r)^2} \right) \right) + \frac{\gamma_{rr}\zeta^2 + \eta_r^2}{1 - \gamma_{rr}} \quad (176)$$

$$\langle E_{rr} \rangle_{\mathcal{D}, \mathbf{w}^*} (\rho = 1) = \frac{1}{1 - \gamma_{rr}} \left(\frac{s(1-c)(1 - \nu_{rr}) + \omega^2}{\nu_{rr}} \right) + \frac{\gamma_{rr}\zeta^2 + \eta_r^2}{1 - \gamma_{rr}} \quad (177)$$

In the ‘‘ridgeless’’ limit where $\lambda \rightarrow 0$, we obtain:

$$\gamma_{rr} = \frac{4\alpha\nu_{rr}}{(\alpha + \nu_{rr} + |\alpha - \nu_{rr}|)^2} \quad (178)$$

$$\langle E_{rr}(\rho = 0) \rangle_{\mathcal{D}, \mathbf{w}^*} = \left\{ \begin{array}{ll} \frac{s(1-c)\nu_{rr}}{\nu_{rr}-\alpha} \left(1 + \frac{s\alpha(1-c)(\alpha-2\nu_{rr})}{\nu_{rr}[s(1-c)+\omega^2]} \right) + \frac{\alpha\zeta^2 + \nu_{rr}\eta_r^2}{\nu_{rr}-\alpha}, & \text{if } \alpha < \nu_{rr} \\ \frac{s(1-c)\alpha}{\alpha-\nu_{rr}} \left(1 - \frac{s(1-c)\nu_{rr}}{s(1-c)+\omega^2} \right) + \frac{\nu_{rr}\zeta^2 + \alpha\eta_r^2}{\alpha-\nu_{rr}}, & \text{if } \alpha > \nu_{rr} \end{array} \right\} \quad (\lambda \rightarrow 0) \quad (179)$$

$$\langle E_{rr}(\rho = 1) \rangle_{\mathcal{D}, \mathbf{w}^*} = \left\{ \begin{array}{ll} \frac{\nu_{rr}}{\nu_{rr}-\alpha} \left(\frac{s(1-c)(1-\nu_{rr})+\omega^2}{\nu_{rr}} \right) + \frac{\alpha\zeta^2 + \nu_{rr}\eta_r^2}{\nu_{rr}-\alpha}, & \text{if } \alpha < \nu_{rr} \\ \frac{\alpha}{\alpha-\nu_{rr}} \left(\frac{s(1-c)(1-\nu_{rr})+\omega^2}{\nu_{rr}} \right) + \frac{\nu_{rr}\zeta^2 + \alpha\eta_r^2}{\alpha-\nu_{rr}}, & \text{if } \alpha > \nu_{rr} \end{array} \right\} \quad (\lambda \rightarrow 0) \quad (180)$$

H.2 Off-Diagonal Terms

In this section, we calculate the off-diagonal error terms $E_{rr'}$ for $r \neq r'$, again making use of our custom Mathematica package to simplify contractions of block matrices of the prescribed form. By the isotropic nature of the equicorrelated data model, $E_{rr'}$ can only depend on the subsampling scheme through ν_{rr} , $\nu_{r'r'}$, and $\nu_{rr'}$. We can thus, without loss of generality, write:

$$\mathbf{A}_r = \begin{pmatrix} \mathbf{I}_{n_r \times n_r} & \mathbf{0}_{n_r \times n_{r'}} & \mathbf{0}_{n_r \times n_s} & \mathbf{0}_{n_r \times l} \\ \mathbf{0}_{n_s \times n_r} & \mathbf{0}_{n_s \times n_{r'}} & \mathbf{I}_{n_s \times n_s} & \mathbf{0}_{n_s \times l} \end{pmatrix} \in \mathbb{R}^{N_r \times M} \quad (181)$$

$$\mathbf{A}_{r'} = \begin{pmatrix} \mathbf{0}_{n_{r'} \times n_r} & \mathbf{I}_{n_{r'} \times n_{r'}} & \mathbf{0}_{n_{r'} \times n_s} & \mathbf{0}_{n_{r'} \times l} \\ \mathbf{0}_{n_s \times n_r} & \mathbf{0}_{n_s \times n_{r'}} & \mathbf{I}_{n_s \times n_s} & \mathbf{0}_{n_s \times l} \end{pmatrix} \in \mathbb{R}^{N_{r'} \times M} \quad (182)$$

where we have defined n_s to be the number of features shared between the readouts, $n_r = N_r - n_s$ and $n_{r'} = N_{r'} - n_s$ and the count of remaining features $l = M - n_r - n_{r'} - n_s$.

Then, to leading order in M , we find:

$$\gamma_{rr'} = \frac{\alpha\nu_{rr'}(s(1-c) + \omega^2)^2}{(\lambda + q_r)(\lambda + q_{r'}) (\nu_{rr} + (s(1-c) + \omega^2)\hat{q}_r) (\nu_{r'r'} + (s(1-c) + \omega^2)\hat{q}_{r'})} \quad (183)$$

Averaging $E_{rr'}$ over $\mathbf{w}_0^* \sim \mathcal{N}(0, \mathbf{I}_M)$, we get:

$$\begin{aligned} \langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 0) &= \frac{\gamma_{rr'}}{1 - \gamma_{rr'}} \zeta^2 + \frac{1}{1 - \gamma_{rr'}} \left(\frac{1}{M} \text{tr} [\mathbb{P}_\perp \Sigma_s \mathbb{P}_\perp] \right) \\ &\quad - \frac{1}{M(1 - \gamma_{rr'})} \text{tr} \left[\mathbb{P}_\perp \Sigma_s \left(\frac{1}{\nu_{rr}} \hat{q}_r \mathbf{A}_r^\top \mathbf{G}_r^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{q}_{r'} \mathbf{A}_{r'}^\top \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \right) \Sigma_s \mathbb{P}_\perp \right] \\ &\quad + \frac{\hat{q}_r \hat{q}_{r'}}{M(1 - \gamma_{rr'})} \frac{1}{\sqrt{\nu_{rr} \nu_{r'r'}}} \text{tr} \left[\mathbb{P}_\perp \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \Sigma_s \mathbb{P}_\perp \right], \end{aligned} \quad (184)$$

$$\begin{aligned} \langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 1) &= \frac{\gamma_{rr'}}{1 - \gamma_{rr'}} \zeta^2 + \frac{1}{M(1 - \gamma_{rr'})} (\mathbf{1}_M^\top \Sigma_s \mathbf{1}_M) \\ &\quad - \frac{1}{M(1 - \gamma_{rr'})} \mathbf{1}_M^\top \Sigma_s \left(\frac{1}{\nu_{rr}} \hat{q}_r \mathbf{A}_r^\top \mathbf{G}_r^{-1} \mathbf{A}_r + \frac{1}{\nu_{r'r'}} \hat{q}_{r'} \mathbf{A}_{r'}^\top \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \right) \Sigma_s \mathbf{1}_M^\top \\ &\quad + \frac{\hat{q}_r \hat{q}_{r'}}{M(1 - \gamma_{rr'})} \frac{1}{\sqrt{\nu_{rr} \nu_{r'r'}}} \mathbf{1}_M^\top \Sigma_s \mathbf{A}_r^\top \mathbf{G}_r^{-1} \tilde{\Sigma}_{rr'} \mathbf{G}_{r'}^{-1} \mathbf{A}_{r'} \Sigma_s \mathbf{1}_M \end{aligned} \quad (185)$$

