Learning Curves for Heterogeneous Feature-Subsampled Ridge Ensembles

Benjamin S. Ruben[1, *] and Cengiz Pehlevan[2, 3, 4, †]

[1] *Biophysics Graduate Program, Harvard University, Cambridge, Massachusetts 02138, USA*
[2] *John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*
[4] *Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, Massachusetts 02138, USA*

(Dated: July 7, 2023)

Feature bagging is a well-established ensembling method which aims to reduce prediction variance by training estimators in an ensemble on random subsamples or projections of features. Typically, ensembles are chosen to be homogeneous, in the sense the the number of feature dimensions available to an estimator is uniform across the ensemble. Here, we introduce heterogeneous feature ensembling, with estimators built on varying number of feature dimensions, and consider its performance in a linear regression setting. We study an ensemble of linear predictors, each fit using ridge regression on a subset of the available features. We allow the number of features included in these subsets to vary. Using the replica trick from statistical physics, we derive learning curves for ridge ensembles with deterministic linear masks. We obtain explicit expressions for the learning curves in the case of equicorrelated data with an isotropic feature noise. Using the derived expressions, we investigate the effect of subsampling and ensembling, finding sharp transitions in the optimal ensembling strategy in the parameter space of noise level, data correlations, and data-task alignment. Finally, we suggest variable-dimension feature bagging as a strategy to mitigate double descent for robust machine learning in practice.

I. INTRODUCTION

Ensembling methods, where one combines predictions from multiple predictors to achieve a stronger prediction, are ubiquitous in machine learning practice [1]. A popular class of ensembling methods (known as attribute bagging [2] as well as the random subspace method [3]) are based on feature subsampling [2–6], where each predictor has access to only a subset of data features, are independently trained on those features, and their predictions are combined to achieve a stronger prediction. For example, the popular random forest method makes use of this strategy [3, 7]. An advantage of these methods is that they allow parallel processing. For example, Feature-Distributed Machine Learning, combine independent predictions made by agents who only see subsets of available features [8].

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 case of feature-subsampled linear ridge regression using methods from statistical physics [9–12]. This allows us to obtain analytical expressions for typical case performance of feature-subsampled linear ridge regression. Analysis of these equations under special cases reveal interesting phenomena involving effects of noise, regularization, and subsampling on prediction performance.

Our findings relate to double-descent [13, 14], which results from over-fitting to noise and poses a serious problem for practical machine learning. Regularization is commonly used to mitigate double descent, however optimal regularization strength depends on data and noise levels [15, 16]. Our theory reveals an alternative strategy. We observe that subsampling shifts the location of a predictor's sample-wise double-descent peak [14, 16, 17]. An interesting consequence of this is that if the predictors are heterogeneous in the number of features they see, they will go through double-descent at different sample-sizes. Therefore, bagging them will lead a mitigation of double-descent, as when one predictor fails, the others will compensate with accurate predictions.

In summary, we make the following original contributions:

- Using the replica trick from statistical physics [9, 11], we derive the generalization error of ensembled least-squares ridge regression with random structured Gaussian data, deterministic feature maps, and a noisy linear teacher function. Our derivation allows for heterogeneity in the rank of the feature maps of the ensemble members.
- We derive explicit formulas which demonstrate that subsampling alters the interpolation threshold of ridge regression.
- We demonstrate benefits of heterogeneous ensembling as a robust method for mitigating double-descent.
- We analyze the role of data correlations, readout noise, and data-task alignment in determining the optimal ensembling strategy in a tractable special case.

* benruben@g.harvard.edu
† cpehlevan@seas.harvard.edu

Related works: A substantial body of work has elucidated the behavior of linear predictors for a variety of feature maps [13, 16, 18–29]. Several recent works have extended this research to characterize the behavior of ensembled regression using solvable models [25, 30, 31]. Ref. [30] derives expressions for the generalization error of generalized linear models, of which ridge ensembles are a special case, in terms of the solutions to a set of self-consistent equations. However, [30] and [25] focus their analysis on the case of isotropic data and Gaussian random masks of *homogeneous* dimensionality. In contrast, we explicitly consider learning from correlated data by ensembles with heterogeneous readout dimensionality. Our work focuses on the effect of feature-wise subsampling. Additional recent works study the performance of ridge ensembles with example-wise subsampling [32, 33] and simultaneous subsampling of features and examples [31]. These works find that subsampling behaves as an implicit regularization, and prove equivalences between optimal ensembling and optimal regularization. In a similar vein, we consider here ensembling as a safeguard against insufficient regularization. Methods from statistical physics have long been used for machine learning theory [10–12]. Relevant work in this domain include [34] which studied ensembling by data-subsampling in linear regression.

II. LERANING CURVES FOR ENSEMBLED RIDGE REGRESSION FROM THE REPLICA METHOD

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 subsampling of equicorrelated features. Later sections illustrate the implications of the derived formulas.

A. Problem Setup

Consider a training set $\mathcal{D} = \{\bar{\boldsymbol{\psi}}_\mu, y^\mu\}_{\mu=1}^P$ of size P. The training examples $\bar{\boldsymbol{\psi}}_\mu \in \mathbb{R}^M$ are drawn from a Gaussian distribution with Gaussian feature noise: $\bar{\boldsymbol{\psi}}_\mu = \boldsymbol{\psi}_\mu + \boldsymbol{\sigma}_\mu$, where $\boldsymbol{\psi}_\mu \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$ and $\boldsymbol{\sigma}_\mu \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$. Data and noise are drawn i.i.d. so that $\mathbb{E}\left[\boldsymbol{\psi}_\mu \boldsymbol{\psi}_\nu^\top\right] = \delta_{\mu\nu}\boldsymbol{\Sigma}_s$ and $\mathbb{E}\left[\boldsymbol{\sigma}_\mu \boldsymbol{\sigma}_\nu^\top\right] = \delta_{\mu\nu}\boldsymbol{\Sigma}_0$. Labels are generated from a noisy teacher function $y_\mu = \frac{1}{\sqrt{M}}\boldsymbol{w}^{*\top}\boldsymbol{\psi}_\mu + \epsilon^\mu$ 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 $\hat{\boldsymbol{w}}_r \in \mathbb{R}^{N_r}$, $r = 1, \dots, k$. 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:

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

The model's prediction $f(\boldsymbol{\psi})$ is an average over k linear predictors. The "measurement matrices" $\boldsymbol{A}_r \in \mathbb{R}^{N_r \times M}$ 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 A_r to coincide with the rows of the identity matrix – the row indices corresponding to indices of the sampled features. The feature noise $\boldsymbol{\sigma} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$ 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 a regular least-squares loss function with ridge regularization:

$$\hat{\boldsymbol{w}}_r = \underset{\boldsymbol{w}_r \in \mathbb{R}^{N_r}}{\arg\min}\left[\sum_{\mu=1}^{P}\left(\frac{1}{\sqrt{N_r}}\boldsymbol{w}_r^\top \boldsymbol{A}_r \bar{\boldsymbol{\psi}}_\mu + \xi_r^\mu - y_\mu\right)^2 + \lambda|\boldsymbol{w}_r^2|\right] \tag{2}$$

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:

$$E_g(\mathcal{D}) = \left\langle \left(f(\boldsymbol{\psi}) - \frac{1}{\sqrt{M}}\boldsymbol{w}^{*\top}\boldsymbol{\psi}\right)^2 \right\rangle \tag{3}$$

Here, the angular brackets represent an average over the data distribution and noise: $\boldsymbol{\psi} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_s)$, $\boldsymbol{\sigma} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$, $\xi_r \sim \mathcal{N}(0, \eta_r^2)$. The generalization error depends on the particular realization of the dataset $\mathcal{D}$ through the learned

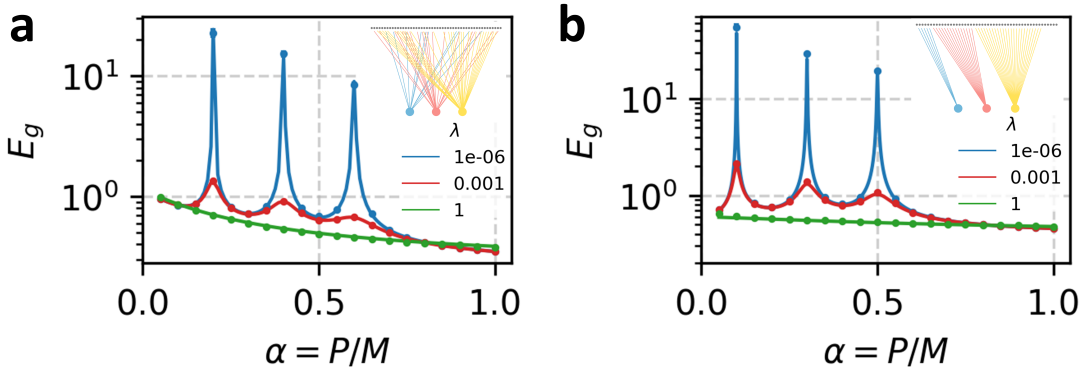

FIG. 1. Comparison between numerically calculated generalization error and theoretical prediction. Dots show results of numerical experiment. Lines are theoretical prediction. (a) Numerical experiment with $[\boldsymbol{\Sigma}_s]_{ij} = .8^{|i-j|}$, $[\boldsymbol{\Sigma}_0]_{ij} = \frac{1}{10}(0.3)^{|i-j|}$, $\zeta = 0.1$, $\eta = 0.2$. We set $k = 3$ with $\nu_1 = 0.2$, $\nu_2 = 0.4$, $\nu_3 = 0.6$. Subsets of feature neurons accessed by each readout are drawn randomly and are permitted to overlap (see inset). Circular markers show the result of numerical experiment with $M = 2000$ feature neurons averaged over 100 trials. Curve shows theoretical prediction, obtained by solving the saddle-point equations 11 numerically. Theory and experiment conducted with a fixed ground-truth readout $\boldsymbol{w}^*$ drawn randomly from an isotropic standard Gaussian distribution (b) Numerical experiment with $[\boldsymbol{\Sigma}_s]_{ij} = (0.6)\delta_{ij} + 0.4$, $[\boldsymbol{\Sigma}_0]_{ij} = .1\delta_{ij}$, $\zeta = 0.1$, $\eta = 0.1$. Ground truth weights are randomly sampled in each trial as in eq. 12 with $\rho = .3$. We set $k = 3$ with $\nu_1 = 0.1$, $\nu_2 = 0.3$, $\nu_3 = 0.5$. Subsets of feature neurons accessed by each readout are are mutually exclusive (see inset). Circular markers show the result of numerical experiment with $M = 5000$ feature neurons averaged over 100 trials. Error bars show the standard error of the mean, and are smaller than the markers. Curve shows analytical prediction obtained in the case of equicorrelated features.

weights $\{\hat{\boldsymbol{w}}^*\}$. We may decompose the generalization error as follows:

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

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

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

B. Main Result

We calculate the generalization error using the replica trick from statistical physics. The result of our calculation is stated in proposition 1. The proof is lengthy, and can be found in the SI.

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

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

$$\boldsymbol{G}_r \equiv \boldsymbol{I}_{N_r} + \hat{q}_r \tilde{\boldsymbol{\Sigma}}_{rr} \tag{7}$$

$$\gamma_{rr'} \equiv \frac{\alpha}{M(\lambda + q_r)(\lambda + q_{r'}))}\,\mathrm{tr}\left[\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r}\right] \tag{8}$$

Then the average generalization error may be written as:

$$\langle E_g(\mathcal{D})\rangle_{\mathcal{D}} = \frac{1}{K^2}\sum_{r,r'=1}^{K}\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D}}, \tag{9}$$

where

$$\begin{aligned}\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D}} =& \frac{\gamma_{rr'}\zeta^2 + \delta_{rr'}\eta_r^2}{1-\gamma_{rr'}} + \frac{1}{1-\gamma_{rr'}}\left(\frac{1}{M}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{w}^*\right) \\ &- \frac{1}{M(1-\gamma_{rr'})}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\left[\frac{1}{\nu_{rr}}\hat{q}_r\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\boldsymbol{A}_r + \frac{1}{\nu_{r'r'}}\hat{q}_{r'}\boldsymbol{A}_{r'}^\top\boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\right]\boldsymbol{\Sigma}_s\boldsymbol{w}^* \\ &+ \frac{\hat{q}_r\hat{q}_{r'}}{M(1-\gamma_{rr'})}\frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\boldsymbol{w}^*\end{aligned} \tag{10}$$

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

$$\hat{q}_r = \frac{\alpha}{\lambda + q_r}, \qquad q_r = \frac{1}{M}\operatorname{tr}\left[\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr}\right]. \tag{11}$$

Proof. We calculate the terms in the generalization error using the replica trick from the statistical physics of disordered systems. The full derivation may be found in the supplemental material. □

We make several remarks on this result:

Remark 1. This is a highly general result which applies to any selection of linear masks $\{\boldsymbol{A}_r\}$. However, we will focus on the case where the $\{\boldsymbol{A}_r\}$ implement subsampling of the feature neurons.

Remark 2. Our result reduces to the results of [35] when $k = 1$ and $\eta = 0$, and may be obtained as a special case of [36] in this limit. In the case where all readout weights have the same dimension $N_r = N, r = 1, \dots, k$, this result may be obtained as a special case of the results of [30]. The novelty in our derivation (and subsequent analysis) is to consider heterogeneity in the values of N_r.

Remark 3. The replica trick [37] is a non-rigorous but standard heuristic in the study of disordered systems. We confirm our results in simulations.

In Figure 1a, we confirm the result of the general calculation by comparing with numerical experiments. Experimental curves are generated by running ridge regression on randomly drawn datasets with $M = 2000$ features and averaging over the resulting error. We use highly structured data, feature noise, label noise, and readout noise (see caption for details). Each of $k = 3$ readouts sees a fixed but randomly drawn subset of features. Theory curves are calculated by solving the fixed-point equations 11 numerically for the chosen $\boldsymbol{\Sigma}_s$, $\boldsymbol{\Sigma}_0$ and $\{\boldsymbol{A}_r\}_{r=1}^k$ then plugging the resulting order parameters into eq. 10.

