Statistical mechanics of support vector regression

Abdulkadir Canatar and SueYeon Chung
Center for Computational Neuroscience, Flatiron Institute, New York, New York 10010, USA
and Center for Neural Science, New York University, New York, New York 10003, USA

(Received 9 January 2025; accepted 23 June 2025; published 4 August 2025)

A key problem in deep learning and computational neuroscience is relating the geometrical properties of neural representations to task performance. Here, we consider this problem for continuous decoding tasks where neural variability may affect task precision. Using methods from statistical mechanics, we study the average-case learning curves for ε-insensitive support vector regression and discuss its capacity as a measure of linear decodability. Our analysis reveals a phase transition in training error at a critical load, capturing the interplay between the tolerance parameter ε and neural variability. We uncover a double-descent phenomenon in the generalization error, showing that ε acts as a regularizer, both suppressing and shifting these peaks. Theoretical predictions are validated both with toy models and deep neural networks, extending the theory of support vector machines to continuous tasks with inherent neural variability.

DOI: 10.1103/78dr-c4xd

I. INTRODUCTION

Understanding how neural systems and machine learning models perform robust decoding in the presence of task-irrelevant variability is a fundamental challenge [1–3]. In real-world tasks, input representations often contain fluctuations that do not affect the target variable of interest. For example, estimating the position of a specific animal from video frames requires ignoring variations due to their sizes, lighting changes, and backgrounds [4]. Such nuisance variability complicates learning, especially when data are limited.

Robust decoding thus demands representations that preserve task-relevant features while remaining tolerant to task-irrelevant variability. In both biological and artificial systems, this tolerance can be viewed as a form of perceptual invariance—the ability to treat small deviations as equivalent for the purposes of a task. Quantifying how much variability a representation can accommodate without loss of decoding precision remains an open problem in both neuroscience and machine learning.

In this work, we introduce a theoretical framework to quantitatively characterize ε-insensitive support vector regression (ε-SVR) [5,6] to analyze the robustness of continuous decoding. In ε-SVR, deviations smaller than a tolerance ε are not penalized during learning, allowing us to directly model how representational variability affects decoding accuracy. The ε-SVR task is similar to ridgeless regression, except that the objective is satisfied when the prediction error lies within a tube of size ε around the target (Fig. 1). Here, the *tube size* ε acts as a hyperparameter that defines the margin of tolerance for input deviations and must be fine-tuned depending on the structure of the variability.

Let P denote the number of training samples, N the dimensionality of the representation, and $\alpha := P/N$ the *load*. In ridgeless regression, when input features and/or targets are noisy, the training error E_{tr} goes through a phase transition, from being exactly zero to being strictly nonzero, at the critical load $\alpha_c = 1$, also known as the interpolation threshold. At this point, double-descent occurs [7]; the generalization error E_g diverges due to overfitting and starts improving again for $\alpha > \alpha_c$. As $\alpha \to \infty$, training and generalization errors approach each other and asymptotically become the same.

By tuning an appropriate tube size, SVR allows fitting more samples ($\alpha_c > 1$) with zero training error by tolerating errors up to ε. Alternatively, one can think of shifting the interpolation threshold beyond $\alpha = 1$ in ridgeless regression.[1] This is always possible by picking an arbitrarily large ε, in which case no learning occurs. Hence, we consider the minimum ε^* for which the training error is zero. This rate is interpreted as the decoding precision of the representation. In other words, the amount of tolerance ε^* required to learn all training examples successfully quantifies how precisely a representation can decode the task. Furthermore, this also defines a form of invariant representation, namely ε-invariance, since the objective function only requires the predictions to be in the ε-vicinity of the exact target.

Using the replica method [8–10], we develop an analytical theory of generalization for ε-SVR that relates the precise geometric properties of representations to their decoding precision of continuous variables. Our theory recovers the previous results on ridge regression in the limit $\varepsilon \to 0$ [11–16]. When applied to our setting, our results also agree with the predictions of the theory by Loureiro *et al.* [17] developed for generic loss functions (see Appendix A 8).

The rest of the paper is organized as follows:

[1]Qualitatively, SVR still differs from ridgeless regression with a shifted interpolation threshold.

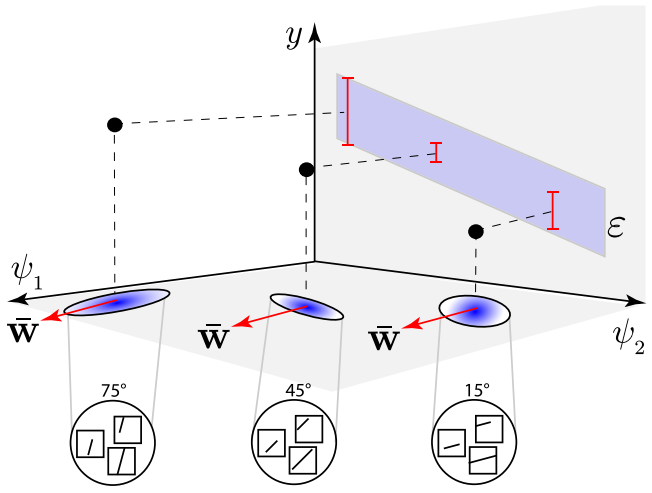

FIG. 1. Decoding the orientation of sticks (degrees) from their pixel representations as an example of invariant regression. For each orientation, labeled circles depict stick images with different variations, such as location and size. Blue regions depict the corresponding neural representations in (ψ_1, ψ_2)-space induced by task-irrelevant variances and depict possible noise geometries with $\boldsymbol{\delta}$ parallel (left), orthogonal (middle), and partially aligned (right) with the coding direction $\bar{\mathbf{w}}$.