Calculating these contractions and traces and expanding to leading order in M , we get:

$$\begin{aligned} \langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 0) &= \frac{s(1-c)}{1-\gamma_{rr'}} \left(1 - \frac{s(1-c)\nu_{rr}\hat{q}_r}{\nu_{rr} + (s(1-c) + \omega^2)\hat{q}_r} - \frac{s(1-c)\nu_{r'r'}\hat{q}_{r'}}{\nu_{r'r'} + (s(1-c) + \omega^2)\hat{q}_{r'}} \right. \\ &\quad \left. + \frac{s(1-c)(s(1-c) + \omega^2)\nu_{rr'}\hat{q}_r\hat{q}_{r'}}{(\nu_{rr} + (s(1-c) + \omega^2)\hat{q}_r)(\nu_{r'r'} + (s(1-c) + \omega^2)\hat{q}_{r'})} \right) \\ &\quad + \frac{\gamma_{rr'}}{1-\gamma_{rr'}} \zeta^2 \end{aligned} \quad (186)$$

$$\langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 1) = \frac{1}{1-\gamma_{rr'}} \left(\frac{s(1-c)(\nu_{rr'} - \nu_{rr}\nu_{r'r'}) + \omega^2\nu_{rr'}}{\nu_{rr}\nu_{r'r'}} \right) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}} \zeta^2 \quad (187)$$

Taking $\lambda \rightarrow 0$ we get the ridgeless limit:

$$\gamma_{rr'} \rightarrow \frac{4\alpha\nu_{rr'}}{(\alpha + \nu_{rr} + |\alpha - \nu_{rr}|)(\alpha + \nu_{r'r'} + |\alpha - \nu_{r'r'}|)} \quad (\lambda \rightarrow 0) \quad (188)$$

$$\langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 0) = \frac{1}{1-\gamma_{rr'}} F_0(\alpha) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}} \zeta^2 \quad (r \neq r') \quad (189)$$

where

$$F_0(\alpha) \equiv \left\{ \begin{array}{ll} \frac{(c-1)s(\nu_r\nu_{r'}((2\alpha-1)(c-1)s+\omega^2)-\alpha^2(c-1)s\nu_{rr'}))}{\nu_r((c-1)s-\omega^2)\nu_{r'}}, & \text{if } \alpha \leq \nu_{rr} \leq \nu_{r'r'} \\ \frac{(c-1)s(\nu_{r'}((c-1)s\nu_r+(\alpha-1)(c-1)s+\omega^2)-\alpha(c-1)s\nu_{rr'})}{((c-1)s-\omega^2)\nu_{r'}}, & \text{if } \nu_{rr} \leq \alpha \leq \nu_{r'r'} \\ \frac{(c-1)s((c-1)s\nu_{r'}-cs\nu_{rr'}+(c-1)s\nu_r-cs+s\nu_{r'r'}+s+\omega^2)}{(c-1)s-\omega^2}, & \text{if } \nu_{rr} \leq \nu_{r'r'} \leq \alpha \end{array} \right\} \quad (190)$$

$$\langle E_{rr'}(\mathcal{D}) \rangle_{\mathcal{D}, \mathbf{w}^*}(\rho = 1) = \frac{1}{1-\gamma_{rr'}} \left(\frac{s(1-c)(\nu_{rr'} - \nu_{rr}\nu_{r'r'}) + \omega^2\nu_{rr'}}{\nu_{rr}\nu_{r'r'}} \right) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}} \zeta^2 \quad (\lambda \rightarrow 0) \quad (191)$$

H.3 Optimal Regularization

Here, we derive the ‘‘locally’’ optimal regularization which minimizes the prediction error of the ensemble members independently. Equivalently we derive the optimal regularization for a single readout ($k = 1$) or the regularization which minimizes the diagonal terms E_{rr} of the generalization error. By differentiating E_{rr} with respect to the regularization λ , one can show that for the optimal regularization, we will have

$$\frac{1}{(1-\gamma_{rr})} \frac{dS_r}{d\lambda} \left[\frac{\alpha^2\nu_{rr}}{\alpha(1-\gamma_{rr})} S_r ((1-\rho^2)I_{rr}^0 + \rho^2 I_{rr}^1 + \zeta^2 + \eta_r^2) + (1-\rho^2)s^2(1-c^2)\nu_{rr}(aS_r - 1) \right] = 0 \quad (192)$$

It is easy to show that $\frac{dS_r}{d\lambda}$ cannot be equal to 0. Setting the term in brackets to zero and solving for λ , we find with the aid of the computer algebra system Mathematica [54]:

$$\lambda^* = a \left(\frac{a(\nu_{rr}(\zeta^2 + \eta^2) + a\rho^2 + s(1-c)\nu_{rr}(1-2\rho^2))}{(1-c)^2\nu_{rr}^2(1-\rho^2)s^2} - 1 \right), \quad (0 < c \leq 1) \quad (193)$$

In the limiting case of isotropic data ($c = 0$), the optimal regularization reduces to:

$$\lambda^* = \frac{1}{\nu_{rr}}(\zeta^2 + \eta^2) + s \left(\frac{1}{\nu_{rr}} - 1 \right), \quad (c = 0) \quad (194)$$

We note that while generalization error is discontinuous at $c = 0$, this result can be quickly obtained from eq. 193 by noticing that the formula for the generalization error with $0 < c \leq 1$ reduces to the formula for generalization error with $c = 0$ when we set $\rho = c = 0$ (when $c = 0$, the generalization error does not depend on ρ because there is no special direction in the data).

H.4 Homogeneous Ensembling with Resource Constraints

We make further simplifications in the special case where $\lambda_r = \lambda$, $\eta_r = \eta$, $\nu_{rr} = \frac{1}{k}$ for all $r = 1, \dots, k$. For simplicity, we also set $\nu_{rr'} = 0$ for all $r \neq r'$ so that ensemble members sample mutually exclusive subsets of the features. We will consider both the ridgeless limit $\lambda \rightarrow 0$ and the case of “locally optimal” regularization $\lambda = \lambda^*$ (see eq. 193). Concretely, we show here that under these special cases, generalization error has the forms given in eqs. 25, 26.