C. 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, as they depend on the solution to a set of in general analytically intractable self-consistent equations for the order parameters. In order to better understand the phenomena captured by these expressions, we simplify them in the tractable special case in which features of the data are equicorrelated:

Proposition 2. *Consider the ensembled ridge regression problem described in section II A, and the result of proposition 1. Consider the special case in which we select the following parameters:*

$$\boldsymbol{w}^* = \sqrt{1-\rho^2}\mathbb{P}_\perp\boldsymbol{w}_0^* + \rho\mathbf{1}_M \tag{12}$$

$$\boldsymbol{w}_0^* \sim \mathcal{N}(0, \boldsymbol{I}_M) \tag{13}$$

$$\boldsymbol{\Sigma}_s = s\left[(1-c)\boldsymbol{I}_M + c\mathbf{1}_M\mathbf{1}_M^\top\right] \tag{14}$$

$$\boldsymbol{\Sigma}_0 = \omega\boldsymbol{I}_M \tag{15}$$

with $c \in [0,1], \rho \in [-1,1]$*. A label noise scale* $\zeta \geq 0$ *and readout noise scales* $\eta_r \geq 0$ *are permitted. Here* $\mathbb{P}_\perp = \boldsymbol{I}_M - \frac{1}{N}\mathbf{1}_M\mathbf{1}_M^\top$ *is a projection matrix which removes the component of* $\boldsymbol{w}_0^*$ *which is parallel to* $\mathbf{1}_M$*. The measurement matrices* $\{\boldsymbol{A}_r\}_{r=1}^k$ *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* $\boldsymbol{A}_r$ *and* $\boldsymbol{A}_{r'}$ *and let* $\nu_{rr'} \equiv n_{rr'}/M$ *remain fixed as* $M \to \infty$*.*

In this case, we may obtain fully analytical formulas for the generalization error as follows. First define the following quantities:

$$a \equiv s(1-c)+\omega \qquad S_r \equiv \frac{\hat{q}_r}{\nu_{rr}+a\hat{q}_r}, \qquad \gamma_{rr'} \equiv \frac{a^2\nu_{rr'}S_rS_{r'}}{\alpha} \tag{16}$$

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

$$\langle E_{rr'}\rangle_{\mathcal{D},\boldsymbol{w}_0^*} = \frac{1}{1-\gamma_{rr'}}\left((1-\rho^2)I^0_{rr'}+\rho^2 I^1_{rr'}\right)+\frac{\gamma_{rr'}\zeta^2+\delta_{rr'}\eta_r^2}{1-\gamma_{rr'}} \tag{17}$$

where we have defined

$$I^0_{rr'} \equiv s(1-c)\left(1-s(1-c)\nu_{rr}S_r - s(1-c)\nu_{r'r'}S_{r'} + as(1-c)\nu_{rr'}S_rS_{r'}\right) \tag{18}$$

$$I^1_{rr'} \equiv \begin{cases} \frac{s(1-c)(\nu_{rr'}-\nu_{rr}\nu_{r'r'})+\omega\nu_{rr'}}{\nu_{rr}\nu_{r'r'}} & \text{if } 0<c\leq 1 \\ I^0_{rr'} & \text{if } c=0 \end{cases} \tag{19}$$

and where the order parameters $\{q_r,\hat{q}_r\}$ may be obtained analytically as the solution (with $q_r>0$) to the following quadratic system of equations:

$$q_r = \frac{a\nu_{rr}}{\nu_{rr}+a\hat{q}_r} \qquad , \qquad \hat{q}_r = \frac{\alpha}{\lambda+q_r} \tag{20}$$

In the "ridgeless" limit where $\lambda\to 0$, we may make the following simplifications:

$$S_r \to \frac{2\alpha}{a\left(\alpha+\nu_{rr}+|\alpha-\nu_{rr}|\right)} \tag{21}$$

$$\gamma_{rr'} \to \frac{4\alpha\nu_{rr'}}{\left(\alpha+\nu_{rr}+|\alpha-\nu_{rr}|\right)\left(\alpha+\nu_{r'r'}+|\alpha-\nu_{r'r'}|\right)} \tag{22}$$

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 [38] and the fact that the data distribution is isotropic in the features so that the form of $\tilde{\boldsymbol{\Sigma}}_{rr}$ and $\tilde{\boldsymbol{\Sigma}}_{rr'}$ depend only on N_r, $N_{r'}$, and $n_{rr'}$. Thus, the result depends only on the values of $\{\nu_{rr'}\}$ and not the identities of the subsampled features. 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 $\boldsymbol{M}=c_1\boldsymbol{I}+c_2\mathbf{1}\mathbf{1}^\top$. This package and the Mathematica notebook used to derive these results will be made available online (see SI) □

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 ρ tunes the alignment of the ground truth weights with the special direction in the covariance matrix. We refer to ρ as the "task alignment", and it can be thought of as a simple proxy for the "task-model alignment" [16] or "code-task alignment" [39]. In Figure 1b, we test these results by comparing the theoretical expressions for generalization error with the results of numerical experiments, finding perfect agreement. Note that in this case, both theory and experiment are averaged over ground-truth weights as well as datasets.

D. Subsampling shifts the double-descent peak of a linear predictor.

Consider the equicorrelated data model in the isotropic limit ($c=0$). Consider a single linear regressor ($k=1$) which connects to a subset of $N=\nu M$ features. In the ridgeless limit where regularization $\lambda\to 0$, and without readout noise or feature noise ($\eta=\omega=0$), the generalization error is given by equation 17 with $\nu_{rr}=\nu$, $s=1$, $\eta_r=\omega=0$ in the $\lambda\to 0$ limit:

$$\langle E_g\rangle_{\mathcal{D},\boldsymbol{w}^*} = \left\{ \begin{array}{ll} \frac{\nu}{\nu-\alpha}\left[(1-\nu_{rr})+\frac{1}{\nu_{rr}}(\alpha-\nu)^2\right]+\frac{\alpha}{\nu-\alpha}\zeta^2, & \text{if } \alpha<\nu \\ \frac{\alpha}{\alpha-\nu}\left[1-\nu\right]+\frac{\nu}{\alpha-\nu}\zeta^2, & \text{if } \alpha>\nu \end{array} \right\} \tag{23}$$

Double descent can arise from two possible sources of variance: explicit label noise ($\zeta>0$) or implicit label noise induced by feature subsampling ($\nu<1$). As $E_g\sim(\alpha-\nu)^{-1}$, we see that the generalization error diverges when $\alpha=\nu$. The subsampling fraction ν thus controls the sample complexity α at which the double-descent peak occurs. 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. While at finite ridge the test error no longer diverges when $\alpha=\nu$, it may still display a distinctive peak.

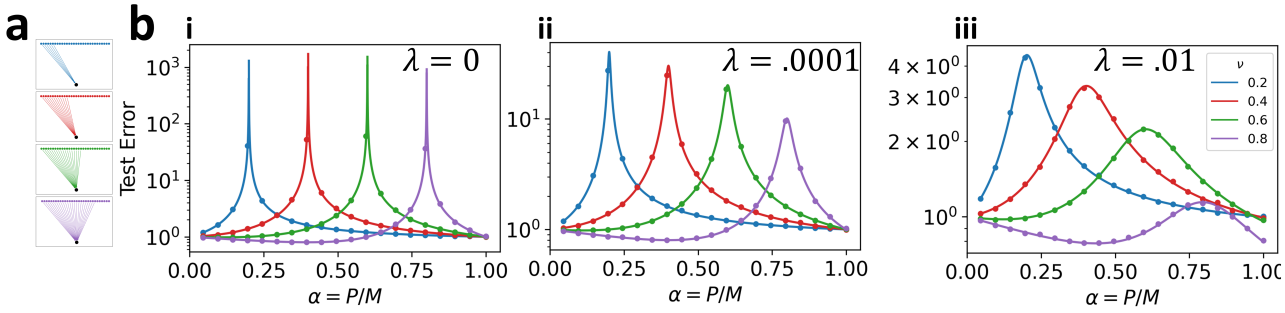

FIG. 2. Subsampling alters the location of the double-descent peak of a linear predictor. (a) Illustrations of subsampled linear predictors with varying subsampling fraction ν. (b) Comparison between experiment and theory for subsampling linear regression on equicorrelated datasets. We choose $[\mathbf{\Sigma}_s]_{ij} = \delta_{ij}$, $[\mathbf{\Sigma}_0]_{ij} = 0$, $\zeta = 0$, $\eta = 0$, and (i) $\lambda = 0$, (ii) $\lambda = 10^{-4}$, (iii) $\lambda = 10^{-2}$. Dots show results of numerical experiment. Lines are analytical prediction.

E. Heterogeneous connectivity mitigates double-descent

The observed phenomenon of 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. Regularization is the canonical strategy employed to mitigate double descent. However, in order to achieve monotonic learning, the regularization parameter must be tuned to the structure of the task and the scale of the label noise [15] – no one choice for the regularization parameter can mitigate double descent for all tasks. 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. Rather than training an ensemble of linear predictors, each with the same interpolation threshold, we may ensemble over predictors with a heterogeneous distribution of interpolation thresholds in the hopes that when one predictor fails, the other members of the ensemble compensate. In Figure 3, we demonstrate that in the absence of a sufficiently regularization, heterogeneous ensembling can mitigate double-descent. Specifically. We define two ensembling strategies: in homogeneous ensembling, each of the k readouts is connected to the same fraction $\nu_{rr} = \frac{1}{k}$ features. In heterogeneous ensembling, the number of features connected by each of the k readouts are drawn i.i.d. from a Gamma distribution with fixed mean $1/k$ and variance σ^2. We denote this $\nu_{rr} \sim \Gamma_{k,\sigma}$. After they are independently drawn, subsampling fractions are re-scaled so that they sum to unity: $\nu_{rr}/\sum_r \nu_{rr} \leftarrow \nu_{rr}$. This ensures fair competition, wherein the total number of readout weights utilized in homogeneous and heterogeneous ensembling are equal. Equivalently, we may consider the readout fractions ν_{rr} to be drawn from a Dirichlet distribution: $(\nu_1, \ldots, \nu_k) \sim \mathrm{Dir}((\sigma k)^{-2}, \ldots, (\sigma k)^{-2})$ [40]. These strategies for connecting readouts to the features are illustrated for $k = 10$ in figures 3 a.i (homogeneous) and 3 a.ii (heterogeneous). The density of the distribution $\Gamma_{k,\sigma}(\nu)$ is plotted in figure 3b for $k = 10$ and varying σ. In figure S1, we apply these ideas to a classification task on the CIFAR-10 dataset. We find that in this nonlinear setting, heterogeneous ensembling prevents catastrophic over-fitting, leading to monotonic learning curves without regularization (see SI for details).

In figure 3c, we use our analytical theory of equicorrelated data (see eqs. 17) to compare the performance of homogeneous and heterogeneous ensembling with $k = 10$. We find that for an under-regularized predictor, (3c.i, c.ii, c.iii) heterogeneous ensembling reduces the height of the double-descent peak. At larger regularization (3c.iv, c.v, c.vi), homogeneous and heterogeneous ensembling perform similarly. We quantify the extent of double-descent through the worst-case error $\max_\alpha(E_g(\alpha))$. We find that as σ increases, the worst-case error decreases monotonically at no cost to the asymptotic error $E_g(\alpha \to \infty)$ (see Fig. 3d,e).

F. Data correlations, readout noise, and task structure determine optimal ensemble size

We now ask whether ensembling is a fruitful strategy – i.e. whether it is preferable to have a single, fully connected readout or multiple sparsely connected readouts. Intuitively, the presence of correlations between features permits subsampling, as measurements from a subset of neurons will also confer information about the state of the others. In addition, ensembling over multiple readouts can average out the readout noise. To quantify these notions, we consider the special case of ensembling over k readouts, each connecting the same fraction $\nu_{rr} = \nu = \frac{1}{k}$ of features in an equicorrelated code with correlation strength c and readout noise scale η, and task alignment ρ. We set the label noise, feature noise, and overlap between readouts to zero ($\zeta = 0$, $\omega = 0$, $\nu_{rr'} = 0$ when $r \neq r'$). In the ridgeless limit, we can then express the error as : $E_g(k) = s(1-c)F(H, k, \rho, \alpha)$, where $H \equiv \frac{\eta^2}{s(1-c)}$ is an effective inverse signal-to-noise ratio