In Sec. II, we review the relevant literature and discuss the similarities with our work.

In Sec. III, we assume Gaussian representations with arbitrary covariance structure and derive analytical expressions for the training and generalization errors of ε-SVR when presented with P training samples. We show that, for fixed P, the training error undergoes a phase transition at a critical value ε^* beyond which it becomes exactly zero. We introduce and numerically test a toy data model and discuss the implications of the theory.

In Sec. IV, we derive optimal learning curves for both ridge regression and support vector regression. We show that optimal ridge regression always outperforms optimal ε-SVR.

In Sec. V, we apply our theory to deep convolutional neural network representations on a continuous decoding task of grating images. We find that training neural networks on natural images allows better discriminability of angles.

II. BACKGROUND

Understanding geometric properties of neural representations of data with invariances is a longstanding problem in both machine learning [18,19] and neuroscience [20]. Here, we review two main lines of prior work related to invariant decoding.

One line of work studies how known symmetries of a particular task can be exploited in designing tailored machine learning models [21–23]. These works demonstrate that regression with kernels that are explicitly invariant under the group actions of the symmetry group of some task significantly improves the sample efficiency. These representations express only task-relevant input variances and hence effectively overcome the curse of dimensionality. Instead of group invariances, here we consider ε-invariance in generic representations, quantifying the amount of decoding precision that must be sacrificed to achieve zero training error. One can also imagine simplifying the decoding problem and discretizing the output variable in a way that input variances do not affect the performance anymore.

A second line of research examines how geometric properties of neural representations relate to invariant prediction, with a particular emphasis on linear decodability under neural variability [20,24,25]. These studies analyzed classification performance using support vector classification (SVC), linking the number of separable object manifolds to the geometry of the representation. In related work, Farrell *et al.* [26] explored how continuous group invariances influence classification capacity. In this paper, we extend this line of inquiry to continuous regression tasks.

III. THEORETICAL FRAMEWORK

We consider a supervised learning setting in which inputs are represented by N-dimensional *center* representations $\bar{\boldsymbol{\psi}}$, and labels $\bar{y}$ are linearly generated via a coding direction $\bar{\mathbf{w}}$, i.e., $\bar{y} = \bar{\mathbf{w}} \cdot \bar{\boldsymbol{\psi}}$. Here, the center representations $\bar{\boldsymbol{\psi}}$ are treated as random variables and capture the task-relevant information from the input data.

In contrast, the learner observes *encoding* representations $\boldsymbol{\psi}$ which include both the task-relevant centers $\bar{\boldsymbol{\psi}}$ and task-irrelevant variations modeled by random variables $\boldsymbol{\delta}$, and assumed to be of the form $\boldsymbol{\psi} \equiv \bar{\boldsymbol{\psi}} + \boldsymbol{\delta}$. The *noise* representations $\boldsymbol{\delta}$ represent nuisance variations around the centers. The goal is to learn a linear predictor $y = \mathbf{w} \cdot \boldsymbol{\psi}$. Similar data models have also been analyzed in prior studies of ridge regression [15–17,27,28].

For each training set instance, P samples $\{\bar{\boldsymbol{\psi}}^\mu, \boldsymbol{\psi}^\mu\}_{\mu=1}^P$ are drawn i.i.d. from their respective distributions, and used to generate labels $\bar{y}^\mu$ and predictions y^μ. Due to the additional noise, the predictions always mismatch with the labels when $P > N$ and overfit to noise when $P < N$. To account for this variance, we consider the ε-SVR algorithm

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2\lambda}\sum_{\mu=1}^{P} [|y^\mu - \bar{y}^\mu| - \varepsilon]_+^2, \tag{1}$$

where ε is the tube size, λ is the ridge parameter, and $[x]_+ \equiv \max(0, x)$. We note that both λ and ε serve as regularizers for regression since both control the effect of training error on the solution. However, we later show that they have qualitatively different effects on generalization (see Appendix A). Although our analysis holds for general λ and ε, here we focus on the ridgeless limit ($\lambda = 0$), where the objective is fitting all training samples exactly within the ε-tube.

Our goal is to compute the average-case training and generalization errors for ε-SVR. By treating the training set as a quenched disorder, we compute these average quantities using the replica method from statistical physics and consider the following partition function:

$$\overline{\log Z} = \overline{\log \int d\mathbf{w}\, e^{-NS(\mathbf{w},\boldsymbol{\psi}^\mu,\bar{\boldsymbol{\psi}}^\mu)+J\mathcal{O}(\mathbf{w})}},$$

$$S = \frac{1}{2N}\|\mathbf{w}\|^2 + \frac{1}{2\lambda N}\sum_{\mu=1}^{P}[|\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu| - \varepsilon]_+^2, \tag{2}$$

where the overline denotes the quenched average over the training set $\{\boldsymbol{\psi}^\mu, \bar{\boldsymbol{\psi}}^\mu\}$ of size P. In the thermodynamic limit $P, N \to \infty$ while keeping $\alpha \equiv P/N \sim \mathcal{O}(1)$, the free energy S in Eq. (2) becomes self-averaging, concentrating around the solution to the problem in Eq. (1). Then, the average of any observable $\langle \mathcal{O}(\mathbf{w}) \rangle$ can be computed by taking derivatives of $\overline{\log Z}$ with respect to source J.