We start by analyzing the saddle-point equations 111, 112. Note that in this special case the saddle point equations will be identical for all $r = 1, \dots, k$. We will therefore suppress the r index. Solving this quadratic system of equations explicitly, we encounter the following radical, which we assign variable name x :

$$x \equiv \sqrt{(a\alpha - a\nu + \lambda\nu)^2 + 4a\lambda\nu^2} \quad (195)$$

In order to simplify this radical to begin extracting the factor of $s(1 - c)$ that appears in eqs. 25, 26, we define a reduced regularization:

$$\Lambda \equiv \frac{\lambda}{s(1 - c)} \quad (196)$$

And substitute Λ , H , W , and Z (recall eq. 24) for λ , η , ω , and ζ , giving

$$x = s(1 - c)\mathcal{X}(\alpha, \nu, W, \Lambda) \quad (197)$$

$$\text{where } \mathcal{X}(\alpha, \nu, W, \Lambda) \equiv \sqrt{(\alpha(W + 1) + \nu(\Lambda - W - 1))^2 + 4\Lambda\nu^2(W + 1)} \quad (198)$$

making the same substitutions, we can then write

$$\hat{q} = [s(1 - c)]^{-1} \mathcal{Q} \quad (199)$$

$$\text{where } \mathcal{Q} \equiv \frac{2\alpha\nu}{(-\alpha(W + 1) + \nu(\Lambda + W + 1) + X)} \quad (200)$$

Continuing to make substitutions in this manner, we arrive at:

$$S = [s(1 - c)]^{-1} \mathcal{S} \quad (201)$$

$$\text{where } \mathcal{S} \equiv \frac{\mathcal{Q}}{\nu + \mathcal{Q}(1 + W)} \quad (202)$$

Finally, we may express the generalization error in a reasonably compact form. In the special case at hand, the generalization error is written:

$$E_k = \frac{1}{k} E_{rr}(\nu_{rr} = \frac{1}{k}) + \frac{k-1}{k} E_{rr'}(\nu_{rr'} = \frac{1}{k} \delta_{rr'}) \quad (203)$$

Substituting 201 for S in eq. 16, and making further substitutions as necessary, we arrive at:

$$E_{rr}(\nu_{rr} = \frac{1}{k}) = s(1 - c) \mathcal{E}_{rr}(k, \alpha, \rho, \Lambda, H, W, Z) \quad (204)$$

where

$$\mathcal{E}_{rr} = \frac{N_1 + N_2}{D} \quad (205)$$

$$N_1 = \alpha(H + 1)k + \mathcal{S}(\mathcal{S}(W + 1)(\alpha + WZ + Z) - 2\alpha) \quad (206)$$

$$N_2 = \alpha\rho^2(k - \mathcal{S})(W(k + \mathcal{S}) + k + \mathcal{S} - 2) \quad (207)$$

$$D = \alpha k - \mathcal{S}^2(W + 1)^2 \quad (208)$$

We also obtain

$$E_{rr'}(\nu_{rr'} = \frac{1}{k}\delta_{rr'}) = s(1 - c)\mathcal{E}_{rr'}(k, \alpha, \rho, \Lambda, H, W, Z) \quad (209)$$

where

$$\mathcal{E}_{rr'} = \frac{2(\rho^2 - 1)\mathcal{S}}{k} - 2\rho^2 + 1 \quad (210)$$

Combining these, we obtain the error of the ensemble as:

$$E_k = s(1 - c)\mathcal{E}(k, \alpha, \rho, \Lambda, H, W, Z) \quad (211)$$

where:

$$\mathcal{E} = \frac{1}{k}\mathcal{E}_{rr} + \frac{k - 1}{k}\mathcal{E}_{rr'} \quad (212)$$

These result are derived in the Mathematica notebook titled “EquiCorrParameterReduction.nb” included with the available code. It follows from eq. 211 that when $\lambda = 0$ error can be written in the form of eq. 22. It follows from equations 211, 209 that at “locally optimal” regularization λ^* , error can be written in the form of eq. 23. To see this, note that the reduced regularization Λ^* which minimizes E_{rr} will only depend on the other arguments of \mathcal{E}_{rr} . Full expressions for the generalization error in the case $\lambda \rightarrow 0$ and $\lambda = \lambda^*$ can be found in the mathematica notebooks “EquiCorrPhaseDiagram_ZeroReg.nb” and “EquiCorrPhaseDiagram_LocalReg.nb” included with the available code. These equations are long and difficult to interperet – nor are they directly used in our code – and so are omitted here.

H.4.1 The Intermediate to Noise-Dominated Transition

The transition between the intermediate regime where $1 < k^* < \infty$ and the noise-dominated regime where $k^* = \infty$ can be studied analytically relatively painlessly in the ridgeless limit $\lambda \rightarrow 0$. The strategy we employ to determine this phase boundary is to examine the large- k asymptotic expansion of E_k to determine whether the error approaches its asymptotic value from below or above. If E_k approaches E_∞ from below, then k^* must be finite. If E_k approaches E_∞ from above, then k^* may be infinite – however, there is still the possibility of $k^* < \infty$ if E_k is non-monotonic in k . In practice, we check the values of E_k for $k = 1, 2, \dots, 100$ and $k = \infty$.

Setting $\Lambda = 0$ and expanding E_k around $k = \infty$, we find:

$$\frac{E_k}{s(1 - c)} = 1 - \rho^2 + \rho^2 W + \frac{((1 + W)^2 \rho^2 + \alpha(-2 + H + HW + 2\rho^2))}{(1 + W)\alpha k} + \mathcal{O}\left(\frac{1}{k^2}\right) \quad (213)$$

Setting the coefficient of k^{-1} to zero, we find the phase boundary as:

$$\alpha = \frac{(1 + W)^2 \rho^2}{2(1 - \rho^2) - H(1 + W)} \quad (214)$$

This equation explains the shapes of the boundaries between the intermediate and noise-dominated regions in the phase diagrams of fig. 4 with $\lambda = 0$ (see black dotted lines in panels b, c, d).

An analytical formula for the boundary between the intermediate and noise-dominated regimes at locally optimal regularization $\lambda = \lambda^*$ cannot be easily obtained. To understand why, we can assess the large- k asymptotic expansion of the generalization error at locally optimal regularization:

$$E_k(\lambda = \lambda^*) = s(1 - c) \left[1 - \rho^2 + \rho^2 W + \frac{H}{k} \right] + \mathcal{O}\left(\frac{1}{k^2}\right) \quad (215)$$

This shows that at large k , when $H > 0$, error always approaches its asymptotic value from above (recall that when $H = 0$ we always have $k^* = 1$, so that there is no phase boundary unless $H > 0$). Thus, determining the phase boundary requires determining the value of k which minimizes E_k , which is not analytically tractable.

H.5 Infinite Data Limit

In this section we consider the behavior of generalization error in the equicorrelated data model as $\alpha \rightarrow \infty$ while keeping the $\lambda \sim \mathcal{O}(1)$. For simplicity, we assume $\nu_{rr'} = 0$ for $r \neq r'$, isotropic features ($c = 0$), no feature noise ($\omega = 0$) and uniform readout noise $\eta_r = \eta$ as in main text Fig. 3. This limit corresponds to data-rich learning, where the number of training examples is large relative to the number of model parameters. In this case, the saddle point equations reduce to:

$$\hat{q}_r \rightarrow \frac{\alpha}{\lambda} \quad (216)$$

$$q_r \rightarrow \frac{\nu_{rr} \lambda}{\alpha} \quad (217)$$

In this limit, we find that $\gamma_{rr'} \rightarrow 0$. Using this, we can simplify the generalization error as follows:

$$E_g = \frac{1}{k^2} \sum_{rr'=1}^k E_{rr'} = s \left[1 - \left(2 - \frac{1}{k}\right) \left(\frac{1}{k} \sum_{r=1}^k \nu_{rr}\right) \right] + \frac{\eta^2}{k} \quad (218)$$

Interestingly, we find that the readout error in this case depends on the subsampling fractions ν_{rr} only through their mean. Therefore, with infinite data, there will be no distinction between homogeneous and heterogeneous subsampling.