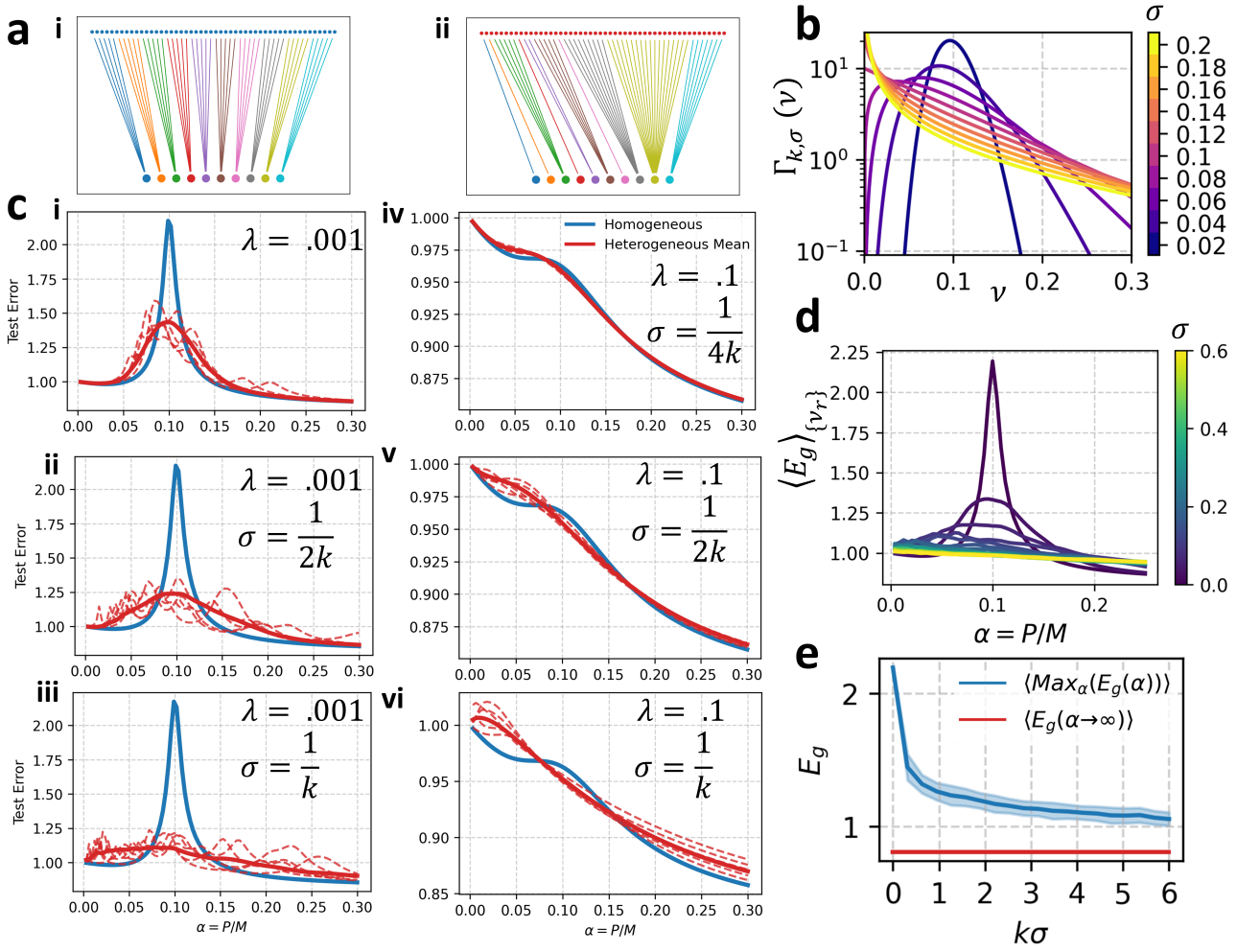

FIG. 3. Homogeneous vs. Heterogeneous Ensembling on equicorrelated data. (a) We compare (i) homogeneous ensembling, in which each readout connects to the same number of feature neurons and (ii) heterogeneous ensembling, in which the number of feature neurons connected by a readout is drawn from a distribution. (b) We use the Gamma distribution with the convention that $\Gamma_{k,\sigma}(\nu)$ is the probability density function of the Gamma distribution with mean k^{-1} and variance σ^2. Shown here for $k = 10$ and σ indicated by the colorbar. (c) Generalization Error Curves for Homogeneous and Heterogeneous ensembling with $k = 10$ and indicated values of λ and σ. Curves are calculated using analytical theory for equicorrelated data with $c = 0$, $\eta = 0$, $\zeta = 0$. Solid blue is the learning curve for homogeneous subsampling. Dotted red curves show loss curve for 5 realizations of the randomly drawn subsampling fractions $\{\nu_{rr}\}_{r=1}^k$. Solid red is the learning curve for heterogeneous ensembling averaged over 100 realizations of the subsampling fractions $\{\nu_{rr}\}_{r=1}^k$ drawn independently from $\Gamma_{k,\sigma}(\nu)$. (d) Average loss curves for heterogeneous ensembling with $k = 10$, $\lambda = 10^{-3}$, and σ indicated by the colorbar. (e) Average worst-case error and asymptotic error as a function of variance for heterogeneous ensembling. Worst-case error is calculated for each realization of the subsampling fractions as $\max_\alpha E_g(\alpha|\{\nu_{rr}\}_{r=1}^k)$. Average worst-case error is the worst-case error averaged over realizations of the subsampling fractions. Shaded region shows standard deviation over realizations of the subsampling fractions.

and $F(H, k, \rho, \alpha)$ is a rational function of its arguments (see SI for full expressions). Therefore, given fixed parameters s, c, ρ, α, the value k^* which minimizes error depends on η, s, and c only through the ratio H.

Using our analytical theory, we plot the optimal number of readouts k in the parameter space of H and ρ (see Fig. 4a). The resulting phase diagrams are naturally 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 $k^* = \infty$. The boundary between the signal-dominated and noise-dominated phases (dotted lines in 4a) can be written $H = (1 - \frac{1}{\alpha})(1 - \rho^2)$ when $\alpha > 1$ and $H = \alpha(1 - \alpha)(1 - \rho^2)$ when $\alpha < 1$. The boundary between the intermediate and noise-dominated phases (dashed lines in 4a) can be written $H = 2 - (2 + \frac{1}{\alpha})\rho^2$. As is evident in these phase diagrams, 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 when there is alignment between the structure of

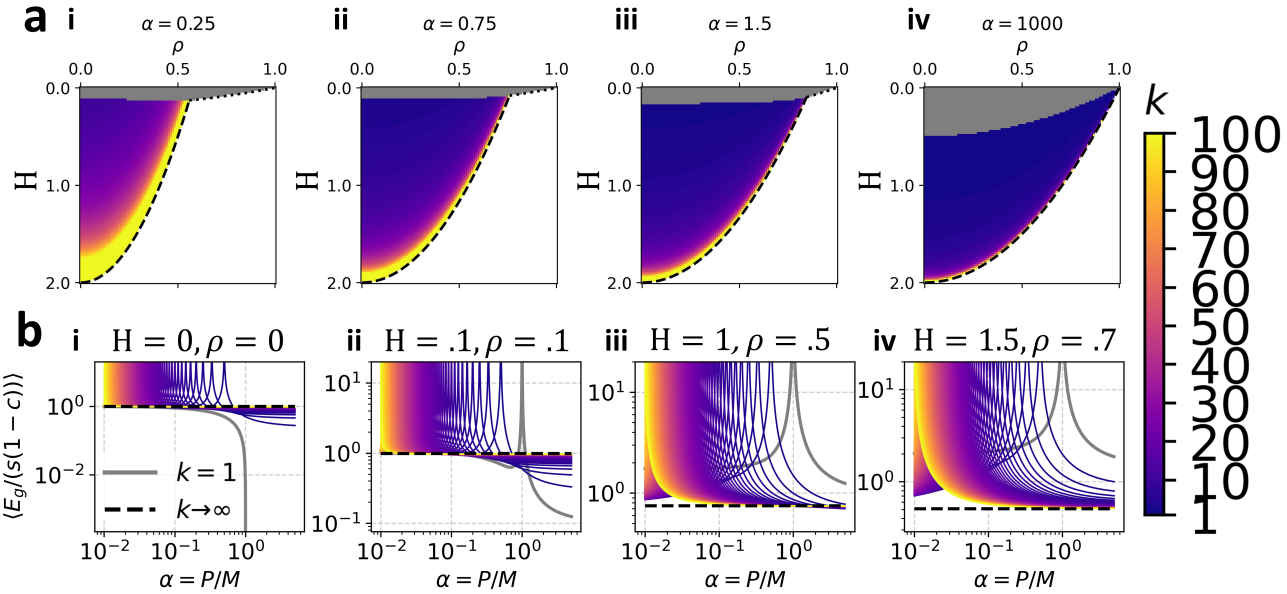

FIG. 4. Noise level and data correlation strength determine optimal readout strategy: Using analytical theory (see eq. 17), we calculate the generalization error of linear predictors on equicorrelated data ($[\mathbf{\Sigma}_s]_{ij} = (1-c)\delta_i j + c$, $0 < c \leq 1$) with readout noise with variance η^2. Ground truth weights are drawn as in eq. 12. For convenience, we set $\lambda = 0$, though results are qualitatively similar with small finite ridge. We consider k readouts, each connecting a fraction $\nu = 1/k$ of the feature neurons, so that the total number of readout weights is conserved. (a) Phase diagrams of optimal k in the parameter space of task alignment ρ and the inverse effective signal-to-noise ratio $H \equiv \frac{\eta^2}{s(1-c)}$. Color indicates the optimal number of readouts k^*, with gray indicating $k^* = 1$ and white indicating $k^* = \infty$ We consider (i) $\alpha = 0.25$, (ii) $\alpha = 0.75$, (iii) $\alpha = 1.5$, (iv) $\alpha = 10^3$. Black lines are analytically derived phase boundaries between regions of parameter space where finite optimal k^* exists and where $k^* = \infty$. Dotted black lines are phase boundaries of the type where k^* jumps discontinuously from 1 to ∞. Dashed black lines are phase boundaries of the type where $k^* \to \infty$ from one side and $k^* = \infty$ on the other. (b) for three choices of the parameters (H, ρ) we plot the learning curve for ensembled linear regression for a variety of k values (see colorbar), as well as $k = \infty$, indicated by the dotted black line. Depending on the region of parameter space, the optimal readout strategy may be to select $k* = 1$, $1 < k* < \infty$, or $k* = \infty$.

the data and the task. Learning curves from each of these phases for varying k are plotted in Fig. 4b. The resulting shifts in the location of the double-descent peak resemble those observed in practice for ensembling methods applied to linear classifiers [6].

III. CONCLUSION

In this paper, we provided a theory of feature-subsampled ensembling techniques focusing on feature-subsampled linear ridge regression. Our technique was the replica method from statistical physics which led us to derive an analytical formula for the typical case generalization error in the aforementioned setting. We solved these equations for a special case which revealed many interesting phenomena.

One of these phenomena relate to double descent [13, 14]. 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 [13] or by cross-validation techniques (see for example [19]). 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. Our analysis of ensembling in noisy neural networks suggests that an ensembling approach may be useful in improving the stability of analog neural networks, where readout noise is a significant problem (see, for example, [41]).

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 and readouts do not overlap. Future work should analyze ensembling for codes with an arbitrary correlation structure, in which readouts access randomly chosen, potentially overlapping subsets of features. This will require to average our expressions for the generalization error over randomly drawn masks $\{A_r\}$. This problem has

been thoroughly studied in the case where the entries of A_r are i.i.d Gaussian [30], as in the ever-popular random feature model. Recent progress on the problem of non-Gaussian projections for a single readout has been made in [42].

IV. ACKNOWLEDGEMENTS

CP and this research were supported by NSF Award DMS-2134157. 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.

[1] G. Kunapuli, *Ensemble Methods for Machine Learning* (Simon and Schuster, 2023).
[2] R. Bryll, R. Gutierrez-Osuna, and F. Quek, Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern recognition **36**, 1291 (2003).
[3] T. K. Ho, The random subspace method for constructing decision forests, IEEE transactions on pattern analysis and machine intelligence **20**, 832 (1998).
[4] Y. Amit and D. Geman, Shape quantization and recognition with randomized trees, Neural computation **9**, 1545 (1997).
[5] G. Louppe and P. 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* (Springer, 2012) pp. 346–361.
[6] M. Skurichina and R. P. W. Duin, Bagging, boosting and the random subspace method for linear classifiers, Pattern Analysis & Applications **5**, 121 (2002).
[7] T. K. Ho, Random decision forests, in *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1 (IEEE, 1995) pp. 278–282.
[8] Y. Hu, D. Niu, J. Yang, and S. Zhou, Fdml: A collaborative machine learning framework for distributed features, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19 (Association for Computing Machinery, New York, NY, USA, 2019) p. 2232–2240.
[9] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific Publishing Company, 1987).
[10] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Physical review A **45**, 6056 (1992).
[11] A. Engel and C. Van den Broeck, *Statistical mechanics of learning* (Cambridge University Press, 2001).
[12] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annual Review of Condensed Matter Physics **11**, 501 (2020).
[13] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proceedings of the National Academy of Sciences **116**, 15849 (2019).
[14] P. Nakkiran, More data can hurt for linear regression: Sample-wise double descent, arXiv preprint arXiv:1912.07242 (2019).
[15] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, Optimal regularization can mitigate double descent, in *International Conference on Learning Representations* (2021).
[16] A. Canatar, B. Bordelon, and C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nature communications **12**, 2914 (2021).
[17] F. F. Yilmaz and R. Heckel, Regularization-wise double descent: Why it occurs and how to eliminate it, in *2022 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2022) pp. 426–431.
[18] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, The Annals of Statistics **50**, 949 (2022).
[19] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2 (Springer, 2009).
[20] J. W. Rocks and P. Mehta, Bias-variance decomposition of overparameterized regression with random linear features, Physical Review E **106**, 025304 (2022).
[21] J. W. Rocks and P. Mehta, Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models, Physical Review Research **4**, 013201 (2022).
[22] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve, Communications on Pure and Applied Mathematics **75**, 10.1002/cpa.22008 (2019), arXiv:1908.05355.
[23] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, Benign overfitting in linear regression, Proceedings of the National Academy of Sciences **117**, 30063 (2020), https://www.pnas.org/doi/pdf/10.1073/pnas.1907378117.
[24] H. Hu and Y. M. Lu, Universality laws for high-dimensional learning with random features, IEEE Transactions on Information Theory **69**, 1932 (2023).
[25] S. D'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, Double trouble in double descent: Bias and variance(s) in the lazy regime, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 2280–2290.

[26] B. Adlam and J. Pennington, Understanding double descent requires a fine-grained bias-variance decomposition, in *Advances in Neural Information Processing Systems*, Vol. 33 (2020) pp. 11022–11032, arXiv:2011.03321.

[27] J. A. Zavatone-Veth, W. L. Tong, and C. Pehlevan, Contrasting random and learned features in deep bayesian linear regression (2022).

[28] B. Bordelon, A. Canatar, and C. Pehlevan, Spectrum dependent learning curves in kernel regression and wide neural networks, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020) pp. 1024–1034, 2002.02561.

[29] J. B. Simon, M. Dickens, D. Karkada, and M. R. DeWeese, The Eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks, arXiv https://doi.org/10.48550/arXiv.2110.03922 (2022), 2110.03922 [cs.LG].

[30] B. Loureiro, C. Gerbelot, M. Refinetti, G. Sicuro, and F. Krzakala, Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR, 2022) pp. 14283–14314.

[31] D. LeJeune, H. Javadi, and R. Baraniuk, The implicit regularization of ordinary least squares ensembles, in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 108, edited by S. Chiappa and R. Calandra (PMLR, 2020) pp. 3525–3535.

[32] J.-H. Du, P. Patil, and A. K. Kuchibhotla, Subsample ridge ensembles: Equivalences and generalized cross-validation (2023), arXiv:2304.13016 [math.ST].

[33] P. Patil, J.-H. Du, and A. K. Kuchibhotla, Bagging in overparameterized learning: Risk characterization and risk monotonization (2022), arXiv:2210.11445 [math.ST].

[34] P. Sollich and A. Krogh, Learning with ensembles: How overfitting can be useful, Advances in neural information processing systems **8** (1995).