The training error $E_{\rm tr}$ and generalization error E_g, the two observables we are interested in, are given by

$$E_{\rm tr} = \overline{\frac{1}{P}\sum_{\mu=1}^{P}[|\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu| - \varepsilon]_+^2},$$
$$E_g = \overline{\langle(\mathbf{w}\cdot\boldsymbol{\psi} - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}})^2\rangle_{\boldsymbol{\psi},\bar{\boldsymbol{\psi}}}}. \tag{3}$$

A. Replica calculation

For simplicity, we assume a joint Gaussian distribution over center and encoding representations:

$$\begin{bmatrix}\boldsymbol{\psi}\\ \bar{\boldsymbol{\psi}}\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\psi}} & \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}^\top & \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}}\end{bmatrix}\right). \tag{4}$$

Note that the solution to the objective in Eq. (2) under this data distribution depends critically on the correlation between centers $\bar{\boldsymbol{\psi}}$ and encoding representations $\boldsymbol{\psi}$. Specifically, in the large-sample limit $P \to \infty$, the optimal weights and generalization error converge to

$$E_\infty = \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} - \mathrm{tr}\mathbf{w}_*\mathbf{w}_*^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \mathbf{w}_* = \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}},$$

where tr denotes the normalized trace operator. The vector $\mathbf{w}_*$ represents the projection of the target weights $\bar{\mathbf{w}}$ into the subspace accessible to the learner, and E_∞ denotes the irreducible error corresponding to the unexplainable variance in the target [15].

By performing the quenched average using this distribution—under a replica symmetric ansatz and saddle-point approximation (see Appendix A)—we obtain a mean-field theory characterized by three effective parameters: the *effective regularization* $\tilde{\lambda}$, the *effective tube size* $\tilde{\varepsilon}$, and the *effective load* $\tilde{\alpha}$:

$$\tilde{\lambda} = \lambda + \mathcal{F}(\tilde{\lambda},\tilde{\varepsilon}), \quad \tilde{\varepsilon} = \varepsilon/\sqrt{\mathcal{G}(\tilde{\lambda},\tilde{\varepsilon})}, \quad \tilde{\alpha} = \alpha f(\tilde{\varepsilon}), \tag{5}$$

where $\tilde{\lambda}$ and $\tilde{\varepsilon}$ satisfy the following coupled self-consistent equations:

$$\mathcal{F}(\tilde{\lambda},\tilde{\varepsilon}) = \mathrm{tr}\mathbf{M}, \quad \mathbf{M} = \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \quad \mathbf{G} = \mathbf{I} + \frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Sigma}_{\boldsymbol{\psi}},$$
$$\mathcal{G}(\tilde{\lambda},\tilde{\varepsilon}) = \frac{E_\infty + \mathrm{tr}\mathbf{M}\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*^\top}{1 - g(\tilde{\varepsilon})\gamma}, \quad \gamma = \partial_{\tilde{\lambda}}\mathcal{F}(\tilde{\lambda},\tilde{\varepsilon}). \tag{6}$$

Here, the scalar functions $f(\tilde{\varepsilon})$ and $g(\tilde{\varepsilon})$ are given by

$$f(\tilde{\varepsilon}) = \langle\theta(|z| - \tilde{\varepsilon})\rangle_z, \quad g(\tilde{\varepsilon}) = \frac{\langle(|z| - \tilde{\varepsilon})^2\theta(|z| - \tilde{\varepsilon})\rangle_z}{\langle\theta(|z| - \tilde{\varepsilon})\rangle_z},$$

where $\langle f(z)\rangle_z$ denotes expectation with respect to a standard normal variable z. Furthermore, the function $\mathcal{G}(\tilde{\lambda},\tilde{\varepsilon})$ coincides with the generalization error E_g from Eq. (3), and the training error is given by

$$E_{\rm tr} = \left(1 - \frac{\mathcal{F}(\tilde{\lambda},\tilde{\varepsilon})}{\tilde{\lambda}}\right)^2 f(\tilde{\varepsilon})g(\tilde{\varepsilon})E_g, \quad E_g = \mathcal{G}(\tilde{\lambda},\tilde{\varepsilon}). \tag{7}$$

The effective regularization $\tilde{\lambda}$ is a renormalized ridge parameter [16,29] that governs the smoothness of the solution and quantifies the implicit regularization of the model [12,13].

The effective load $\tilde{\alpha}$ gives a measure of the actual number of samples that contribute to learning, and decreases monotonically as a function of the effective tube size $\tilde{\varepsilon}$. Intuitively, larger tube sizes require sampling more points since it becomes less likely to hit the tube boundary. A similar notion was also introduced in Ref. [30] when the learning is restricted to a subset of training examples.

The effective tube size $\tilde{\varepsilon}$, however, is a renormalized measure of tolerance and depends inversely on generalization error, implying that the number of samples required to improve generalization increases as the model's predictivity improves [Eq. (5)].

Finally, $\tilde{\varepsilon} = \varepsilon/\sqrt{E_g}$ can also be interpreted as a form of effective discriminability varying as a function of load α. In neuroscience, *discriminability* $d' \equiv \varepsilon/\sqrt{E_\infty}$ is a common metric used to quantify the perceptible contrast between two stimuli ε-apart from each other [31,32].

B. Analysis of a toy data model

To analyze our results and build intuition, we apply our theory to investigate the decoding precision in the presence of structured noise, which correlates with the coding direction. We consider a toy data model where the centers are drawn from $\bar{\boldsymbol{\psi}} \sim \mathcal{N}(0, \mathbf{I})$ and the labels are generated by $\bar{y} = \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}$ with $\mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top = 1$. The representations seen by the linear model are of the form $\boldsymbol{\psi} = \bar{\boldsymbol{\psi}} + \sigma\boldsymbol{\delta}$ where the noise $\boldsymbol{\delta}$ is drawn from a structured Gaussian distribution with zero mean and covariance

$$\boldsymbol{\Sigma}_{\boldsymbol{\delta}} = (1-\beta)\mathbf{P} + \beta(\mathbf{I} - \mathbf{P}), \quad \mathbf{P} \equiv \mathbf{I} - \frac{1}{N}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top, \tag{8}$$

where σ controls the noise strength. Through the parameter β, this noise model interpolates between the cases where the noise is completely orthogonal ($\beta = 0$) or parallel ($\beta = 1$) to the coding direction $\bar{\mathbf{w}}$. A schematic of this data model is illustrated in Fig. 1.