I Theoretical Learning Curves and Optimal Subsampling Phase Diagrams

Here, we provide additional learning curves and phase diagrams of k^* such as those in Fig. 4a,b,c,d, exploring more parameter values for the task-model alignment ρ and the Reduced noises H , W , and Z . Generalization errors are calculated for a homogeneous ensemble of k linear regressors, as described in sections 5, H.4. We also show diagrams of generalization error E_{k^*} .

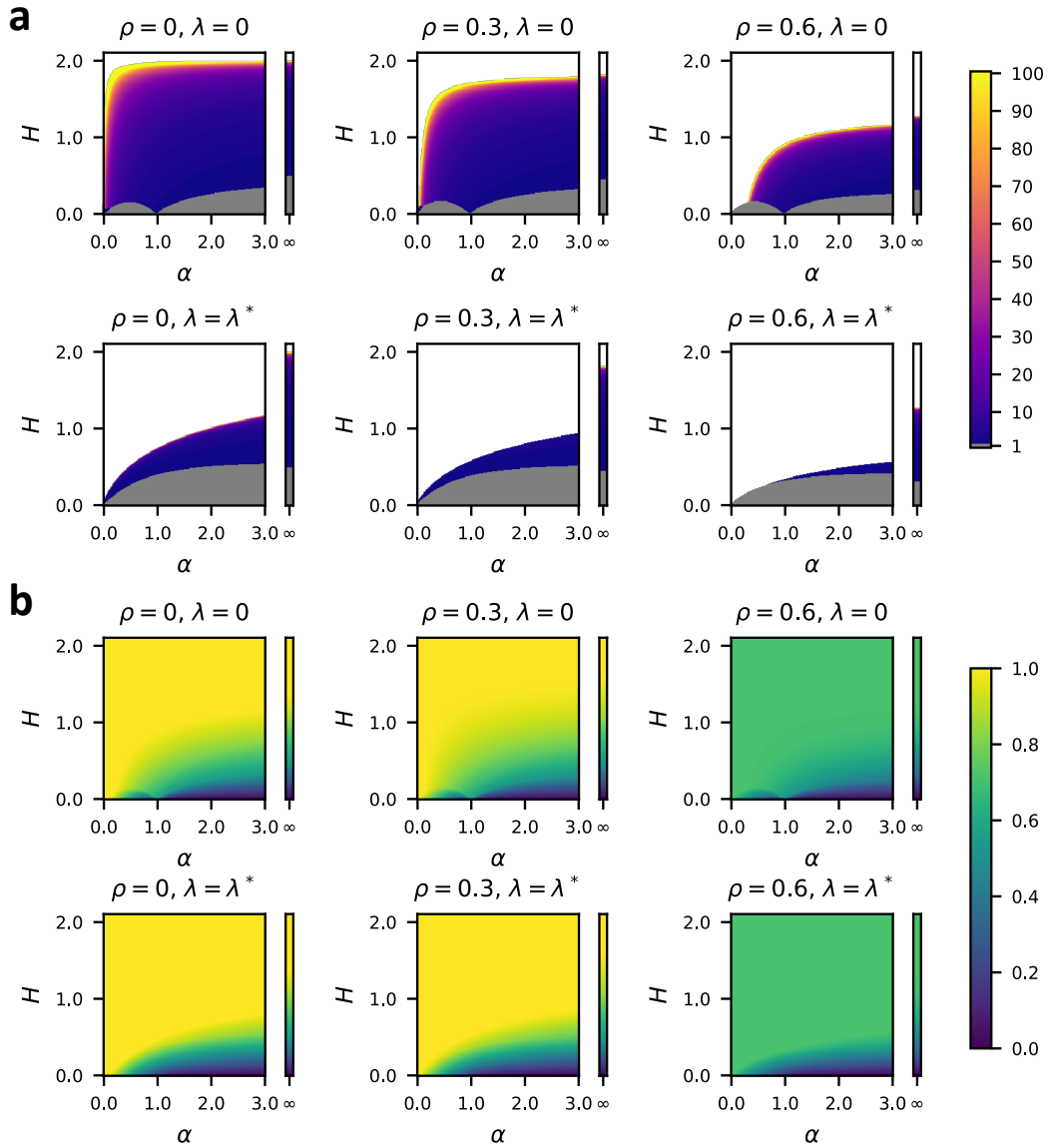


Figure S6: (a) Optimal ensemble size k^* (eqs. 25, 26) in the parameter space of sample size α and reduced readout noise scale H setting $W = Z = 0$. Grey indicates $k^* = 1$ and white indicates $k^* \rightarrow \infty$, with intermediate values given by the colorbar. Appended vertical bars show $\alpha \rightarrow \infty$. ρ and λ indicated in panel titles. $\lambda = \lambda^*$ denotes the ‘locally optimal’ regularization (see section H.3) (b) Optimal generalization error E_{k^*} for the same parameter values in (a).