[35] A. Atanasov, B. Bordelon, S. Sainathan, and C. Pehlevan, The onset of variance-limited behavior for networks in the lazy and rich regimes, in *The Eleventh International Conference on Learning Representations* (2023).

[36] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborova, Learning curves of generic features maps for realistic datasets with a teacher-student model, in *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (2021).

[37] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (WORLD SCIENTIFIC, 1986).

[38] J. Sherman and W. J. Morrison, Adjustment of an inverse matrix corresponding to a change in one element of a given matrix, The Annals of Mathematical Statistics **21**, 124 (1950).

[39] B. Bordelon and C. Pehlevan, Population codes enable learning from few examples by shaping inductive bias, eLife **11**, 10.7554/elife.78606 (2022).

[40] L. Devroye, *Non-Uniform Random Variate Generation* (Springer New York, 1986).

[41] D. Janke and D. 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, 2020).

[42] F. Bach, High-dimensional analysis of double descent for linear regression with random projections (2023), arXiv:2303.01372 [cs.LG].

[43] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Tech. Rep. (University of Toronto, 2009).

[44] J. R. Silvester, Determinants of block matrices, The Mathematical Gazette **84**, 460 (2000).

[45] T.-T. Lu and S.-H. Shiou, Inverses of 2 x 2 block matrices, Computers & Mathematics with Applications **43**, 119 (2002).

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019) pp. 8024–8035.

Appendix A: Application to CIFAR10 Classification Task

In this section, we demonstrate that the 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 show that ensembling with heterogeneous readout connectivity can mitigate double-descent in the CIFAR10 classification task [43] without regularization. The experimental setup is as follows:

To obtain a useful feature map, we first train a deep, fully connected multi-layer perceptron (MLP) to solve the CIFAR10 classification task. The MLP has three hidden layers of size $N_h = 1000$. We use a training set of 50,000 images $\boldsymbol{x}_\mu \in \mathbb{R}^{N_0}$, $N_0 = 3072$ from $N_{out} = 10$ classes. The targets are assigned as one-hot vectors We write this network as follows:

$$\boldsymbol{h} = \mathrm{ReLU}(N_{in}^{-1/2}\boldsymbol{W}_{in}\boldsymbol{x}) \tag{A1}$$

$$\boldsymbol{h}_2 = \mathrm{ReLU}(M^{-1/2}\boldsymbol{W}_1\boldsymbol{h}_1) \tag{A2}$$

$$\boldsymbol{\psi}(\boldsymbol{x}) = \mathrm{ReLU}(M^{-1/2}\boldsymbol{W}_2\boldsymbol{h}_2) \tag{A3}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = M^{-1/2}\boldsymbol{W}_{out}\boldsymbol{\psi}(\boldsymbol{x}) \tag{A4}$$

Where $\boldsymbol{W}_{in} \in \mathbb{R}^{N_h \times N_{in}}$, $\boldsymbol{W}_1 \in \mathbb{R}^{N_h \times M}$, $\boldsymbol{W}_2 \in \mathbb{R}^{N_h \times M}$, $\boldsymbol{W}_{out} \in \mathbb{R}^{10 \times M}$ We use a MSE loss function:

$$\ell(\{\boldsymbol{W}_{in}, \boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_{out}\}) = \sum_\mu (\hat{\boldsymbol{y}}(\boldsymbol{x}_\mu) - \boldsymbol{y}_\mu) \tag{A5}$$

The predicted class is then assigned as the class corresponding to the component of $\hat{\boldsymbol{y}}$ with maximum value. Training for 2000 steps with full-batch gradient descent and the Adam optimizer at a learning rate of .001, the network achieves a training accuracy of 90% and a test accuracy of $\sim$ %50. The learned features $\boldsymbol{\psi}(\boldsymbol{x})$ are then extracted and new readout weights are trained using the homogeneous or heterogeneous ensembling strategies with modifications for handling multiple readout classes. An ensemble of k predictions is made:

$$\boldsymbol{y}^r(\boldsymbol{x}) = \frac{1}{\sqrt{N_r}}\boldsymbol{W}_r\boldsymbol{A}_r\boldsymbol{\psi}(\boldsymbol{x}) \tag{A6}$$

for $r = 1, \ldots, k$. Here, each $\boldsymbol{A}_r \in \mathbb{R}^{N_r \times N_{in}}$ implements subsampling of a randomly drawn subset of N_r features, and $\boldsymbol{w}_r \in \mathbb{R}^{N_{out} \times N_r}$ predict the class of the input from these subsampled features. The weights $\boldsymbol{W}_r$ are trained independently using a pseudoinverse rule, which is equivalent to ridge regression in the limit of zero regularization. Finally, the predictions of the ensemble of readouts are combined using a mean with a nonlinear threshold:

$$\hat{y}(x) = \sum_{r=1}^{k} \phi(\boldsymbol{y}^r(\boldsymbol{x})) \tag{A7}$$

$$\phi(x) = \frac{1}{2}\left(1 + \tanh(5(x - \frac{1}{2}))\right) \tag{A8}$$

In supplemental figure 1, we demonstrate the performance of re-learning ensembles of readout weights using the homogeneous and heterogeneous ensembling strategies. To review, in homogeneous ensembling, the subsampling fractions $\nu_{rr} = \frac{N_r}{M}$ are chosen as $\nu_{rr} = 1/k$, $r = 1, \ldots, k$. However, in heterogeneous subsampling, the weights are drawn from a gamma distribution with mean $1/k$ and variance σ^2, then re-scaled so that $\sum_r \nu_{rr} = 1$. In figure S1, we use $\sigma = 1/(2k)$.

We that the heterogeneous ensembling approach leads to a smooth learning curve without double descent. This effect is most pronounced in the plots of "test accuracy", which is the probability of incorrect classification of a test example. While homogeneous ensembling shifts the double-descent peak toward smaller P, heterogeneous ensembling eliminated the peak. Because the data is heavily correlated, there is no cost to the prediction performance at large P.

Appendix B: 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;

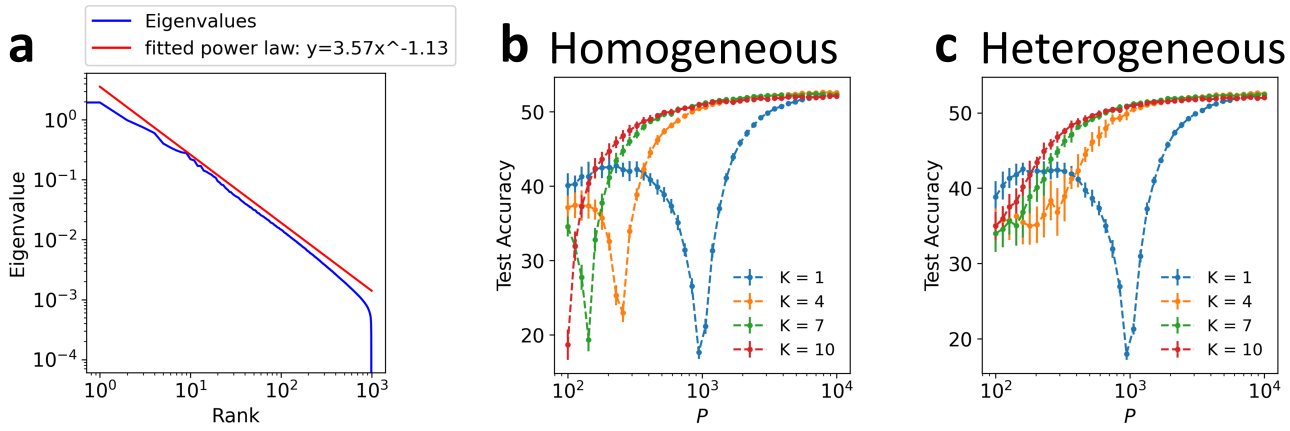

FIG. 5. Homogeneous and Heterogeneous Subsampling applied to CIFAR10 demonstrates the benefit of heterogeneous ensembling. (a) Eigenvalue spectrum of the empirical feature covariance matrix $\mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{\psi}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{\psi}(\boldsymbol{x}))(\boldsymbol{\psi}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}\boldsymbol{\psi}(\boldsymbol{x}))^\top\right]$. Spectrum is well fit by a power law with exponent $\lambda_n \propto n^{-\alpha}$ with exponent $\alpha = 1.13$. (b,c) Test Accuracy defined as the fraction of test examples which are incorrectly classified by the ensemble of readouts using homogeneous ensembling strategy (b) and heterogeneous ensembling strategy (c) with $\sigma = \frac{1}{2k}$.

$$Z = \int \prod_r d\boldsymbol{w} \exp\left(-\frac{\beta}{2}\sum_r E_t^r - \frac{M\beta}{2}\sum_{r,r'} J_{rr'} E_{rr'}(\boldsymbol{w_r}, \boldsymbol{w}_{r'})\right) \tag{B1}$$

$$E_t^r = \sum_{\mu=1}^{P}\left(\frac{1}{\sqrt{N_r}}\boldsymbol{w}_r^\top \boldsymbol{A}_r \bar{\boldsymbol{\psi}}_\mu + \xi_r - y_\mu\right)^2 + \lambda|\boldsymbol{w}_r^2| \tag{B2}$$

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

$$\langle \log(Z[\mathcal{D}])\rangle_{\mathcal{D}} = \lim_{n\to 0}\frac{1}{n}\log\left(\langle Z^n\rangle_{\mathcal{D}}\right) \tag{B3}$$

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: $\mathcal{D} = \{\boldsymbol{\psi}_\mu, \boldsymbol{\sigma}^\mu, \epsilon^\mu\}_{\mu=1}^P$. 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 $n \to 0$.

1. Diagonal Terms

We start by calculating E_{rr} for some fixed choice of r. Noting that the diagonal terms of the generalization error E_{rr} only depend on the learned weights $\boldsymbol{w}_r$, and the loss function separates over the readouts, we may consider the Gibbs measure over only these weights:

$$Z = \int d\boldsymbol{w}_r \exp\left(-\frac{\beta}{2\lambda}E_r^t - \frac{JM\beta}{2}E_{rr}(\boldsymbol{w}_r)\right) \tag{B4}$$

$$\langle Z^n\rangle_{\mathcal{D}} = \int \prod_a d\boldsymbol{w}_r^a \mathbb{E}_{\{\boldsymbol{\psi}_\mu,\boldsymbol{\sigma}^\mu,\epsilon^\mu\}} \exp\left(-\frac{\beta M}{2\lambda}\sum_{\mu,a}\frac{1}{M}\left[\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\left(\boldsymbol{\psi}_\mu+\boldsymbol{\sigma}^\mu\right)-\boldsymbol{w}^{*\top}\boldsymbol{\psi}_\mu-\sqrt{M}(\epsilon^\mu-\xi_r^\mu)\right]^2 -\frac{\beta}{2}\sum_a|\boldsymbol{w}_r^a|^2-\frac{JM\beta}{2}\sum_a E_{rr}(\boldsymbol{w}^a)\right) \tag{B5}$$

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

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

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

$$\langle h_\mu^{ra}h_\nu^{rb}\rangle = \delta_{\mu\nu}Q_{ab}^{rr} \tag{B6}$$

$$\begin{aligned} Q_{ab}^{rr} = \frac{1}{M}\Bigg[&\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r-\boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{A}_r^\top\boldsymbol{w}_r^b-\boldsymbol{w}^*\right)\\ &+\frac{1}{\nu_{rr}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_r^\top\boldsymbol{w}_r^b+M(\zeta^2+\eta_r^2)\Bigg]\end{aligned} \tag{B7}$$

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

$$\boldsymbol{H}_\mu^r = \begin{bmatrix} h_\mu^{r1}\\ h_\mu^{r2}\\ \vdots\\ h_\mu^{rn}\end{bmatrix} \in \mathbb{R}^n \tag{B8}$$

$$\langle Z^n\rangle_{\mathcal{D}} = \int\prod_a d\boldsymbol{w}_r^a\mathbb{E}_{\{\psi_\mu,\boldsymbol{\sigma}^\mu,\boldsymbol{\epsilon}^\mu\}}\exp\left(-\frac{\beta}{2\lambda}\sum_\mu\boldsymbol{H}_\mu^{r\top}\boldsymbol{H}_\mu^r-\frac{\beta}{2}\sum_a|\boldsymbol{w}_r^a|^2-\frac{JM\beta}{2}\sum_a E_{rr}(\boldsymbol{w}^a)\right) \tag{B9}$$

Integrating over the $\boldsymbol{H}_\mu^r$ we get:

$$\langle Z^n\rangle_{\mathcal{D}} = \int\prod_a d\boldsymbol{w}_r^a\exp\left(-\frac{P}{2}\log\det\left(\boldsymbol{I}_n+\frac{\beta}{\lambda}\boldsymbol{Q}^{rr}\right)-\frac{\beta}{2}\sum_a|\boldsymbol{w}_r^a|^2-\frac{JM\beta}{2}\sum_a E_{rr}(\boldsymbol{w}_r)\right) \tag{B10}$$

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

$$\begin{aligned} 1 = \prod_{ab'}\int dQ_{ab}^{rr}\delta\Bigg(Q_{ab}^{rr}-\frac{1}{M}\Bigg[&\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r-\boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{A}_r^\top\boldsymbol{w}_r^b-\boldsymbol{w}^*\right)\\ &+\frac{1}{\nu_{rr}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_r^\top\boldsymbol{w}_r^b+M(\zeta^2+\eta_r^2)\Bigg]\Bigg)\end{aligned} \tag{B11}$$

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

$$\begin{aligned} 1 = \prod_{ab}\int\frac{1}{4\pi i/M}dQ_{ab}^{rr}d\hat{Q}_{ab}^{rr}\exp\Bigg(\frac{M}{2}\hat{Q}_{ab}^{rr}\Bigg(Q_{ab}^{rr}-\frac{1}{M}\Bigg[&\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r-\boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{A}_r^\top\boldsymbol{w}_r^b-\boldsymbol{w}^*\right)\\ &+\frac{1}{\nu_{rr}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_r^\top\boldsymbol{w}_r^b+M(\zeta^2+\eta_r^2)\Bigg]\Bigg)\Bigg)\end{aligned} \tag{B12}$$