For this model, the training and generalization errors from Eq. (3) simplify to the following forms (see Appendix C):

$$E_g = \frac{1}{1 - \frac{g(\tilde{\varepsilon})\tilde{\alpha}}{(\tilde{\alpha}+\tilde{\lambda})^2}}\left(E_\infty + \frac{1 - E_\infty}{\left(1 + \frac{\tilde{\alpha}}{\tilde{\lambda}}\frac{1+\beta\sigma^2}{1+(1-\beta)\sigma^2}\right)^2}\right),$$
$$E_{\rm tr} = \left(1 - \frac{1}{\tilde{\alpha}+\tilde{\lambda}}\right)^2 f(\tilde{\varepsilon})g(\tilde{\varepsilon})E_g, \quad E_\infty = \frac{\beta\sigma^2}{1+\beta\sigma^2}, \tag{9}$$

where the self-consistent equations for $\tilde{\lambda}$ and $\tilde{\varepsilon}$ in Eq. (5) can be solved numerically. Furthermore, in the ridgeless limit $\lambda \to 0$, the expression for $\tilde{\lambda}$ simplifies to

$$\tilde{\lambda} = \begin{cases}1 - \tilde{\alpha}, & \tilde{\alpha} < 1,\\ 0, & \tilde{\alpha} \geqslant 1.\end{cases} \tag{10}$$

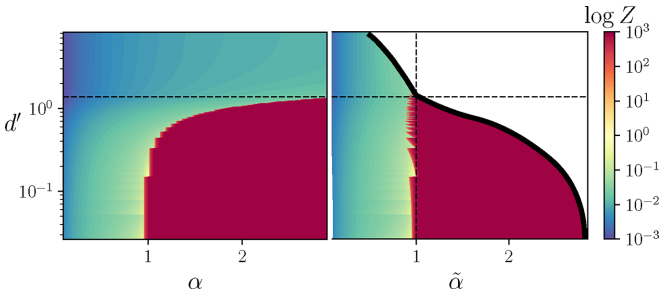

FIG. 2. Phase plots of the theoretical values of $\overline{\log Z}$ against $d' = \varepsilon/\sqrt{E_\infty}$. (Left) For fixed E_∞, increasing ε delays phase transition, and for fixed ε, increasing noise variability (σ) causes a phase transition for smaller sample sizes. (Right) Same plot but as a function of $\tilde{\alpha}$. The vertical dashed line indicates the phase boundary at $\tilde{\alpha} = 1$, the thick solid curve is the boundary $\alpha = \alpha_{\max} = 3$, and the horizontal line indicates their intersection corresponding to the minimum d' to achieve zero training error.

Equations (9) and (10) imply that in the over-parameterized regime, when the effective sample size is smaller than the number of parameters ($\tilde{\alpha} < 1$), the training error vanishes identically and goes through a first-order phase transition at $\tilde{\alpha} = 1$ corresponding to the interpolation threshold. Beyond this threshold ($\tilde{\alpha} > 1$), E_{tr} is nonzero and approaches the generalization error for large $\tilde{\alpha}$. Furthermore, when $\tilde{\alpha} > 1$, the generalization error E_g becomes proportional to E_∞, and thus identically vanishes when either β or σ is zero. In Fig. 2, we visualize this phase transition by plotting the free energy across model parameters.

In the limit $\varepsilon \to 0$, our theory recovers the equations from previous work on ridge regression [12,13,28]. However, we discover additional phenomena in SVR that are lacking in ridge regression and are reported below.

C. Model capacity and effective sample size

Our analysis reveals that the phase boundary is governed by the effective load $\tilde{\alpha} = 1$ (see Fig. 2). This critical point is clearly observed for both training error E_{tr} and generalization error E_g in Fig. 3, where the theoretical predictions closely match empirical results.

In ridge regression ($\varepsilon = 0$), the training error undergoes a phase transition at $\alpha = 1$. For SVR with $\varepsilon > 0$, the transition shifts to a higher value $\alpha > 1$ [Fig. 3(a)]. This is expected since larger tube sizes allow for more *capacity* in the sense that more samples can be fit satisfying $E_{\text{tr}} = 0$. Hence, the transition for nonzero ε occurs at some critical load $\alpha_c > 1$,

$$\alpha_c^{-1} = f(\tilde{\varepsilon}), \tag{11}$$

so that $\tilde{\alpha} = 1$ [see Fig. 3(a), inset]. This critical value α_c represents the maximum number of training examples that can be fit within an ε-tube and, in analogy to classification capacity [24], can be used as a measure that ties representational geometry to the precision of continuous decoding tasks.

Computing the capacity α_c requires solving the coupled equations in Eq. (5) which are, in general, analytically intractable. However, it can be exactly solved for our toy model Eq. (9) at the critical point $\tilde{\alpha} = 1$ (see Appendix C). In this case, the solution asymptotically has the form

$$\tilde{\varepsilon} \approx \begin{cases} d' - 1/d', & d' \gg 1, \\ d'^2\sqrt{\frac{2}{\pi}}, & d' \ll 1. \end{cases} \tag{12}$$

Through Eq. (11), this result explicitly links the capacity α_c to discriminability d', and shows that lower discriminability (larger tube sizes) permits fitting more training samples without error.

Furthermore, it allows us to connect the model's capacity to its representational properties since d' explicitly depends on model parameters σ (noise) and β (task correlation). For fixed tube size ε, discriminability d' increases with more noise and task correlation. However, when the noise geometry is completely orthogonal to the task ($\beta = 0$), the noise magnitude does not affect d' and hence the capacity.