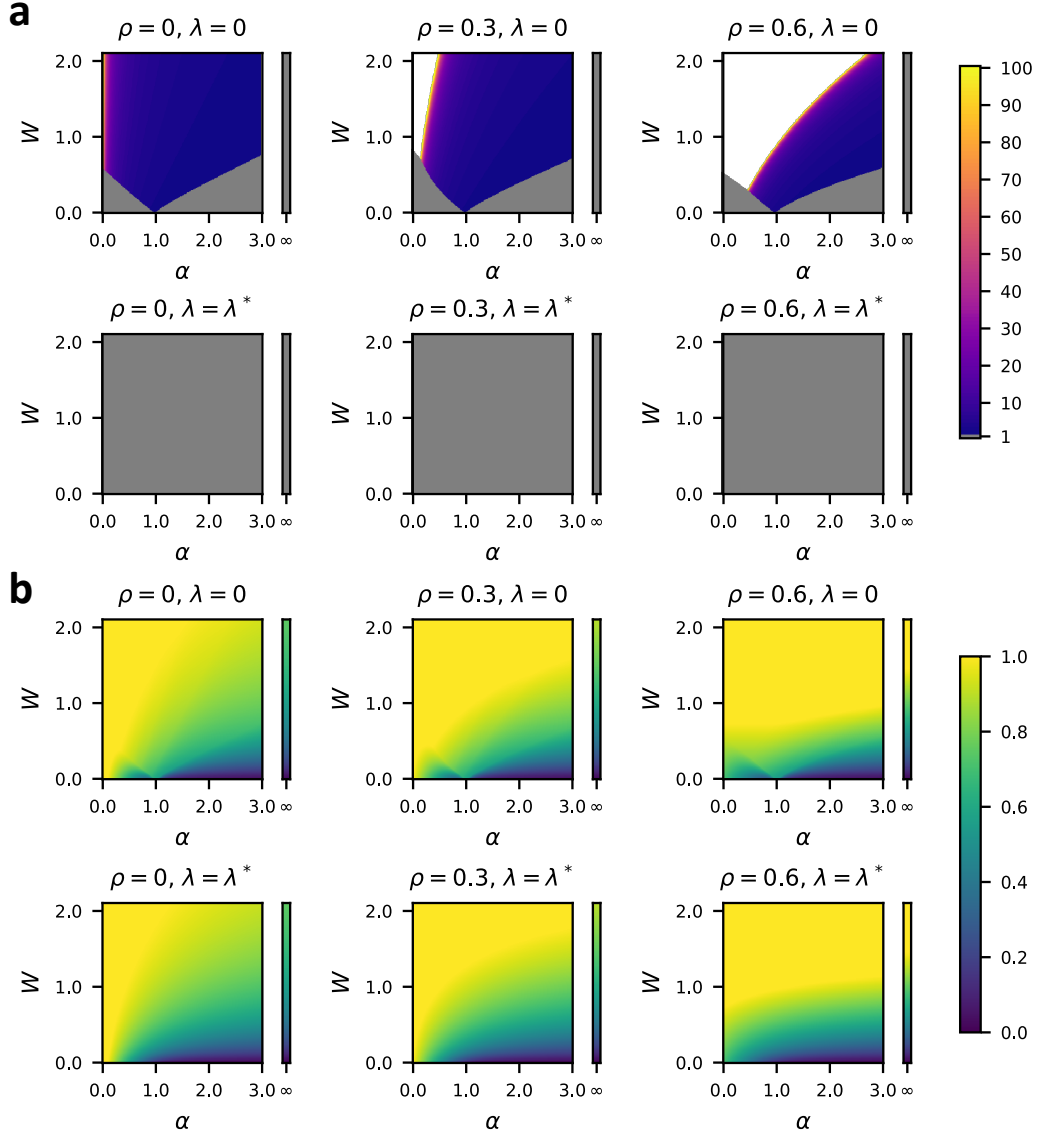


Figure S7: (a) Optimal ensemble size k^* (eqs. 25, 26) in the parameter space of sample size α and reduced feature noise scale W setting $H = Z = 0$. Grey indicates $k^* = 1$ and white indicates $k^* \rightarrow \infty$, with intermediate values given by the colorbar. Appended vertical bars show $\alpha \rightarrow \infty$. ρ and λ indicated in panel titles. $\lambda = \lambda^*$ denotes the ‘locally optimal’ regularization (see section H.3) (b) Optimal generalization error E_{k^*} for the same parameter values in (a).

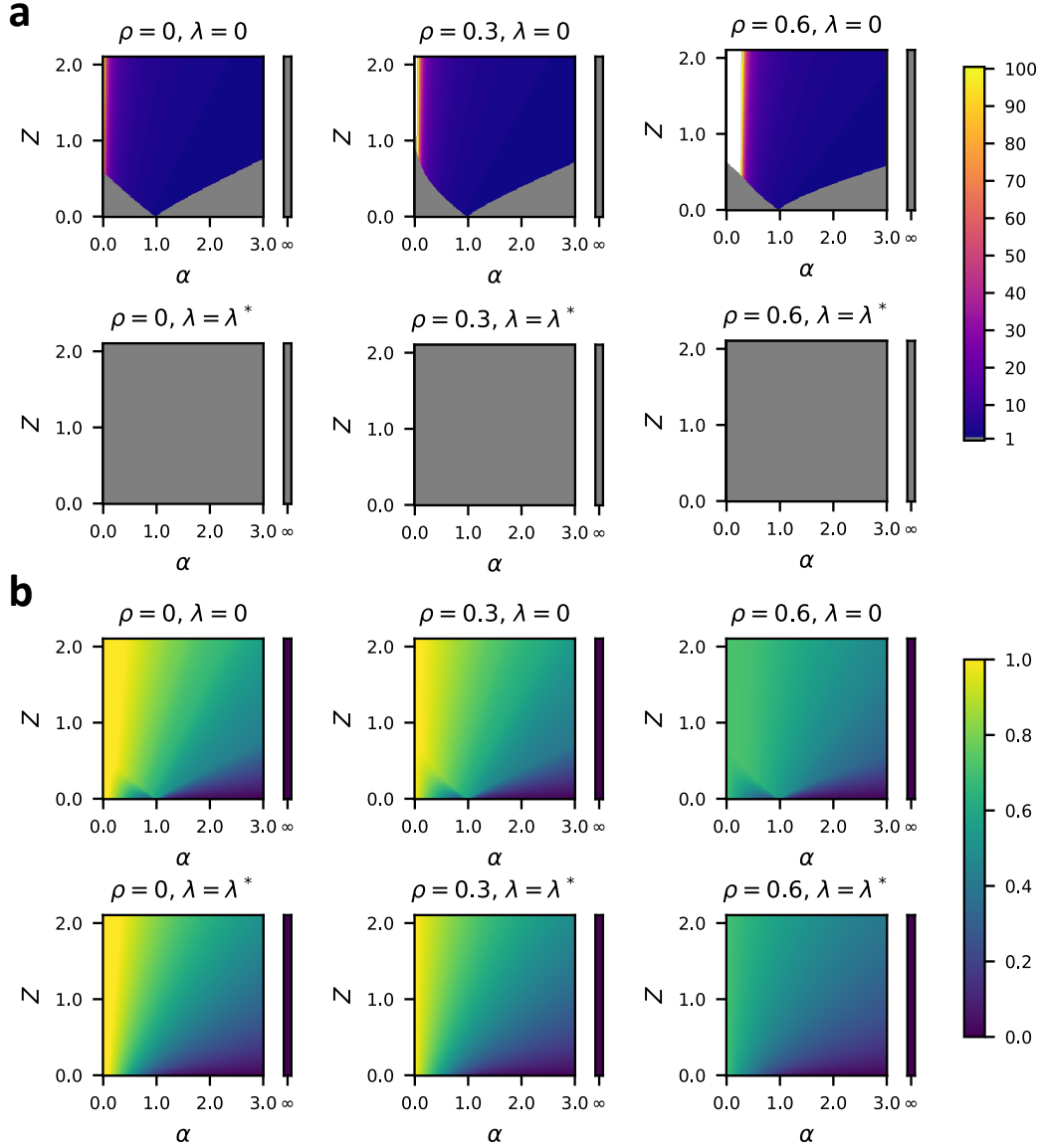


Figure S8: (a) Optimal ensemble size k^* (eqs. 25, 26) in the parameter space of sample size α and reduced label noise scale Z setting $H = W = 0$. Grey indicates $k^* = 1$ and white indicates $k^* \rightarrow \infty$, with intermediate values given by the colorbar. Appended vertical bars show $\alpha \rightarrow \infty$. ρ and λ indicated in panel titles. $\lambda = \lambda^*$ denotes the ‘locally optimal’ regularization (see section H.3) (b) Optimal generalization error E_{k^*} for the same parameter values in (a).