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

$$\begin{aligned}
&\langle Z^n \rangle_{\mathcal{D}} \propto \\
&\int \prod_{ab} dQ_{ab}^{rr} d\hat{Q}_{ab}^{rr} \exp\left(-\frac{P}{2} \log\det\left(\boldsymbol{I}_n + \frac{\beta}{\lambda} \boldsymbol{Q}^{rr} \right) + \frac{1}{2} \sum_{ab} M \hat{Q}_{ab}^{rr} Q_{ab}^{rr} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr} - \zeta^2) \right) \\
&\int \prod_a d\boldsymbol{w}_r^a \exp\left(-\frac{\beta}{2} \sum_a |\boldsymbol{w}_r^a|^2 - \frac{1}{2} \sum_{ab} \hat{Q}_{ab}^{rr} \left[\left(\frac{1}{\sqrt{\nu_{rr}}} \boldsymbol{w}_r^{a\top} \boldsymbol{A}_r - \boldsymbol{w}^{*\top} \right) \boldsymbol{\Sigma}_s \left(\frac{1}{\sqrt{\nu_{rr}}} \boldsymbol{A}_r^\top \boldsymbol{w}_r^b - \boldsymbol{w}^* \right) \right. \right. \\
&\left. \left. + \frac{1}{\nu_{rr}} \boldsymbol{w}_r^{a\top} \boldsymbol{A}_r \boldsymbol{\Sigma}_0 \boldsymbol{A}_r^\top \boldsymbol{w}_r^b + M(\zeta^2 + \eta_r^2) \right] \right)
\end{aligned} \tag{B13}$$

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

$$\dot{\boldsymbol{w}_r} \equiv \begin{bmatrix} \boldsymbol{w}_r^1 \\ \vdots \\ \boldsymbol{w}_r^n \end{bmatrix} \tag{B14}$$

$$T^r \equiv \beta \boldsymbol{I}_n \otimes \boldsymbol{I}_{N_r} + \hat{\boldsymbol{Q}}^{rr} \otimes \left(\frac{1}{\nu_{rr}} \boldsymbol{A}_r (\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_0) \boldsymbol{A}_r^\top \right) \tag{B15}$$

$$V^r \equiv (\hat{\boldsymbol{Q}}^{rr} \otimes \boldsymbol{I}_{N_r})(\boldsymbol{1}_n \otimes \frac{1}{\sqrt{\nu_{rr}}} \boldsymbol{A}_r \boldsymbol{\Sigma}_s \boldsymbol{w}^*) \tag{B16}$$

We then have for the integral over w

$$\int d\dot{\boldsymbol{w}_r} \exp\left(-\frac{1}{2} \dot{\boldsymbol{w}_r}^\top T^r \dot{\boldsymbol{w}_{r'}} + V^{r\top} \dot{\boldsymbol{w}_r} \right) \tag{B17}$$

$$= \exp\left(\frac{1}{2} V^{r\top} (T^r)^{-1} V^r - \frac{1}{2} \log\det(T^r) \right) \tag{B18}$$

We can finally write the replicated partition function as:

$$\begin{aligned}
&\langle Z^n \rangle_{\mathcal{D}} \propto \\
&\int \prod_{ab} dQ_{ab}^{rr} d\hat{Q}_{ab}^{r} \exp\left(-\frac{P}{2} \log\det\left(\boldsymbol{I}_n + \frac{\beta}{\lambda} \boldsymbol{Q}^{rr} \right) + \frac{1}{2} \sum_{ab} M \hat{Q}_{ab}^{rr} Q_{ab}^{rr} - \frac{JM\beta}{2} \sum_a (Q_{aa}^{rr} - \zeta^2) \right) \\
&\exp\left(\frac{1}{2} V^{r\top} (T^r)^{-1} V^r - \frac{1}{2} \log\det(T^r) - \frac{1}{2} \sum_{ab} \hat{Q}_{ab}^{rr} (M(\zeta^2 + \eta_r^2) + \boldsymbol{w}^{*\top} \Sigma_s \boldsymbol{w}^*) \right)
\end{aligned} \tag{B19}$$

We now make the following replica-symmetric ansatz:

$$\beta Q_{ab}^{rr} = q \delta_{ab} + q_0 \tag{B20}$$

$$\beta^{-1} \hat{Q}_{ab}^{rr} = \hat{q} \delta_{ab} + \hat{q}_0 \tag{B21}$$

which is well-motivated because the loss function is convex. We may then rewrite the partition function as follows:

$$\langle Z^n \rangle_{\mathcal{D}} = \exp\left(-\frac{nM}{2} \mathfrak{g}\left[q, q_0, \hat{q}, \hat{q}_0 \right] \right) \tag{B22}$$

Where the effective action is written:

$$\mathfrak{g}\left[q,q_0,\hat{q},\hat{q}_0\right]=\alpha\left[\log(1+\frac{q}{\lambda})+\frac{q_0}{\lambda+q}\right]-(q\hat{q}+q\hat{q}_0+q_0\hat{q})+J\left[(q+q_0)-\beta\zeta^2\right]$$
$$-\frac{\beta}{\nu_{rr}M}\hat{q}^2\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\boldsymbol{w}^*+\frac{1}{M}\left[\log\det(\boldsymbol{G})+\hat{q}_0\,\mathrm{tr}[\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}}]\right]+\beta\hat{q}\left(\zeta^2+\eta_r^2+\frac{1}{M}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{w}^*\right) \tag{B23}$$

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

$$E_{rr}=\frac{\partial}{\partial J}\lim_{\beta\to\infty}\frac{1}{\beta}\mathfrak{g}[q,q_0,\hat{q},\hat{q}_0] \tag{B24}$$

Where g is evaluated at the values of $q,q_0,\hat{q},\hat{q}_0$ which minimize g, in accordance with the method of steepest descent.

$$\begin{aligned}E_{rr}&=\partial_J\left(\frac{1}{M}\boldsymbol{w}^{\star\top}\left[\hat{q}\boldsymbol{\Sigma}_s-\frac{1}{\nu_{rr}}\hat{q}^2\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right]\boldsymbol{w}^*+\hat{q}(\zeta^2+\eta_r^2)-J\zeta^2\right)\\&=[\partial_J\hat{q}]\left(\frac{1}{M}\boldsymbol{w}^{\star\top}\left[\boldsymbol{\Sigma}_s-\frac{2}{\nu_{rr}}\hat{q}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s+\frac{1}{\nu_{rr}}\hat{q}^2\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right]\boldsymbol{w}^*+\zeta^2+\eta_r^2\right)-\zeta^2\end{aligned} \tag{B25}$$

To complete the calculation, we need to find $\partial_J\hat{q}$. We may do this by examining two of the saddle-point equations:

$$\frac{\partial\mathfrak{g}}{\partial q_0}=0=\frac{\alpha}{\lambda+q}-\hat{q}+J \qquad \Rightarrow\hat{q}=\frac{\alpha}{\lambda+q}+J \tag{B26}$$
$$\frac{\partial\mathfrak{g}}{\partial\hat{q}_0}=0=-q+\frac{1}{M}\,\mathrm{tr}\left[\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}}\right] \qquad \Rightarrow q=\frac{1}{M}\,\mathrm{tr}\left[\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}}\right] \tag{B27}$$

These two equations may in principle be solved for the dominant values of q and $\hat{q}$. Letting $\kappa=\lambda+q$, we get:

$$\partial_J\hat{q}=-\frac{\alpha}{\kappa^2}\partial_J q+1 \tag{B28}$$
$$\partial_J q=-\frac{1}{M}\partial_J\hat{q}\,\mathrm{tr}\left[(\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}})^2\right] \tag{B29}$$

Solving this system of equations, we find $\partial_J\hat{q}=\frac{1}{1-\gamma}$ where $\gamma\equiv\frac{\alpha}{M\kappa^2}\,\mathrm{tr}\left[(\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}})^2\right]$

$$E_{rr}=\frac{1}{1-\gamma}\frac{1}{M}\boldsymbol{w}^{\star\top}\left[\boldsymbol{\Sigma}_s-\frac{2}{\nu_{rr}}\hat{q}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s+\frac{1}{\nu_{rr}}\hat{q}^2\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right]\boldsymbol{w}^*+\frac{\gamma\zeta^2+\eta_r^2}{1-\gamma} \tag{B30}$$
$$=\frac{1}{1-\gamma}\frac{1}{M}\boldsymbol{w}^{\star\top}\left[\boldsymbol{\Sigma}_s-\frac{1}{\nu_{rr}}\hat{q}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s-\frac{1}{\nu_{rr}}\hat{q}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}^{-2}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right]\boldsymbol{w}^*+\frac{\gamma\zeta^2+\eta_r^2}{1-\gamma} \tag{B31}$$

2. Off-Diagonal Terms

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

$$Z=\int d\boldsymbol{w}_r d\boldsymbol{w}_{r'}\exp\left(-\frac{\beta}{2\lambda}(E_r^t+E_{r'}^t)-\frac{JM\beta}{2}E_{rr'}(\boldsymbol{w}_r,\boldsymbol{w}_{r'})\right) \tag{B32}$$
$$\tag{B33}$$

$$\begin{aligned}\langle Z^n\rangle_{\mathcal{D}}=&\int\prod_a d\boldsymbol{w}_r^a d\boldsymbol{w}_{r'}^a\mathbb{E}_{\{\psi_\mu,\boldsymbol{\sigma}^\mu,\epsilon^\mu\}}\\&\exp\left(-\frac{\beta M}{2\lambda}\sum_{\mu,a}\frac{1}{M}\left[\left(h_\mu^{ra}\right)^2+\left(h_\mu^{r'a}\right)^2\right]-\frac{\beta}{2}\sum_a\left[|\boldsymbol{w}_r^a|^2+|\boldsymbol{w}_{r'}^a|^2\right]-\frac{JM\beta}{2}\sum_a E_{rr'}(\boldsymbol{w}_r^a,\boldsymbol{w}_{r'}^a)\right)\end{aligned} \tag{B34}$$

Where the h_μ^{ra} are defined as before. Next we must perform the averages over quenched disorder. We first integrate over $\{\boldsymbol{\psi}_\mu, \epsilon_\mu\}$. To do so, we note that the h_μ^{ra} are Gaussian random variables with covariance structure:

$$\langle h_\mu^{ra} h_\nu^{r'b} \rangle = \delta_{\mu\nu} Q_{ab}^{rr'} \tag{B35}$$

$$\begin{aligned} Q_{ab}^{rr'} &= \frac{1}{M}\left[\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r - \boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{r'r'}}}\boldsymbol{A}_{r'}^\top \boldsymbol{w}_{r'}^b - \boldsymbol{w}^*\right)\right. \\ &\left. + \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_{r'}^\top\boldsymbol{w}_{r'}^b + M\zeta^2\right] \end{aligned} \tag{B36}$$

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

$$\boldsymbol{H}_\mu = \begin{bmatrix} \boldsymbol{H}_\mu^r \\ \boldsymbol{H}_\mu^{r'} \end{bmatrix} \in \mathbb{R}^{2n} \tag{B37}$$

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} = &\int \prod_a d\boldsymbol{w}_r^a d\boldsymbol{w}_{r'}^a \mathbb{E}_{\{\psi_\mu, \boldsymbol{\sigma}^\mu, \epsilon^\mu\}} \\ &\exp\left(-\frac{\beta}{2\lambda}\sum_\mu \boldsymbol{H}_\mu^\top \boldsymbol{H}_\mu - \frac{\beta}{2}\sum_a\left[|\boldsymbol{w}_r^a|^2 + |\boldsymbol{w}_{r'}^a|^2\right] - \frac{JM\beta}{2}\sum_a E_{rr'}(\boldsymbol{w}_r^a, \boldsymbol{w}_{r'}^a)\right) \end{aligned} \tag{B38}$$

Integrating over $\boldsymbol{H}_\mu$ we get:

$$\begin{aligned} \langle Z^n \rangle_{\mathcal{D}} = &\int \prod_a d\boldsymbol{w}_r^a d\boldsymbol{w}_{r'}^a \\ &\exp\left(-\frac{P}{2}\log\det\left(\boldsymbol{I}_{2n} + \frac{\beta}{\lambda}\boldsymbol{Q}\right) - \frac{\beta}{2}\sum_a\left[|\boldsymbol{w}_r^a|^2 + |\boldsymbol{w}_{r'}^a|^2\right] - \frac{JM\beta}{2}\sum_a E_{rr}(\boldsymbol{w}_r^a, \boldsymbol{w}_{r'}^a)\right) \end{aligned} \tag{B39}$$

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

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}^{rr} & \boldsymbol{Q}^{rr'} \\ \boldsymbol{Q}^{rr'} & \boldsymbol{Q}^{r'r'} \end{bmatrix} \tag{B40}$$

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

$$\begin{aligned} 1 = \prod_{ab}\int dQ_{ab}^{rr'}\,\delta\left(Q_{ab}^{rr'} - \frac{1}{M}\left[\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r - \boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{r'r'}}}\boldsymbol{A}_{r'}^\top\boldsymbol{w}_{r'}^b - \boldsymbol{w}^*\right)\right.\right. \\ \left.\left. + \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_{r'}^\top\boldsymbol{w}_{r'}^b + M\zeta^2\right]\right) \end{aligned} \tag{B41}$$

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

$$\begin{aligned} 1 = \prod_{ab}\int \frac{1}{4\pi i/M} dQ_{ab}^{rr'} d\hat{Q}_{ab}^{rr'} \exp\left(\frac{M}{2}\hat{Q}_{ab}^{rr'}\left(Q_{ab}^{rr'} - \frac{1}{M}\left[\left(\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r - \boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{r'r'}}}\boldsymbol{A}_{r'}^\top\boldsymbol{w}_{r'}^b - \boldsymbol{w}^*\right)\right.\right.\right. \\ \left.\left.\left. + \frac{1}{\nu_{rr}}\boldsymbol{w}_r^{a\top}\boldsymbol{A}_r\boldsymbol{\Sigma}_0\boldsymbol{A}_{r'}^\top\boldsymbol{w}_{r'}^b + M\zeta^2\right]\right)\right) \end{aligned} \tag{B42}$$