D. Generalization and double-descent

The analysis of E_g reveals a well-known phenomenon known as *double-descent* [7,33] where the generalization degrades as the number of samples approaches the critical point $\tilde{\alpha} = 1$ [Fig. 3(b)]. Essentially, double-descent implies overfitting to noisy samples at the model capacity and signals a sharp transition between the over- and under-parameterized regimes.

In regression, this overfitting can be avoided by tuning the ridge parameter λ, which penalizes solutions with large norms [34]. In our case, although we set the ridge parameter

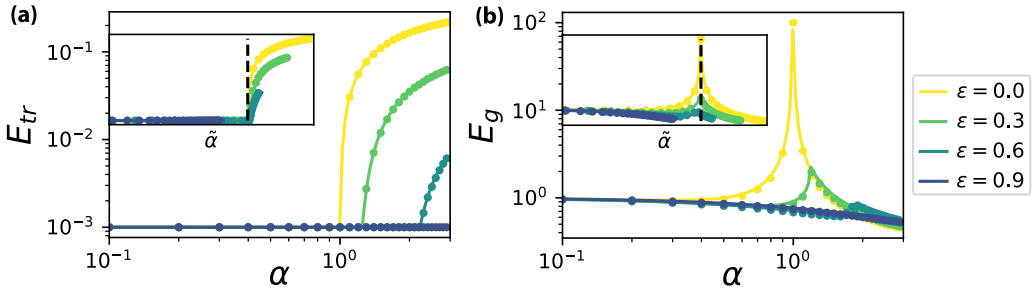

FIG. 3. Theoretical (lines) and experimental (dots) learning curves for the toy model with $\beta = 0.5$ and $\sigma = 1$. Insets show the same curves but as a function of $\tilde{\alpha}$. (a) With increasing ε-tube size, training error decreases due to increasingly relaxed definition of being correct, therefore allowing more samples to be fit. (b) However, generalization error displays a nonmonotonicity called the double-descent. In this case, the effect of increasing ε amounts to keeping the problem in the *overparameterized* regime.

to zero, turning on ε also acts like a regularization. However, in contrast to the ridge penalty, which only suppresses the double-descent peak, ε additionally shifts its location, implying that the tube size effectively controls whether the model operates in an over- or under-parameterized regime [Fig. 3(b)].

Similar behavior has been previously observed in many ridge-regression settings where increasing the number of effective parameters of a model helps shift the double-descent peak [15,27,35]. Based on these observations, we may also consider ε as a parameter that tunes the effective model size.

IV. OPTIMAL RIDGE REGRESSION VERSUS OPTIMAL SUPPORT VECTOR REGRESSION

Determined by the critical point in training error, the capacity condition in Eq. (11) determines either the maximum allowable load α for a fixed tolerance ε, or the minimum possible tolerance ε required to fit a dataset of size α. However, at the critical point ($\tilde{\alpha} = 1$), the generalization error typically diverges [Fig. 2(b)], implying that the model completely overfits at capacity with poor generalization. For this reason, we define the *optimal* load $\alpha_{\rm opt} < \alpha_c$ as the maximum number of samples that can be fit with zero training error while maintaining acceptable generalization.

Here, we analyze this quantity by minimizing the generalization error in Eq. (3) with respect to model hyperparameters. Treating ε as a free parameter, we compute the optimal tolerance $\varepsilon_{\rm opt}$ as a function of sample size and obtain analytical results for optimal ε-SVR. For comparison, we also derive learning curves for optimal ridge regression by minimizing E_g with respect to λ [36] and compare their generalization performances. Since we get analytical expressions for both $\varepsilon_{\rm opt}$ and $\lambda_{\rm opt}$, our results furthermore help bypass computationally intensive hyperparameter sweeps for ridge regression and ε-SVR [37].

For optimal SVR, we set $\lambda = 0$ and solve for $\partial_\varepsilon E_g = 0$. While the analytical calculation for optimal ε is tedious (see Appendix B), we can obtain a new set of self-consistent equations replacing the ones in Eq. (5). For optimal SVR, we get the following self-consistent equations:

$$\tilde{\lambda} = \mathrm{tr}\mathbf{M}, \quad h(\tilde{\varepsilon}_{\rm opt}) = \frac{\mathcal{A}(\tilde{\varepsilon}_{\rm opt}, \tilde{\lambda})}{1-\gamma}, \tag{13}$$

where

$$\mathcal{A}(\tilde{\varepsilon}, \tilde{\lambda}) = \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\mathrm{tr}\mathbf{M}^2}\left(g(\tilde{\varepsilon}) - \frac{\tilde{\lambda}}{C}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}\right),$$
$$h(\tilde{\varepsilon}) \equiv \tilde{\varepsilon}\left(\tilde{\varepsilon} + \frac{f'}{f} - \frac{f}{f'}\right). \tag{14}$$

While being complicated, these self-consistent equations can be easily solved numerically.

For optimal ridge regression, we set $\varepsilon = 0$ and solve for $\partial_\lambda E_g = 0$. In this case, the self-consistent equation simply reduces to $\mathcal{A}(\tilde{\varepsilon}, \tilde{\lambda}) = 0$ (see Appendix B) and yields

$$\tilde{\lambda} = C\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}, \tag{15}$$

replacing the original equation $\tilde{\lambda} = \lambda + \mathrm{tr}\mathbf{M}$ in Eq. (5).