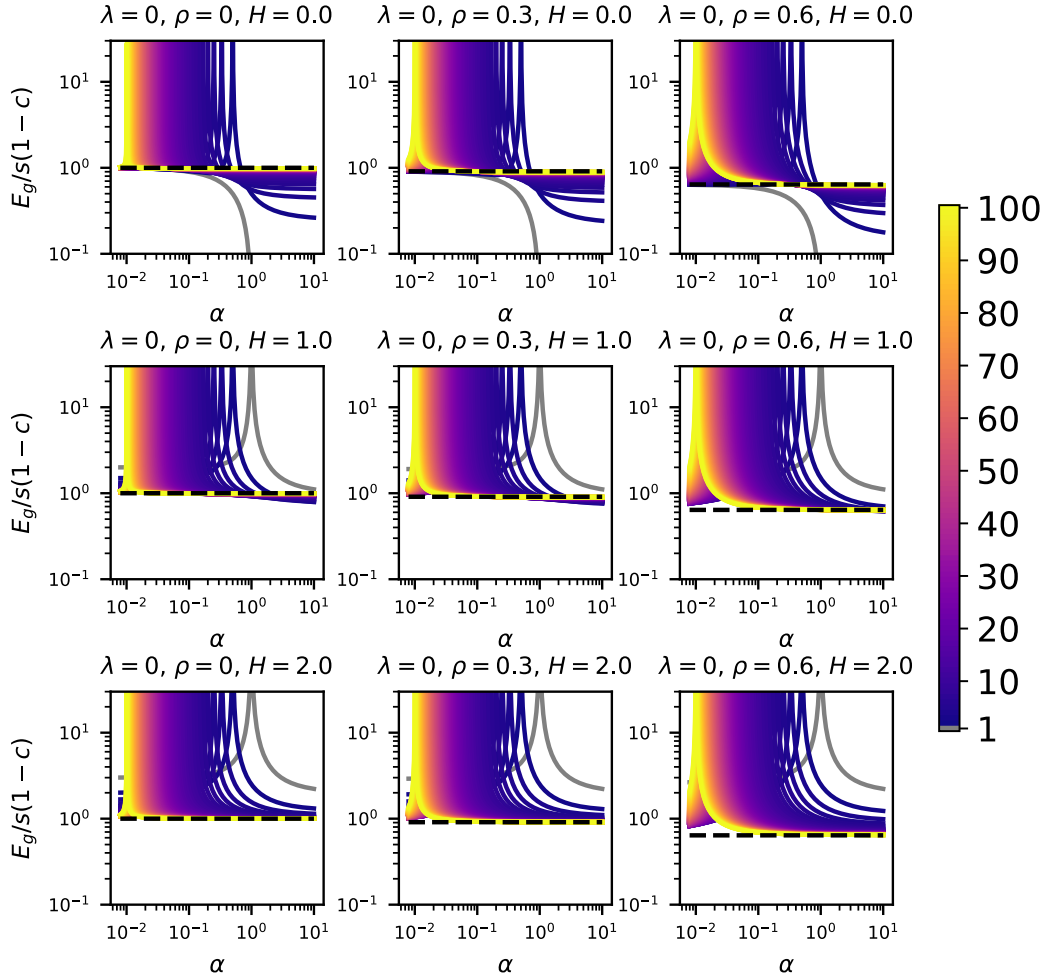


Figure S9: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = 0$ and $W = Z = 0$ (given by eq. 22) for linear ridge ensembles of varying size k . ρ and H values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar.

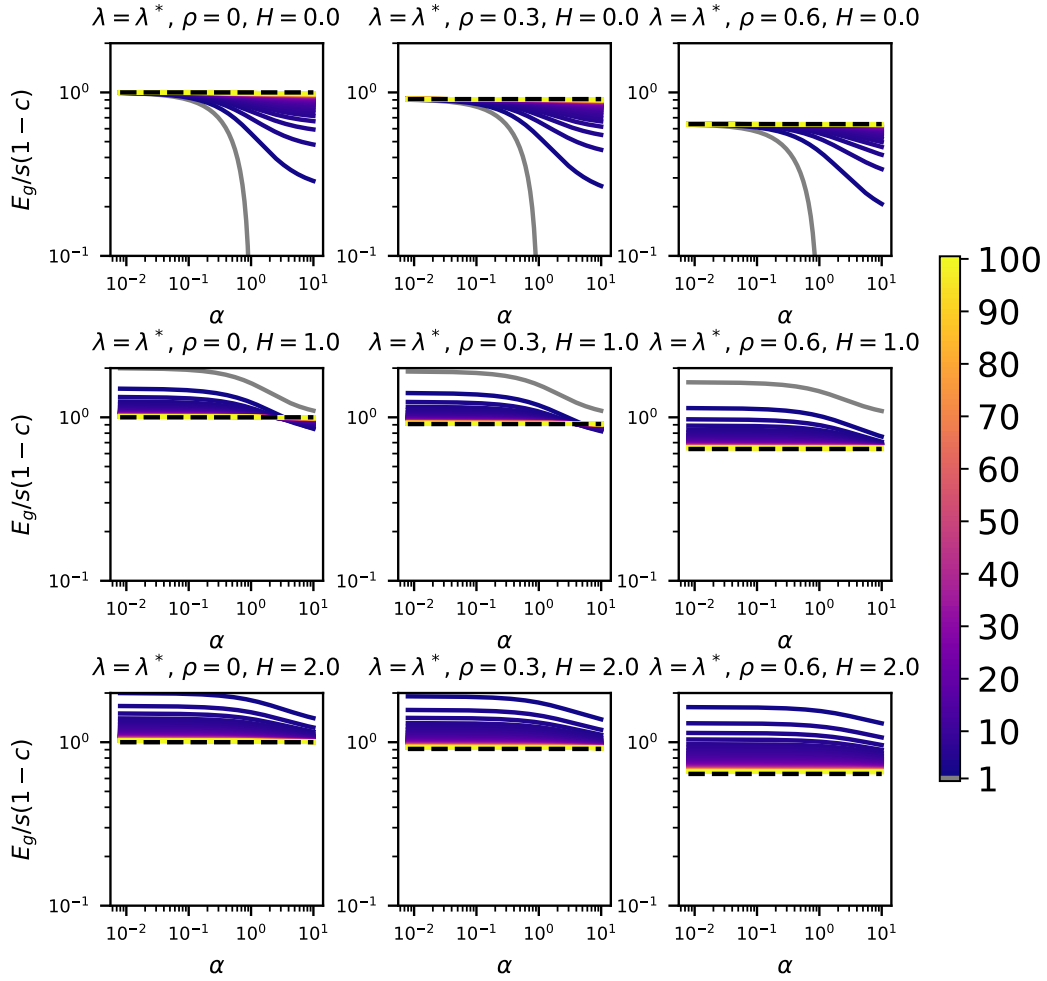


Figure S10: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = \lambda^*$ and $W = Z = 0$ (given by eq. 23) for linear ridge ensembles of varying size k . ρ and H values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar.