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

$$\langle Z^n\rangle_{\mathcal{D}} \propto \int \prod_{abr_1r_2} dQ_{ab}^{r_1r_2} d\hat{Q}_{ab}^{r_1r_2}$$
$$\exp\left(-\frac{P}{2}\log\det\left(\boldsymbol{I}_{2n}+\frac{\beta}{\lambda}\boldsymbol{Q}\right)+\frac{1}{2}\sum_{abr_1r_2} M\hat{Q}_{ab}^{r_1r_2}Q_{ab}^{r_1r_2}-\frac{JM\beta}{2}\sum_a(Q_{aa}^{rr'}-\zeta^2)\right)$$
$$\int\prod_a d\boldsymbol{w}_r^a d\boldsymbol{w}_{r'}^a \exp\left(-\frac{\beta}{2}\sum_a\left[|\boldsymbol{w}_r^a|^2+|\boldsymbol{w}_{r'}^a|^2\right]-\frac{1}{2}\sum_{abr_1r_2}\hat{Q}_{ab}^{r_1r_2}\left[\left(\frac{1}{\sqrt{\nu_{r_1}}}\boldsymbol{w}_{r_1}^{a\top}\boldsymbol{A}_{r_1}-\boldsymbol{w}^{*\top}\right)\boldsymbol{\Sigma}_s\left(\frac{1}{\sqrt{\nu_{r_2}}}\boldsymbol{A}_{r_2}^{\top}\boldsymbol{w}_{r_2}^b-\boldsymbol{w}^*\right)\right.\right.$$
$$\left.\left.+\frac{1}{\sqrt{\nu_{r_1}\nu_{r_2}}}\boldsymbol{w}_{r_1}^{a\top}\boldsymbol{A}_{r_1}\boldsymbol{\Sigma}_0\boldsymbol{A}_{r_2}^{\top}\boldsymbol{w}_{r_2}^b+M\zeta^2\right]\right) \tag{B43}$$

Where sums over r_1 and r_2 run over $\{r, r'\}$.
In order to perform the Gaussian integral over the $\{\boldsymbol{w}_r^a\}$, we unfold in two steps. We first define the following:

$$\boldsymbol{w}_r^{\cdot} \equiv \begin{bmatrix}\boldsymbol{w}_r^1\\ \vdots\\ \boldsymbol{w}_r^n\end{bmatrix} \tag{B44}$$
$$[\hat{\boldsymbol{Q}}^{rr'}]_{ab} \equiv \hat{Q}_{ab}^{rr'} \tag{B45}$$
$$\tilde{\boldsymbol{\Sigma}}_{rr'} \equiv \frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\boldsymbol{A}_r[\boldsymbol{\Sigma}_s+\boldsymbol{\Sigma}_0]\boldsymbol{A}_{r'}^{\top} \tag{B46}$$
$$T^{rr'} \equiv \beta\delta_{rr'}\boldsymbol{I}_n\otimes\boldsymbol{I}_{N_r}+\hat{\boldsymbol{Q}}^{rr'}\otimes\tilde{\boldsymbol{\Sigma}}_{rr'} \tag{B47}$$

Unfolding over the replica indices, we then get:

$$\begin{aligned}&\langle Z^n\rangle_{\mathcal{D}} \propto \int\prod_{abr_1r_2} dQ_{ab}^{r_1r_2}d\hat{Q}_{ab}^{r_1r_2}\\ &\exp\left(-\frac{P}{2}\log\det\left(\boldsymbol{I}_{2n}+\frac{\beta}{\lambda}\boldsymbol{Q}\right)+\frac{1}{2}\sum_{abr_1r_2}M\hat{Q}_{ab}^{r_1r_2}Q_{ab}^{r_1r_2}-\frac{JM\beta}{2}\sum_a(Q_{aa}^{rr'}-\zeta^2)\right)\\ &\exp\left(-\frac{1}{2}\sum_{abr_1r_2}\hat{Q}_{ab}^{r_1r_2}\left(\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{w}^*+M\zeta^2\right)\right)\\ &\int d\boldsymbol{w}_r^{\cdot}d\boldsymbol{w}_{r'}^{\cdot}\exp\left(-\frac{1}{2}\sum_{r_1r_2}\boldsymbol{w}_{r_1}^{\cdot\top}T^{r_1r_2}\boldsymbol{w}_{r_2}+\sum_{r_1r_2}\left[(\hat{\boldsymbol{Q}}^{r_1r_2}\otimes\boldsymbol{I}_{N_{r_1}})(\boldsymbol{1}_n\otimes\frac{1}{\sqrt{\nu_{r_1}}}\boldsymbol{A}_{r_1}\boldsymbol{\Sigma}_s\boldsymbol{w}^*)\right]^{\top}\boldsymbol{w}_{r_1}\right)\end{aligned} \tag{B48}$$

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:

$$\boldsymbol{w} \equiv \begin{bmatrix}\boldsymbol{w}_r^{\cdot}\\ \boldsymbol{w}_{r'}^{\cdot}\end{bmatrix} \tag{B49}$$
$$T \equiv \begin{bmatrix}T^{rr} & T^{rr'}\\ T^{r'r} & T^{r'r'}\end{bmatrix} \tag{B50}$$
$$V \equiv \begin{bmatrix}\left((\hat{\boldsymbol{Q}}^{rr}+\hat{\boldsymbol{Q}}^{rr'})\otimes\boldsymbol{I}_{N_r}\right)\left(\boldsymbol{1}_n\otimes\frac{1}{\sqrt{\nu_{rr}}}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\boldsymbol{w}^*\right)\\ \left((\hat{\boldsymbol{Q}}^{r'r'}+\hat{\boldsymbol{Q}}^{r'r})\otimes\boldsymbol{I}_{N_{r'}}\right)\left(\boldsymbol{1}_n\otimes\frac{1}{\sqrt{\nu_{r'r'}}}\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\boldsymbol{w}^*\right)\end{bmatrix} \tag{B51}$$

The integral over w then becomes:

$$\int d\boldsymbol{w}\exp\left(-\frac{1}{2}\boldsymbol{w}^{\top}T\boldsymbol{w}+V^{\top}\boldsymbol{w}\right)\propto\exp\left(\frac{1}{2}V^{\top}T^{-1}V-\frac{1}{2}\log\det T\right) \tag{B52}$$

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:

$$\beta Q_{ab}^{r_1 r_2} = Q^{r_1 r_2}\delta_{ab} + q^{r_1 r_2} \tag{B53}$$
$$\beta^{-1}\hat{Q}_{ab}^{r_1 r_2} = \hat{Q}^{r_1 r_2}\delta_{ab} + \hat{q}^{r_1 r_2} \tag{B54}$$

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

$$\beta \boldsymbol{Q}^{rr} = q\boldsymbol{I} + Q\mathbf{1}\mathbf{1}^\top \tag{B55}$$
$$\beta \boldsymbol{Q}^{r'r'} = r\boldsymbol{I} + R\mathbf{1}\mathbf{1}^\top \tag{B56}$$
$$\beta \boldsymbol{Q}^{rr'} = c\boldsymbol{I} + C\mathbf{1}\mathbf{1}^\top \tag{B57}$$
$$\beta^{-1} \hat{\boldsymbol{Q}}^{rr} = \hat{q}\boldsymbol{I} + \hat{Q}\mathbf{1}\mathbf{1}^\top \tag{B58}$$
$$\beta^{-1} \hat{\boldsymbol{Q}}^{r'r'} = \hat{r}\boldsymbol{I} + \hat{R}\mathbf{1}\mathbf{1}^\top \tag{B59}$$
$$\beta^{-1} \hat{\boldsymbol{Q}}^{rr'} = \hat{c}\boldsymbol{I} + \hat{C}\mathbf{1}\mathbf{1}^\top \tag{B60}$$

In order to simplify $\log\det\left(\lambda \boldsymbol{I}_{2n} + \beta \boldsymbol{Q}\right)$, we note that this is a symmetric 2-by-2-block matrix, where each block commutes with all other blocks. We may then use [44]'s result to simplify.

$$\log\det\left(\lambda \boldsymbol{I}_{2n} + \beta \boldsymbol{Q}\right) = n\left[\log\left((\lambda+q)(\lambda+r) - c^2\right) + \frac{(\lambda+q)R + (\lambda+r)Q - 2cC}{(\lambda+q)(\lambda+r) - c^2}\right] + \mathcal{O}(n^2) \tag{B61}$$

$$\sum_{abr_1r_2} \hat{\boldsymbol{Q}}_{ab}^{r_1r_2}\boldsymbol{Q}_{ab}^{r_1r_2} = n\left[\left(q\hat{q} + \hat{q}Q + q\hat{Q}\right) + \left(r\hat{r} + \hat{r}R + r\hat{R}\right) + 2\left(c\hat{c} + \hat{c}C + c\hat{C}\right)\right] + \mathcal{O}(n^2) \tag{B62}$$

$$\sum_a \left(\boldsymbol{Q}_{aa}^{rr'} - \zeta^2\right) = n\left[\frac{1}{\beta}(c+C) - \zeta^2\right] + \mathcal{O}(n^2) \tag{B63}$$

$$\sum_{abr_1r_2} \hat{\boldsymbol{Q}}_{ab}^{r_1r_2} = \beta n\left[\hat{q} + \hat{r} + 2\hat{c}\right] \tag{B64}$$

$$\begin{aligned}\log\det(T) = n\Bigg[&\log(\beta) + \log\det\begin{bmatrix}\boldsymbol{G}_{rr} & \boldsymbol{G}_{rr'}\\ \boldsymbol{G}_{r'r} & \boldsymbol{G}_{r'r'}\end{bmatrix}\\ &+\hat{q}\,\mathrm{tr}(\boldsymbol{D}_{rr}\tilde{\boldsymbol{\Sigma}}_{rr} + \hat{c}\,\mathrm{tr}(\boldsymbol{D}_{rr'}\tilde{\boldsymbol{\Sigma}}_{r'r}) + \hat{r}\,\mathrm{tr}(\boldsymbol{D}_{r'r'}\tilde{\boldsymbol{\Sigma}}_{r'r'}) + \hat{c}\,\mathrm{tr}(\boldsymbol{D}_{r'r}\tilde{\boldsymbol{\Sigma}}_{rr'})\Bigg] + \mathcal{O}(n^2)\end{aligned} \tag{B65}$$

$$\text{where } \boldsymbol{G}_{rr} = \boldsymbol{I}_{N_r} + \hat{q}\tilde{\boldsymbol{\Sigma}}_{rr} \quad \boldsymbol{G}_{r'r'} = \boldsymbol{I}_{N_{r'}} + \hat{r}\tilde{\boldsymbol{\Sigma}}_{r'r'} \quad \boldsymbol{G}_{rr'} = \hat{c}\tilde{\boldsymbol{\Sigma}}_{rr'} \quad \boldsymbol{G}_{r'r} = \hat{c}\tilde{\boldsymbol{\Sigma}}_{r'r} \tag{B66}$$

and where the $\boldsymbol{D}_{r_1r_2}$ matrices are defined implicitly through the following equation:

$$\begin{bmatrix}\boldsymbol{G}_{rr} & \boldsymbol{G}_{rr'}\\ \boldsymbol{G}_{r'r} & \boldsymbol{G}_{r'r'}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{D}_{rr} & \boldsymbol{D}_{rr'}\\ \boldsymbol{D}_{r'r} & \boldsymbol{D}_{r'r'}\end{bmatrix} \tag{B67}$$

The $\boldsymbol{D}_{r_1r_2}$ matrices can be expressed in terms of the $\boldsymbol{G}_{r_1r_2}$ and their inverses by the standard 2×2 block matrix inversion formula (see, for example, [45]). Applying this formula gives the following:

$$\boldsymbol{D}_{rr} = \left[\boldsymbol{I} + \hat{q}\tilde{\boldsymbol{\Sigma}}_{rr} - \hat{c}^2\tilde{\boldsymbol{\Sigma}}_{rr'}\left(\boldsymbol{I} + \hat{r}\tilde{\boldsymbol{\Sigma}}_{r'r'}\right)^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r}\right]^{-1} \tag{B68}$$

$$\boldsymbol{D}_{r'r'} = \left[\boldsymbol{I} + \hat{r}\tilde{\boldsymbol{\Sigma}}_{r'r'} - \hat{c}^2\tilde{\boldsymbol{\Sigma}}_{r'r}\left(\boldsymbol{I} + \hat{q}\tilde{\boldsymbol{\Sigma}}_{rr}\right)^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\right]^{-1} \tag{B69}$$

$$\boldsymbol{D}_{rr'} = -\hat{c}\boldsymbol{D}_{rr}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}^{-1}_{r'r'} \tag{B70}$$

$$\boldsymbol{D}_{r'r} = -\hat{c}\boldsymbol{D}_{r'r'}\tilde{\boldsymbol{\Sigma}}_{r'r}\boldsymbol{G}^{-1}_{rr} \tag{B71}$$

$$V^\top T^{-1} V = n\beta \boldsymbol{w}^{*\top}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{G}_{rr} & \boldsymbol{G}_{rr'} \\ \boldsymbol{G}_{r'r} & \boldsymbol{G}_{r'r'}\end{bmatrix}^{-1}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}\boldsymbol{w}^* + \mathcal{O}(n^2) \tag{B72}$$

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

$$\langle Z^n\rangle_{\mathcal{D}} = \exp\left(-\frac{nM}{2}\mathfrak{g}\left[q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C}\right]\right) \tag{B73}$$

Where the effective action is written:

$$\mathfrak{g}\left[q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C}\right] = \tag{B74}$$

$$\alpha\left[\log\left((\lambda+q)(\lambda+r) - c^2\right) + \frac{(\lambda+q)R + (\lambda+r)Q - 2Gg}{(\lambda+q)(\lambda+r) - c^2}\right] \tag{B75}$$

$$-\left[\left(q\hat{q} + \hat{q}Q + q\hat{Q}\right) + \left(r\hat{r} + \hat{r}R + r\hat{R}\right) + 2\left(c\hat{c} + \hat{c}C + c\hat{C}\right)\right] \tag{B76}$$

$$+ J(c+C) - \beta J\zeta^2 \tag{B77}$$

$$+ \beta\left[\hat{q} + \hat{r} + 2\hat{c}\right]\left(\frac{1}{M}\boldsymbol{w}^{*\top}\Sigma\boldsymbol{w}^* + \zeta^2\right) \tag{B78}$$

$$-\beta\boldsymbol{w}^{*\top}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}^\top \boldsymbol{G}^{-1}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}\boldsymbol{w}^* \tag{B79}$$

$$+\frac{1}{M}\left[\log(\beta) + \log\det(\boldsymbol{G}) + \hat{Q}\,\mathrm{tr}(\boldsymbol{D}_{rr}\tilde{\boldsymbol{\Sigma}}_{rr}) + \hat{C}\,\mathrm{tr}(\boldsymbol{D}_{rr'}\tilde{\boldsymbol{\Sigma}}_{r'r}) + \hat{R}\,\mathrm{tr}(\boldsymbol{D}_{r'r'}\tilde{\boldsymbol{\Sigma}}_{r'r'}) + \hat{C}\,\mathrm{tr}(\boldsymbol{D}_{r'r}\tilde{\boldsymbol{\Sigma}}_{rr'})\right] \tag{B80}$$