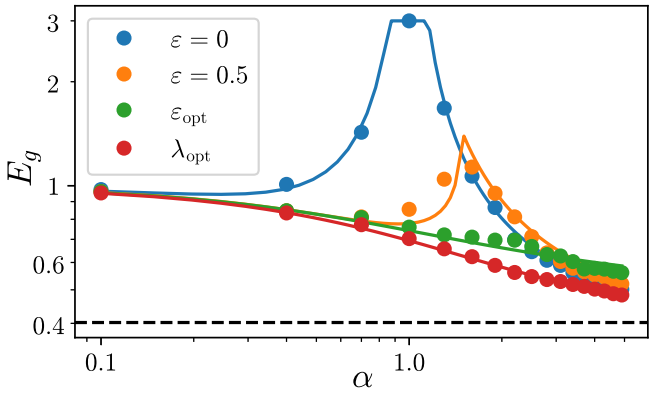

FIG. 4. Optimal learning curves for the toy model. Choosing optimal λ for ridge regression always outperforms optimal ε-SVR. Lines indicate theory, and dots indicate experiments. The dashed line indicates E_∞. Optimal λ and ε change as a function of α, and their values are reported in Fig. 6.

In Fig. 4, we test our theory and observe perfect agreement with experiments. Interestingly, optimal ridge regression consistently outperforms optimal ε-SVR in terms of generalization. Note that the optimal ridge regression always maintains a nonzero training error, while optimal SVR ensures zero training error for all α. Here, optimal SVR yields worse generalization performance because it stops learning as soon as the constraints are satisfied, while in optimal ridge regression, the model is forced to learn finer details since the training error never vanishes.

V. REAL DATA APPLICATIONS

Although our theory is derived in the self-averaging limit $N \to \infty$ and $\alpha \sim \mathcal{O}(1)$, we find that it agrees well with experiments when it is applied to real data with finite P and N.

To test the applicability of our theory, we designed an experiment where the task is to decode the orientation of grating images from their deep neural network representations. We generated grating images with orientations θ uniformly sampled from $[0, \pi)$ and augmented them with task-irrelevant attributes such as spatial frequency and phase (see Appendix E). We extracted their neural network representations from the hidden layers of various ResNet architectures and evaluated their decoding performance using optimal ε-SVR.

In Fig. 5, we show the optimal $\varepsilon_{\rm opt}$ and generalization error for both randomly initialized models and trained models on natural scenes. Here, $\varepsilon_{\rm opt}$ is normalized by the range of the decoding interval L such that the case $\varepsilon_{\rm opt}/L > 1$ implies that decoding any orientation is impossible while fitting all training examples. For random networks, discriminability improves with depth. In contrast, for trained networks, intermediate layers yield better discriminability, which then degrades in deeper layers. This trend—where early and middle layers better support decoding of simple stimuli—has also been observed in prior work [4].

Our theoretical framework allows the decomposition of generalization error into interpretable spectral components, offering insights into how and why decoding performance

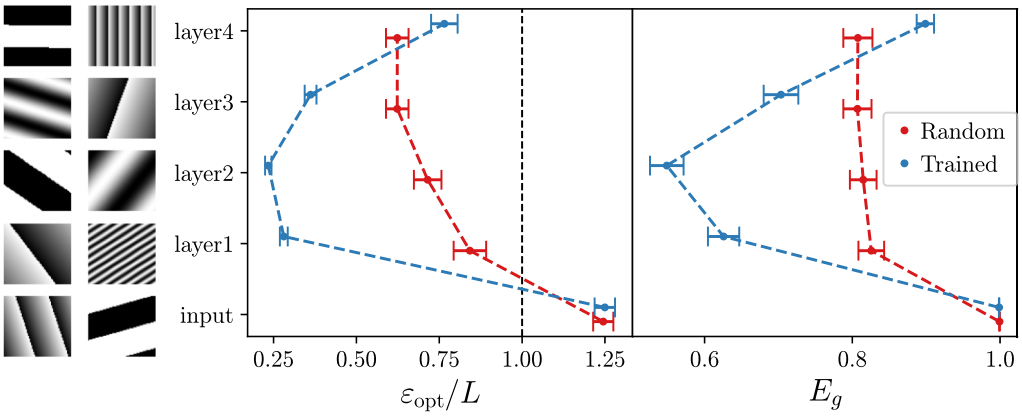

FIG. 5. SVR analysis of layers extracted from trained and random ResNet models on the task of decoding grating orientations.

evolves across network layers. In Appendix D, we demonstrate that the error mode geometry framework from Ref. [38] can be directly applied to our setting (see Appendix E for error mode geometry analysis).

VI. DISCUSSION

Invariant decoding remains a fundamental challenge in both machine learning and neuroscience, as real-world inputs frequently vary in ways that do not alter the target attribute. This study bridges the gap between representational geometry and its influence on continuous decoding tasks. By utilizing ε-SVR as a geometric probe in high-dimensional feature spaces, we employed average-case analysis to explore how the geometry of noise representations impacts decoding performance.

In particular, our analysis revealed that the phase transition in training error can be controlled by the tube size parameter ε, which also determines the desired precision of the task. Larger ε values increase the model's capacity by allowing more samples to be fit without error, but this comes at the cost of reduced precision in continuous variable decoding, as indicated by increased generalization error. This trade-off between training and generalization errors underscores the importance of tuning ε to optimize both model fit and predictive accuracy.

We also derived analytical expressions for optimal learning curves for both ε-SVR and ridge regression. While prior work also derived analytical formulas for the optimal ridge parameter [36,39–43], these results were limited to simplified data designs. Here, our framework allowed us to extend these results to generic Gaussian data with arbitrary covariance and compute optimal hyperparameters analytically for both ridge strength λ and tube width ε without a hyperparameter search.

While our analysis currently relies on Gaussian assumptions and linear decoding strategies, future investigations should extend beyond these assumptions and address more realistic coding scenarios in distributed neural representations.