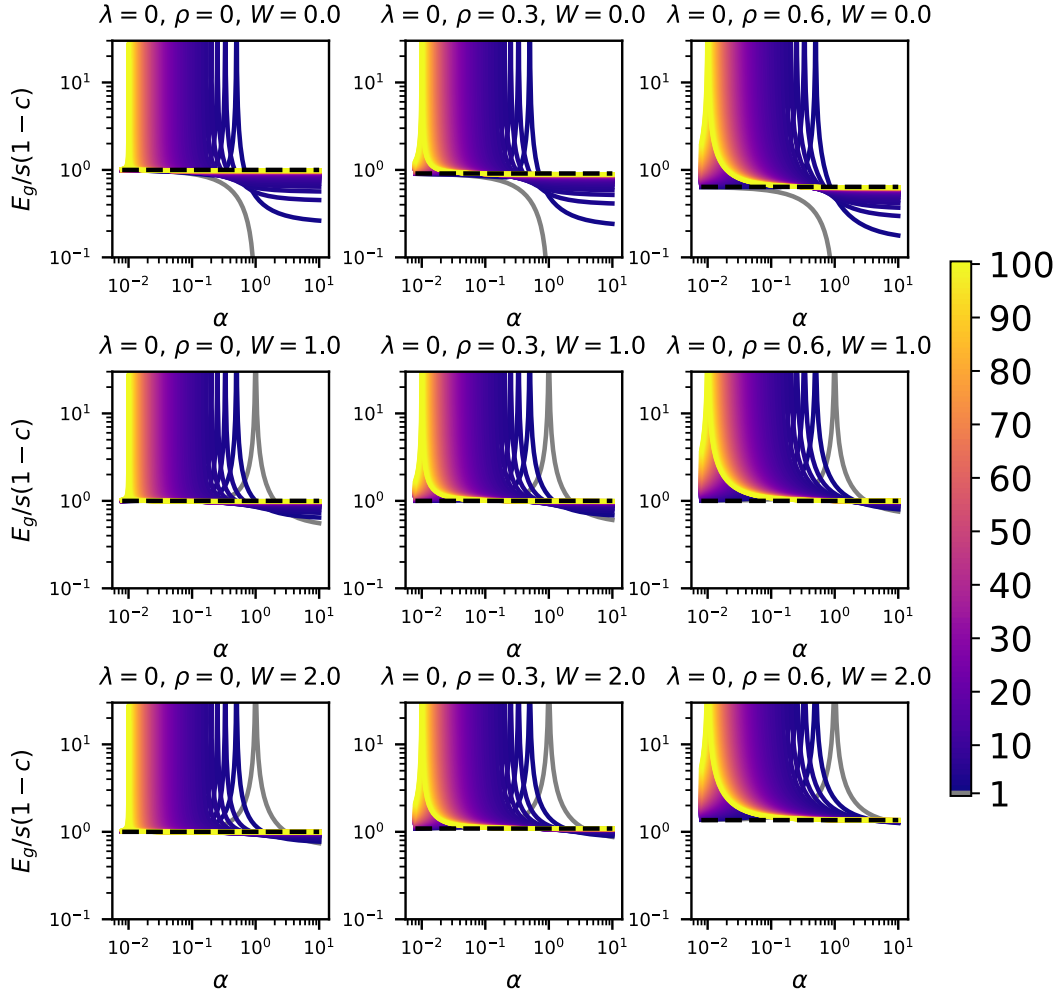


Figure S11: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = 0$ and $H = Z = 0$ (given by eq. 22) for linear ridge ensembles of varying size k . ρ and W values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar.

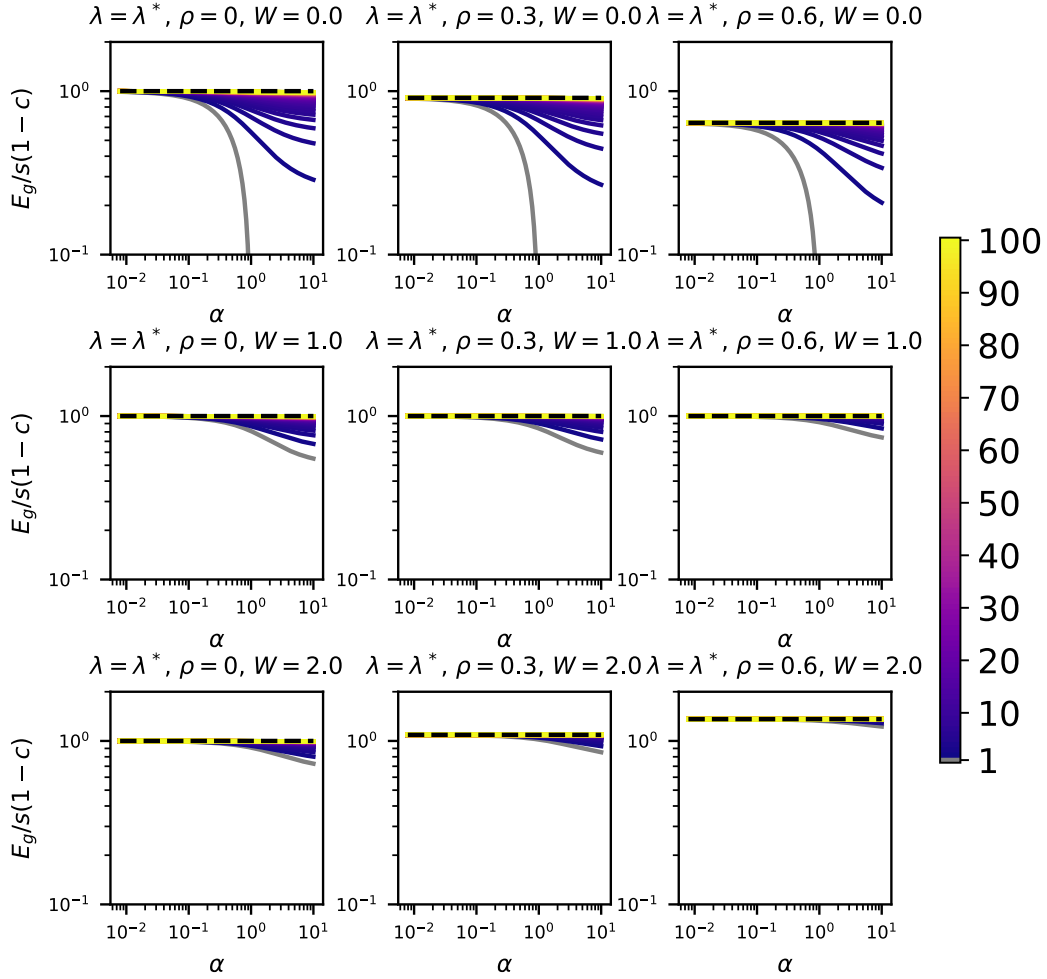


Figure S12: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = \lambda^*$ and $H = Z = 0$ (given by eq. 23) for linear ridge ensembles of varying size k . ρ and W values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar.