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

We have:

$$E_{rr'} = \frac{\partial}{\partial J}\lim_{\beta\to\infty}\frac{1}{\beta}\mathfrak{g}\left[q, Q, r, R, c, C, \hat{q}, \hat{Q}, \hat{r}, \hat{R}, \hat{c}, \hat{C}\right] \tag{B81}$$

Where g is evaluated at the values of $Q, q, R, r, C, c, \hat{Q}, \hat{q}, \hat{R}, \hat{r}, \hat{C}, \hat{c}$ which minimize $\mathfrak{g}$, in accordance with the method of steepest descent (and thus implicitly depend on J). This gives:

$$E_{rr'} = -\zeta^2 + \left[\frac{\partial\hat{q}}{\partial J} + \frac{\partial\hat{r}}{\partial J} + 2\frac{\partial\hat{c}}{\partial J}\right]\left(\frac{1}{M}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{w}^* + \zeta^2\right) \tag{B82}$$

$$-\frac{2}{M}\boldsymbol{w}^{*\top}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\frac{\partial\hat{q}}{\partial J} + \frac{\partial\hat{c}}{\partial J})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\frac{\partial\hat{r}}{\partial J} + \frac{\partial\hat{c}}{\partial J})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}^\top \boldsymbol{G}^{-1}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}\boldsymbol{w}^* \tag{B83}$$

$$+\frac{1}{M}\boldsymbol{w}^{*\top}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}^\top \boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial J}\boldsymbol{G}^{-1}\begin{bmatrix}\frac{1}{\sqrt{\nu_{rr}}}(\hat{q}+\hat{c})\boldsymbol{A}_r\boldsymbol{\Sigma}_s \\ \frac{1}{\sqrt{\nu_{r'r'}}}(\hat{r}+\hat{c})\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\end{bmatrix}\boldsymbol{w}^* \tag{B84}$$

$$\frac{\partial \boldsymbol{G}}{\partial J} = \frac{\partial \boldsymbol{G}}{\partial \hat{q}}\frac{\partial \hat{q}}{\partial J} + \frac{\partial \boldsymbol{G}}{\partial \hat{r}}\frac{\partial \hat{r}}{\partial J} + \frac{\partial \boldsymbol{G}}{\partial \hat{c}}\frac{\partial \hat{c}}{\partial J} = \begin{bmatrix} \frac{\partial \hat{q}}{\partial J}\tilde{\boldsymbol{\Sigma}}_{rr} & \frac{\partial \hat{c}}{\partial J}\tilde{\boldsymbol{\Sigma}}_{rr'} \\ \frac{\partial \hat{c}}{\partial J}\tilde{\boldsymbol{\Sigma}}_{r'r} & \frac{\partial \hat{r}}{\partial J}\tilde{\boldsymbol{\Sigma}}_{r'r'} \end{bmatrix} \tag{B85}$$

All that remains is to calculate the saddle-point values of $\hat{q}, \hat{r}, \hat{c}$ and their derivatives with respect to J at $J = 0$.

$$\frac{\partial \mathfrak{g}}{\partial Q} = 0 = \frac{\alpha(\lambda + r)}{(\lambda + q)(\lambda + r) - C^2} - \hat{q} \tag{B86}$$

$$\frac{\partial \mathfrak{g}}{\partial \hat{Q}} = 0 = -q + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{D}_{rr}\tilde{\boldsymbol{\Sigma}}_{rr}\right) \tag{B87}$$

$$\frac{\partial \mathfrak{g}}{\partial R} = 0 = \frac{\alpha(\lambda + q)}{(\lambda + q)(\lambda + r) - C^2} - \hat{r} \tag{B88}$$

$$\frac{\partial \mathfrak{g}}{\partial \hat{R}} = 0 = -r + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{D}_{r'r'}\tilde{\boldsymbol{\Sigma}}_{r'r'}\right) \tag{B89}$$

$$\frac{\partial \mathfrak{g}}{\partial g} = 0 = \frac{-2\alpha C}{(\lambda + q)(\lambda + r) - C^2} - 2\hat{C} + J \tag{B90}$$

$$\frac{\partial \mathfrak{g}}{\partial \hat{c}} = 0 = -2C + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{D}_{rr'}\tilde{\boldsymbol{\Sigma}}_{r'r} + \boldsymbol{D}_{r'r}\tilde{\boldsymbol{\Sigma}}_{rr'}\right) \tag{B91}$$

These 6 equations can in principle be solved for $\{q, r, c, \hat{q}, \hat{r}, \hat{c}\}$ and their derivatives with respect to J. Note that when $J = 0$, the saddle point equations B90, B91 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:

$$0 = \frac{\alpha}{(\lambda + q)} - \hat{q} \tag{B92}$$

$$0 = -q + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr}\right) \tag{B93}$$

and for readout r':

$$0 = \frac{\alpha}{(\lambda + r)} - \hat{r} \tag{B94}$$

$$0 = -r + \frac{1}{M}\operatorname{tr}\left(\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r'}\right) \tag{B95}$$

These are equivalent to the saddle-point equations for a single readout given in equation B87, B86 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 $C > 0$). Differentiating the saddle point equations with respect to J will give a 6 by 6 system of equations for the derivatives of the order parameters. With foresight, we first calculate $\partial_J \boldsymbol{D}$

$$\partial_J \boldsymbol{D} = \partial_J \boldsymbol{G}^{-1} = -\boldsymbol{G}^{-1}(\partial_J \boldsymbol{G})\boldsymbol{G}^{-1} \tag{B96}$$

Evaluated at $J = 0$, we have the following:

$$\partial_J \boldsymbol{D}_{rr} = -\partial_J \hat{q}\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr}\boldsymbol{G}_{rr}^{-1} \tag{B97}$$

$$\partial_J \boldsymbol{D}_{r'r'} = -\partial_J \hat{r}\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r'}\boldsymbol{G}_{r'r'}^{-1} \tag{B98}$$

$$\partial_J \boldsymbol{D}_{rr'} = -\partial_J \hat{c}\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'r'}^{-1} \tag{B99}$$

$$\partial_J \boldsymbol{D}_{r'r'} = -\partial_J \hat{c}\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r}\boldsymbol{G}_{rr}^{-1} \tag{B100}$$

Differentiating equations B86, B87, B88, B89, B90, B91 and evaluating at $J, c, \hat{c} = 0$ we get:

$$0 = \partial_J \hat{q} + \frac{\alpha}{(\lambda+q)^2}\partial_J q \tag{B101}$$

$$0 = \partial_J q + \partial_J \hat{q}\frac{1}{M}\,\mathrm{tr}\left[\left(\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr}\right)^2\right] \tag{B102}$$

$$0 = \partial_J \hat{r} + \frac{\alpha}{(\lambda+r)^2}\partial_J r \tag{B103}$$

$$0 = \partial_J r + \partial_J \hat{r}\frac{1}{M}\,\mathrm{tr}\left[\left(\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r'}\right)^2\right] \tag{B104}$$

$$\frac{1}{2} = \partial_J \hat{c} + \frac{\alpha}{(\lambda+q)(\lambda+r)}\partial_J c \tag{B105}$$

$$0 = \partial_J c + \partial_J \hat{c}\frac{1}{M}\,\mathrm{tr}\left[\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r}\right] \tag{B106}$$

Solving these equations, we obtain:

$$\partial_J \hat{q} = 0 \tag{B107}$$

$$\partial_J \hat{r} = 0 \tag{B108}$$

$$\partial_J \hat{c} = \frac{1}{2(1-\gamma)} \tag{B109}$$

$$\text{where } \gamma \equiv \frac{\alpha}{(\lambda+q)(\lambda+r)}\frac{1}{M}\,\mathrm{tr}\left[\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'r'}^{-1}\tilde{\boldsymbol{\Sigma}}_{r'r}\right] \tag{B110}$$

We may then simplify the expression for the generalization error as follows:

$$\begin{aligned} E_{rr'} = {} & \frac{\gamma}{1-\gamma}\zeta^2 + \frac{1}{1-\gamma}\left(\frac{1}{M}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{w}^*\right) \\ & - \frac{1}{M(1-\gamma)}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\left[\frac{1}{\nu_{rr}}\hat{q}\boldsymbol{A}_r^\top\boldsymbol{G}_{rr}^{-1}\boldsymbol{A}_r + \frac{1}{\nu_{r'r'}}\hat{r}\boldsymbol{A}_{r'}^\top\boldsymbol{G}_{r'r'}^{-1}\boldsymbol{A}_{r'}\right]\boldsymbol{\Sigma}_s\boldsymbol{w}^* \\ & + \frac{1}{M(1-\gamma)}\hat{q}\hat{r}\frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\boldsymbol{w}^{*\top}\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_{rr}^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'r'}^{-1}\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\boldsymbol{w}^* \end{aligned} \tag{B111}$$

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

Appendix C: 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:

$$\boldsymbol{\Sigma}_s = s\left[(1-c)\boldsymbol{I}_M + c\mathbf{1}_M\mathbf{1}_M^\top\right] \tag{C1}$$

$$\boldsymbol{\Sigma}_0 = \omega\boldsymbol{I}_M \tag{C2}$$

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

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 $\mathbf{1}_M$:

$$\boldsymbol{w}^* = \sqrt{1-\rho^2}\mathbb{P}_\perp \boldsymbol{w}_0^* + \rho \mathbf{1}_M \tag{C3}$$

$$\boldsymbol{w}_0^* \sim \mathcal{N}(0, \boldsymbol{I}_M) \tag{C4}$$

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

$$\langle E_{rr'} \rangle = (1-\rho^2)E_{rr'}(\rho=0) + \rho^2 E_{rr'}(\rho=1) \tag{C5}$$

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 $a\delta_{MN}\boldsymbol{I}_M + b\mathbf{1}_M\mathbf{1}_N^\top$, 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.

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 $\boldsymbol{A}_r = \begin{pmatrix}\boldsymbol{I}_{N_r} & \mathbf{0}\end{pmatrix} \in \mathbb{R}^{N_r \times M}$ where $\mathbf{0}$ denotes a matrix of all zeros, of the appropriate dimensionality.

We start by simplifying the saddle-point equations B87,B86. Expanding $\frac{1}{M}\operatorname{tr}\left(\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr}\right)$ and keeping only leading order terms, the saddle-point equations for q_r and $\hat{q}_r$ reduce to:

$$q_r = \frac{\nu_{rr}\left(s(1-c)+\omega\right)}{\hat{q}_r(s(1-c)+\omega) + \nu_r} \tag{C6}$$

$$\hat{q}_r = \frac{\alpha}{\lambda + q_r} \tag{C7}$$

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

$$q_r = \frac{\sqrt{a^2\alpha^2 + 2a\alpha(\lambda-a)\nu_r + (a+\lambda)^2\nu_r^2} - a\alpha + (a-\lambda)\nu_r}{2\nu_r} \tag{C8}$$

$$\hat{q}_r = \frac{\sqrt{a^2\alpha^2 + 2a\alpha(\lambda-a)\nu_r + (a+\lambda)^2\nu_r^2} + a\alpha - (a+\lambda)\nu_r}{2a\lambda} \tag{C9}$$

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 equatios can be applied to each of the k readouts.

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

$$\gamma_{rr} = \frac{a^2\alpha\nu_r}{\left(\lambda+q_r\right)^2\left(a\hat{q}_r+\nu_r\right)^2} \tag{C10}$$

$$\begin{aligned}\langle E_{rr}\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=0) =& \frac{1}{1-\gamma_{rr}}\frac{1}{M}\operatorname{tr}\left[\mathbb{P}_\perp\left(\boldsymbol{\Sigma}_s - \frac{2}{\nu_{rr}}\hat{q}_r\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s + \frac{1}{\nu_{rr}}\hat{q}_r^2\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{G}_r^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right)\mathbb{P}_\perp\right] \\ &+ \frac{\gamma_{rr}}{1-\gamma_{rr}}\zeta^2 + \eta_r^2,\end{aligned} \tag{C11}$$

$$\begin{aligned}\langle E_{rr}\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=1) =& \frac{1}{1-\gamma_{rr}}\frac{1}{M}\mathbf{1}_M^\top\left[\boldsymbol{\Sigma}_s - \frac{2}{\nu_{rr}}\hat{q}_r\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s + \frac{1}{\nu_{rr}}\hat{q}_r^2\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{G}_r^{-1}\boldsymbol{A}_r\boldsymbol{\Sigma}_s\right]\mathbf{1}_m \\ &+ \frac{\gamma_{rr}}{1-\gamma_{rr}}\zeta^2 + \eta_r^2,\end{aligned} \tag{C12}$$

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:

$$\langle E_{rr}\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=0) = \frac{1}{1-\gamma_{rr}}\left(s(1-c)\left(1-\frac{(1-c)s\hat{q}_r\nu_r\left(\hat{q}_r(s(1-c)+\omega)+2\nu_r\right)}{\left(\hat{q}_r(s(1-c)+\omega)+\nu_r\right)^2}\right)\right)+\frac{\gamma_{rr}\zeta^2+\eta_r^2}{1-\gamma_{rr}} \tag{C13}$$

$$\langle E_{rr}\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=1) = \frac{1}{1-\gamma_{rr}}\left(\frac{s(1-c)(1-\nu_{rr})+\omega}{\nu_{rr}}\right)+\frac{\gamma_{rr}\zeta^2+\eta_r^2}{1-\gamma_{rr}} \tag{C14}$$