Our theoretical framework reveals fundamental tradeoffs between precision and robustness in neural computations. The relationship we uncovered between representational geometry and decoding performance not only enhances our understanding of SVR algorithms but could also inform the design of more robust regression models and understand the nature of neural decoding algorithms where precise representations must be maintained despite inherent variability.

ACKNOWLEDGMENTS

The authors thank Chi-Ning Chou for helpful comments on the manuscript. This work was supported by the Center for Computational Neuroscience at the Flatiron Institute of the Simons Foundation, as well as by a Alfred Sloan Foundation Fellowship and an Award by Esther A. and Joseph Klingenstein Fund and Simons Foundation (to S.C.). All experiments were performed on the high-performance computing cluster at the Flatiron Institute of Simons Foundation.

DATA AVAILABILITY

The data that support the findings of this article are openly available [44].

APPENDIX A: PROBLEM SETUP

Here, we study a family of regression problems and their generalization properties using the replica method. In our notation, we consider a training set $\mathcal{D} = \{(\mathbf{x}^\mu, \bar{y}^\mu)\}_{\mu=1}^P$ consisting of P i.i.d. drawn inputs $\mathbf{x}^\mu$ and labels $\bar{y}^\mu$. In this work, we assume that the target labels are related to the inputs via

$$\bar{y}^\mu = \bar{\mathbf{w}} \cdot \bar{\boldsymbol{\psi}}^\mu, \quad \bar{\boldsymbol{\psi}}^\mu \equiv \bar{\boldsymbol{\psi}}(\mathbf{x}^\mu) \in \mathbb{R}^N, \quad \bar{\mathbf{w}} \in \mathbb{R}^N, \tag{A1}$$

where $\bar{\boldsymbol{\psi}}^\mu$ denote N-dimensional *center* features as a function of input samples, and $\bar{\mathbf{w}}$ denotes the *coding direction*. However, we consider a linear predictor

$$y^\mu = \mathbf{w} \cdot \boldsymbol{\psi}^\mu, \quad \boldsymbol{\psi}^\mu \equiv \boldsymbol{\psi}(\mathbf{x}^\mu) \in \mathbb{R}^N, \quad \mathbf{w} \in \mathbb{R}^N, \tag{A2}$$

where $\boldsymbol{\psi}^\mu$ denote the N-dimensional *encoding* features available to the predictor, and $\mathbf{w}$ is to be learned.

Our generic regression task solves the following problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\lambda}\sum_{\mu=1}^{P} \ell(\mathbf{w} \cdot \boldsymbol{\psi}^\mu - \bar{\mathbf{w}} \cdot \bar{\boldsymbol{\psi}}^\mu), \tag{A3}$$

where $\ell(x)$ is the loss function. Below we list possible choices for $\ell(x)$:

(1) *Kernel Ridge Regression (KRR)*: For ridge regression, $\ell(x) = \frac{1}{2}x^2$ and the optimization problem becomes

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2\lambda}\sum_{\mu=1}^{P}(\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)^2. \tag{A4}$$

Here, $\lambda \to 0$ limit corresponds to the ridgeless (least-squares) regression, and the case with linear features [$\boldsymbol{\psi}(\mathbf{x}) = \mathbf{x}$] corresponds to linear regression.

(2) *Support Vector Regression (SVR)*: Specifically, we consider ε-insensitive SVR where the loss vanishes if the error is less than some positive tolerance variable ε. Generically, the loss function is $\ell(x) = \frac{1}{n!}\max(0, |x| - \varepsilon)^n$, and the optimization problem becomes

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{n!\lambda}\sum_{\mu=1}^{P}\max(0, |\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu| - \varepsilon)^n, \tag{A5}$$

where $n \in \mathbb{N}_0$ are nonnegative integers. Note that the case where $n = 2$ and $\varepsilon = 0$ is identical to KRR above.

A few notes are in order:

(1) Our calculations require averaging over this inputs $\mathbf{x}^\mu$, which is intractable for general feature maps. However, we make a *Gaussian assumption* that the statistics of $\bar{\boldsymbol{\psi}}(\mathbf{x}^\mu)$ and $\boldsymbol{\psi}(\mathbf{x}^\mu)$ can be approximated by their second moments:

$$\begin{aligned}
\langle\bar{\boldsymbol{\psi}}(\mathbf{x})\bar{\boldsymbol{\psi}}(\mathbf{x})^\top\rangle_{\mathbf{x}} &= \frac{1}{N}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}},\\
\langle\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^\top\rangle_{\mathbf{x}} &= \frac{1}{N}\boldsymbol{\Sigma}_{\boldsymbol{\psi}},\\
\langle\boldsymbol{\psi}(\mathbf{x})\bar{\boldsymbol{\psi}}(\mathbf{x})^\top\rangle_{\mathbf{x}} &= \frac{1}{N}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}},
\end{aligned} \tag{A6}$$

and the final results will depend only on these quantities. This approximation simplifies our calculations, yields excellent agreement with experiments, and is also justified by the *Gaussian equivalence* theorems in high dimensional statistics where the covariance of nonlinear features (kernels) can be summarized by an equivalent Gaussian statistics [45–48].

(2) In our formulation, we excluded the effect of label noise in the training set [13] to avoid clutter, since the effect of label noise with variance σ^2 can equivalently be studied by shifting $\bar{\boldsymbol{\psi}}^\mu \to \bar{\boldsymbol{\psi}}^\mu + \mathbf{n}^\mu$ with an independent random vector $\mathbf{n}$ with $\mathbb{E}(\bar{\mathbf{w}}\cdot\mathbf{n})^2 = \sigma^2$.