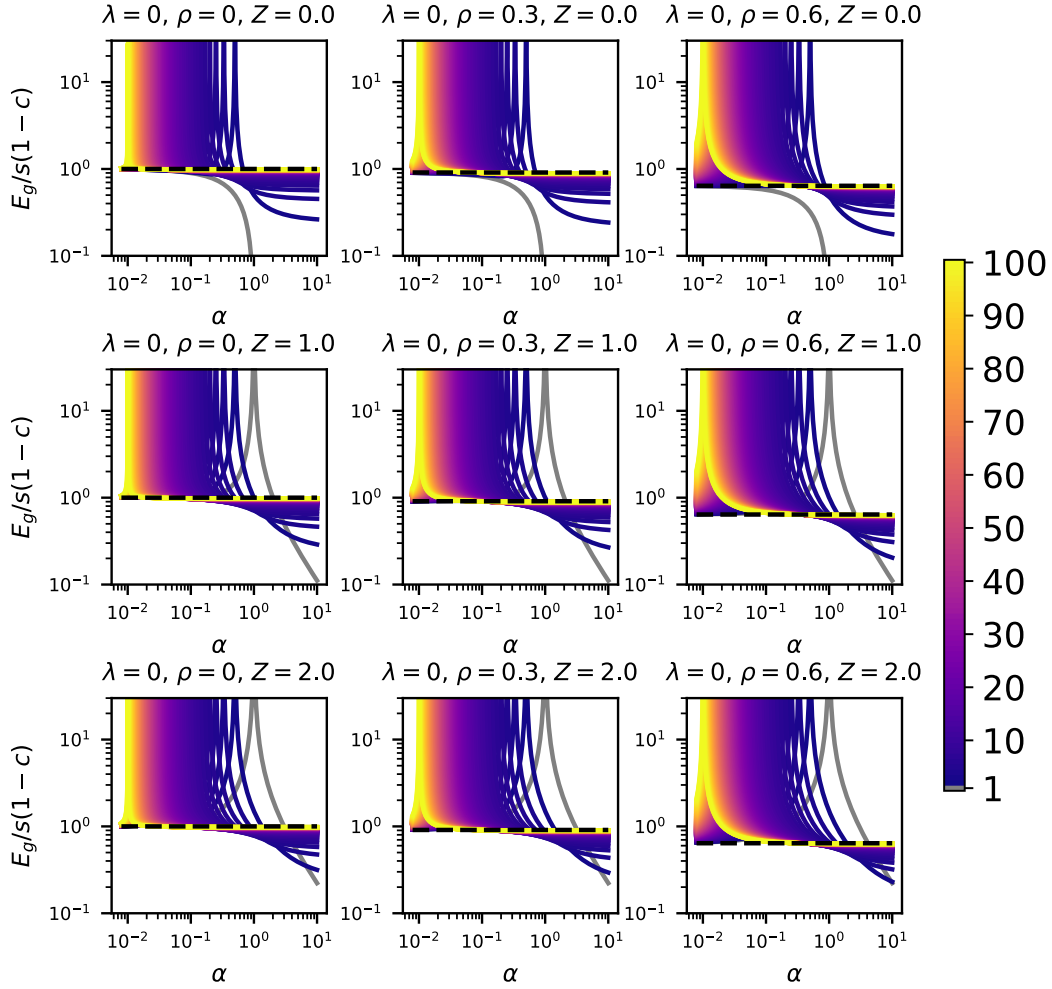


Figure S13: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = 0$ and $H = W = 0$ (given by eq. 22) for linear ridge ensembles of varying size k . ρ and Z values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar

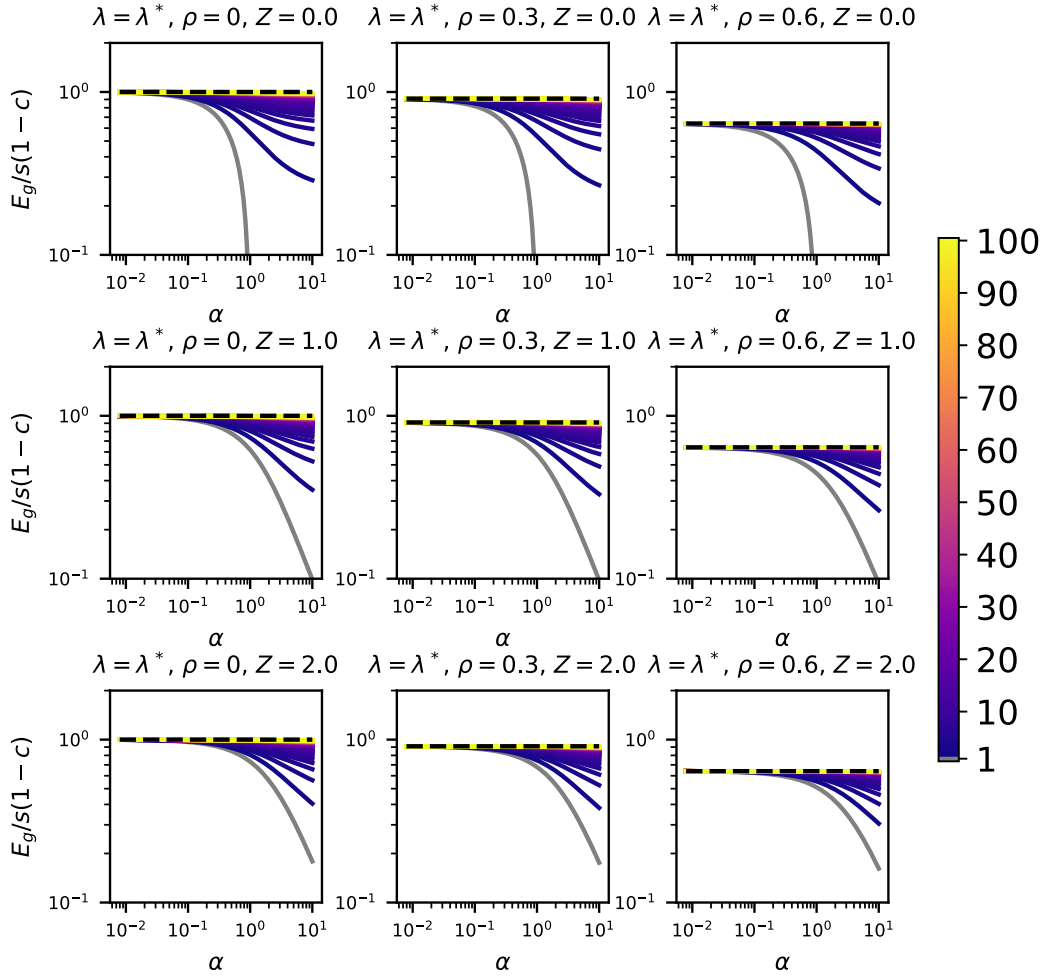


Figure S14: Reduced generalization errors $E_g/s(1-c)$ with $\lambda = \lambda^*$ and $H = W = 0$ (given by eq. 23) for linear ridge ensembles of varying size k . ρ and Z values indicated above plots. Grey lines indicate $k = 1$, dashed black lines $k \rightarrow \infty$, and intermediate k values by the colorbar.