In the "ridgeless" limit where $\lambda \to 0$, we obtain:

$$\gamma_{rr} = \frac{4\alpha\nu_{rr}}{(\alpha+\nu_{rr}+|\alpha-\nu_{rr}|)^2} \tag{C15}$$

$$\langle E_{rr}(\rho=0)\rangle_{\mathcal{D},\boldsymbol{w}^*} = \begin{cases} \frac{s(1-c)\nu_{rr}}{\nu_{rr}-\alpha}\left(1+\frac{s\alpha(1-c)(\alpha-2\nu_{rr})}{\nu_{rr}[s(1-c)+\omega]}\right)+\frac{\alpha\zeta^2+\nu_{rr}\eta_r^2}{\nu_{rr}-\alpha}, & \text{if } \alpha<\nu_{rr} \\ \frac{s(1-c)\alpha}{\alpha-\nu_{rr}}\left(1-\frac{s(1-c)\nu_{rr}}{s(1-c)+\omega}\right)+\frac{\nu_{rr}\zeta^2+\alpha\eta_r^2}{\alpha-\nu_{rr}}, & \text{if } \alpha>\nu_{rr} \end{cases} \quad (\lambda\to 0) \tag{C16}$$

$$\langle E_{rr}(\rho=1)\rangle_{\mathcal{D},\boldsymbol{w}^*} = \begin{cases} \frac{\nu_{rr}}{\nu_{rr}-\alpha}\left(\frac{s(1-c)(1-\nu_{rr})+\omega}{\nu_{rr}}\right)+\frac{\alpha\zeta^2+\nu_{rr}\eta_r^2}{\nu_{rr}-\alpha}, & \text{if } \alpha<\nu_{rr} \\ \frac{\alpha}{\alpha-\nu_{rr}}\left(\frac{s(1-c)(1-\nu_{rr})+\omega}{\nu_{rr}}\right)+\frac{\nu_{rr}\zeta^2+\alpha\eta_r^2}{\alpha-\nu_{rr}}, & \text{if } \alpha>\nu_{rr} \end{cases} \quad (\lambda\to 0) \tag{C17}$$

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:

$$\boldsymbol{A}_r = \begin{pmatrix} \boldsymbol{I}_{n_r\times n_r} & \boldsymbol{0}_{n_r\times n_{r'}} & \boldsymbol{0}_{n_r\times n_s} & \boldsymbol{0}_{n_r\times l} \\ \boldsymbol{0}_{n_s\times n_r} & \boldsymbol{0}_{n_s\times n_{r'}} & \boldsymbol{I}_{n_s\times n_s} & \boldsymbol{0}_{n_s\times l} \end{pmatrix} \in \mathbb{R}^{N_r\times M} \tag{C18}$$

$$\boldsymbol{A}_{r'} = \begin{pmatrix} \boldsymbol{0}_{n_{r'}\times n_r} & \boldsymbol{I}_{n_{r'}\times n_{r'}} & \boldsymbol{0}_{n_{r'}\times n_s} & \boldsymbol{0}_{n_{r'}\times l} \\ \boldsymbol{0}_{n_s\times n_r} & \boldsymbol{0}_{n_s\times n_{r'}} & \boldsymbol{I}_{n_s\times n_s} & \boldsymbol{0}_{n_s\times l} \end{pmatrix} \in \mathbb{R}^{N_{r'}\times M} \tag{C19}$$

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:

$$\gamma_{rr'} = \frac{\alpha\nu_{rr'}(s(1-c)+\omega)^2}{(\lambda+q_r)(\lambda+q_{r'})\left(\nu_{rr}+(s(1-c)+\omega)\hat{q}_r\right)\left(\nu_{r'r'}+(s(1-c)+\omega)\hat{q}_{r'}\right)} \tag{C20}$$

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

$$\begin{aligned} \langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=0) =& \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2 + \frac{1}{1-\gamma_{rr'}}\left(\frac{1}{M}\operatorname{tr}\left[\mathbb{P}_\perp\boldsymbol{\Sigma}_s\mathbb{P}_\perp\right]\right) \\ &- \frac{1}{M(1-\gamma_{rr'})}\operatorname{tr}\left[\mathbb{P}_\perp\boldsymbol{\Sigma}_s\left(\frac{1}{\nu_{rr}}\hat{q}_r\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\boldsymbol{A}_r+\frac{1}{\nu_{r'r'}}\hat{q}_{r'}\boldsymbol{A}_{r'}^\top\boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\right)\boldsymbol{\Sigma}_s\mathbb{P}_\perp\right] \\ &+ \frac{\hat{q}_r\hat{q}_{r'}}{M(1-\gamma_{rr'})}\frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\operatorname{tr}\left[\mathbb{P}_\perp\boldsymbol{\Sigma}_s\boldsymbol{A}_r^\top\boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s\mathbb{P}_\perp\right], \end{aligned} \tag{C21}$$

$$\begin{aligned}\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=1) =& \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2 + \frac{1}{M(1-\gamma_{rr'})}\left(\mathbf{1}_M^\top \boldsymbol{\Sigma}_s \mathbf{1}_M\right) \\ &- \frac{1}{M(1-\gamma_{rr'})}\mathbf{1}_M^\top \boldsymbol{\Sigma}_s \left(\frac{1}{\nu_{rr}}\hat{q}_r \boldsymbol{A}_r^\top \boldsymbol{G}_r^{-1}\boldsymbol{A}_r + \frac{1}{\nu_{r'r'}}\hat{q}_{r'}\boldsymbol{A}_{r'}^\top \boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\right)\boldsymbol{\Sigma}_s \mathbf{1}_M^\top \\ &+ \frac{\hat{q}_r\hat{q}_{r'}}{M(1-\gamma_{rr'})}\frac{1}{\sqrt{\nu_{rr}\nu_{r'r'}}}\mathbf{1}_M^\top \boldsymbol{\Sigma}_s \boldsymbol{A}_r^\top \boldsymbol{G}_r^{-1}\tilde{\boldsymbol{\Sigma}}_{rr'}\boldsymbol{G}_{r'}^{-1}\boldsymbol{A}_{r'}\boldsymbol{\Sigma}_s \mathbf{1}_M\end{aligned} \tag{C22}$$

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

$$\begin{aligned}\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=0) = \frac{s(1-c)}{1-\gamma_{rr'}}\Bigg(1 &- \frac{s(1-c)\nu_{rr}\hat{q}_r}{\nu_{rr}+(s(1-c)+\omega)\hat{q}_r} - \frac{s(1-c)\nu_{r'r'}\hat{q}_{r'}}{\nu_{r'r'}+(s(1-c)+\omega)\hat{q}_{r'}} \\ &+ \frac{s(1-c)(s(1-c)+\omega)\nu_{rr'}\hat{q}_r\hat{q}_{r'}}{(\nu_{rr}+(s(1-c)+\omega)\hat{q}_r)(\nu_{r'r'}+(s(1-c)+\omega)\hat{q}_{r'})}\Bigg) \\ &+ \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2\end{aligned} \tag{C23}$$

$$\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=1) = \frac{1}{1-\gamma_{rr'}}\left(\frac{s(1-c)(\nu_{rr'}-\nu_{rr}\nu_{r'r'})+\omega\nu_{rr'}}{\nu_{rr}\nu_{r'r'}}\right) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2 \tag{C24}$$

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

$$\gamma_{rr'} \to \frac{4\alpha\nu_{rr'}}{(\alpha+\nu_{rr}+|\alpha-\nu_{rr}|)(\alpha+\nu_{r'r'}+|\alpha-\nu_{r'r'}|)} \qquad (\lambda\to 0) \tag{C25}$$

$$\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=0) = \frac{1}{1-\gamma_{rr'}}F_0(\alpha) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2 \quad (r\neq r') \tag{C26}$$

where

$$F_0(\alpha) \equiv \left\{\begin{array}{ll} \frac{(c-1)s\left(\nu_r\nu_{r'}((2\alpha-1)(c-1)s+\omega)-\alpha^2(c-1)s\nu_{rr'}\right)}{\nu_r((c-1)s-\omega)\nu_{r'}}, & \text{if } \alpha\le\nu_{rr}\le\nu_{r'r'} \\ \frac{(c-1)s(\nu_{r'}((c-1)s\nu_r+(\alpha-1)(c-1)s+\omega)-\alpha(c-1)s\nu_{rr'})}{((c-1)s-\omega)\nu_{r'}}, & \text{if } \nu_{rr}\le\alpha\le\nu_{r'r'} \\ \frac{(c-1)s((c-1)s\nu_{r'}-cs\nu_{rr'}+(c-1)s\nu_r-cs+s\nu_{rr'}+s+\omega)}{(c-1)s-\omega}, & \text{if } \nu_{rr}\le\nu_{r'r'}\le\alpha \end{array}\right\} \tag{C27}$$

$$\langle E_{rr'}(\mathcal{D})\rangle_{\mathcal{D},\boldsymbol{w}^*}(\rho=1) = \frac{1}{1-\gamma_{rr'}}\left(\frac{s(1-c)(\nu_{rr'}-\nu_{rr}\nu_{r'r'})+\omega\nu_{rr'}}{\nu_{rr}\nu_{r'r'}}\right) + \frac{\gamma_{rr'}}{1-\gamma_{rr'}}\zeta^2 \qquad (\lambda\to 0) \tag{C28}$$

3. Phase Transitions in Uniform Resource-Constrained Ensembling.

We make further simplifications in the special case where $\omega=\zeta=0$, $\eta_r=\eta$, $\nu_{rr}=\frac{1}{k}$ for all $r=1,\dots k$, and $\nu_{rr'}=0$ for all $r\neq r'$. in the ridgeless limit $\lambda\to 0$.

$$E_k = \frac{1}{k}E_{rr}(\nu_{rr}=1/k) + \frac{k-1}{k}E_{rr'}(\nu_{rr}=\nu_{r'r'}=1/k, \nu_{rr'}=0) \tag{C29}$$

Or, in full:

$$E_k = \left\{\begin{array}{ll} -\frac{\alpha(c-1)k^2s\left(2\alpha\left(\rho^2-1\right)-2\rho^2+1\right)-(c-1)ks\left(\alpha^2\left(\rho^2-1\right)-\alpha-\rho^2+1\right)+\eta^2}{k(\alpha k-1)}, & \text{if } \alpha\le\frac{1}{k} \\ \frac{(c-1)(k-1)s\left(\alpha k^2\left(\rho^2-1\right)+k\left(-\alpha\rho^2+\alpha-2\rho^2+1\right)+2\left(\rho^2-1\right)\right)+\alpha\eta^2k^2}{k^2(\alpha k-1)}, & \text{if } \alpha\ge\frac{1}{k} \end{array}\right\} \quad (r\neq r') \tag{C30}$$

To find the boundary between the signal-dominated and noise-dominated regions, we set $E_\infty = E_1$, and rearrange to get:

$$H = \left\{ \begin{array}{ll} \alpha(1-\alpha)(1-\rho^2), & \text{if } 0 < \alpha < 1 \\ \frac{(\alpha-1)(1-\rho^2)}{\alpha}, & \text{if } \alpha > 1 \end{array} \right\} \tag{C31}$$

To find the boundary between the intermediate and noisy phases, we set: $E_{k+1} = E_k$ and take the limit $k \to \infty$, then rearrange to get:

$$H = 2 - \left(\frac{1+2\alpha}{\alpha}\right)\rho^2 \tag{C32}$$

4. Infinite Data Limit

In this section we consider the behavior of generalization error in the equicorrelated data model as $\alpha \to \infty$ 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 figure 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:

$$\hat{q}_r \to \frac{\alpha}{\lambda} \tag{C33}$$

$$q_r \to \frac{\nu_{rr}\lambda}{\alpha} \tag{C34}$$

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

$$E_g = \frac{1}{k^2}\sum_{rr'=1}^{k} E_{rr'} = s\left[1 - \left(2 - \frac{1}{k}\right)\left(\frac{1}{k}\sum_{r=1}^{k}\nu_{rr}\right)\right] + \frac{\eta^2}{k} \tag{C35}$$

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.

Appendix D: Numerical Experiments

Numerical experiments are performed by generating synthetic datasets by drawing data randomly from multivariate Gaussian distributions, assigning feature noise and noisy labels. Writing the training set in terms of a data matrix $\boldsymbol{\Psi} \in \mathbb{R}^{M \times P}$ in which column μ consist of the training point $\boldsymbol{\psi}_\mu$ and the labels are organized into a column vector $\boldsymbol{y}$ such that $\boldsymbol{y}_\mu = y_\mu$, the learned weights are calculated as:

$$\hat{\boldsymbol{w}} = \boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\top\boldsymbol{\Psi} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{y} \tag{D1}$$

In the ridgeless case, a pseudoinverse is used:

$$\hat{\boldsymbol{w}} = \boldsymbol{\Psi}^\dagger \boldsymbol{y} \tag{D2}$$

Numerical experiments were performed using the PyTorch library [46]. The code used to perform numerical experiments and generate plots is provided in a zip file with this submission, and will be made publicly available on GitHub upon acceptance of this manuscript. Numerical computations necessary for this work may be performed in a small amount of time (less than one hour) using an Nvidia GPU.

1. Details of Heterogeneous Subsampling Theory

In this section, we describe the method used to calculate loss curves for heterogeneous subsampling experiments seen in main text fig. 3. In each trial, Subsampling fractions $\{\nu_1, \ldots, \nu_k\}$ are generated according to the following process:

1. Each fraction ν_{rr} is drawn independently from a Γ distribution with mean $\frac{1}{k}$ and variance σ^2. $\nu_{rr} \sim \Gamma_{k,\sigma}$
2. The fractions are re-scaled in order to sum to unity: $\nu_{rr} \to \nu_{rr}/(\nu_1 + \cdots + \nu_k)$

Equivaltly, the fractions ν_{rr} are drawn from a Dirichlet distribution [40]. Then, the loss curves are calculated from the given fractions $\{\nu_{rr}\}$ using equations C16, C17, C23, C24. The dotted red lines show the loss curves for 5 single trials. The solid red lines show the average loss curves from 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:

$$\Gamma_{k,\sigma} \equiv \Gamma\left(\text{shape } = (k\sigma)^{-2}, \text{scale } = k\sigma^2\right) \tag{D3}$$

Appendix E: Code Availability

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.