(3) We assume that the encoding representations depend on centers via

$$\boldsymbol{\psi}(\mathbf{x}) = \mathbf{A}\bar{\boldsymbol{\psi}}(\mathbf{x}) + \boldsymbol{\delta}(\mathbf{x}). \tag{A7}$$

Here, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a deterministic projection matrix that models the relation between centers and encoding features. The noise representation $\boldsymbol{\delta}(\mathbf{x}) \in \mathbb{R}^M$ models the *variations* in the encoding features, and is the source of randomness in $\boldsymbol{\psi}(\mathbf{x})$ for each fixed input $\mathbf{x}$.

(4) While, in general, $\boldsymbol{\delta}(\mathbf{x})$ should be thought of as a random feature conditioned on each input $\mathbf{x}$, here we consider a noise model independent of centers. The simplest model for $\boldsymbol{\delta}(\mathbf{x})$ with input correlations is a linear random feature model $\boldsymbol{\delta}(\mathbf{x}) = \mathbf{B}\bar{\boldsymbol{\psi}}(\mathbf{x}) + \boldsymbol{\delta}_0$ where $\mathbf{B} \in \mathbb{R}^{M\times N}$ is a random matrix, and $\boldsymbol{\delta}_0$ is a random vector which are drawn independently from their respective distributions. The study of this model requires computing another quenched average over random features $\mathbf{B}$ and is left for future work. See Refs. [15,35] for the analysis of a similar model in ridge regression setting.

(5) In particular, we will apply our theory to the case where $\boldsymbol{\psi}(\mathbf{x}) = \bar{\boldsymbol{\psi}}(\mathbf{x}) + \boldsymbol{\delta}$ where $\boldsymbol{\delta} \in \mathbb{R}^N$ models the variances in the inputs that are irrelevant to the target. The dimensionalities of the features $\boldsymbol{\psi}$ and $\bar{\boldsymbol{\psi}}$ are chosen to be the same for notational convenience, however, can be trivially applied to a more general setting with $\boldsymbol{\psi} = \mathbf{A}\bar{\boldsymbol{\psi}} + \boldsymbol{\delta}$, since our final result will depend only on the covariance matrices in Eq. (A6) and the coding direction $\bar{\mathbf{w}}$.

1. Replica calculation

We set up a general framework for our calculation. The measure for the weights are denoted as $d\mu(\mathbf{w})$, and the loss function $\ell(\mathbf{w}\cdot\boldsymbol{\psi} - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}})$ is kept generic. Then, the partition function is given by

$$\begin{aligned}
Z &= \int d\tilde{\mu}(\mathbf{w}) e^{-\frac{\beta}{\lambda}\sum_{\mu=1}^{P}\ell(\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)},\\
d\tilde{\mu}(\mathbf{w}) &\equiv d\mu(\mathbf{w}) e^{-\beta\mathbf{w}^\top\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}}},
\end{aligned} \tag{A8}$$

where we defined $d\tilde{\mu}(\mathbf{w})$ by including a source term to extract the mean and variance of the optimized weights.

First, we integrate in auxiliary variables to simplify the computation:

$$H_\mu \equiv \mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu \tag{A9}$$

and get

$$\begin{aligned}
Z &= \int d\tilde{\mu}(\mathbf{w})\prod_\mu dH_\mu\, e^{-\frac{\beta}{\lambda}\ell(H_\mu)}\delta(H_\mu - \mathbf{w}\cdot\boldsymbol{\psi}^\mu + \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)\\
&= \int d\tilde{\mu}(\mathbf{w})[d\mathcal{H}_\mu][d\hat{H}_\mu]\, e^{i\sum_\mu \hat{H}_\mu H_\mu} e^{-i\sum_\mu \hat{H}_\mu(\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)},
\end{aligned} \tag{A10}$$

where we defined $d\mathcal{H}_\mu \equiv dH_\mu\, e^{-\frac{\beta}{\lambda}\ell(H_\mu)}$ and introduced the convention where $[dx^{a,b,\cdots}]$ represents product over free indices $[dx^{a,b,\cdots}] \equiv \prod_{a,b,\cdots} dx^{a,b,\cdots}$. For example, $[d\mathcal{H}_\mu] \equiv \prod_\mu d\mathcal{H}_\mu$ and $[d\hat{H}_\mu] \equiv \prod_\mu d\hat{H}_\mu$.

Next, we replicate the partition function to perform averages over the data distribution:

$$\begin{aligned}
Z^n = \int &[d\tilde{\mu}(\mathbf{w}^a)]\big[d\mathcal{H}_\mu^a\big]\big[d\hat{H}_\mu^a\big]\\
&\times\ e^{i\sum_{a,\mu}\hat{H}_\mu^a H_\mu^a} e^{-i\sum_{a,\mu}\hat{H}_\mu^a(\mathbf{w}^a\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)}.
\end{aligned} \tag{A11}$$

Next, we define the $(2N)$-dimensional feature vector $\boldsymbol{\gamma}^\mu$ and $(2N)$-dimensional weight vector $\mathcal{W}^a$ as

$$\boldsymbol{\gamma}^\mu = \begin{pmatrix}\boldsymbol{\psi}^\mu\\ \bar{\boldsymbol{\psi}}^\mu\end{pmatrix}, \quad \mathcal{W}^a = \begin{pmatrix}\mathbf{w}^a\\ -\bar{\mathbf{w}}\end{pmatrix}, \tag{A12}$$

so that for each sample, we have

$$\mathbf{w}^a\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu = \mathcal{W}^a\cdot\boldsymbol{\gamma}^\mu. \tag{A13}$$

At this point, we make the Gaussian assumption where the statistics of $\boldsymbol{\gamma}^\mu$ can be captured by its first and second moments:

$$\langle \boldsymbol{\gamma} \rangle = 0, \quad \langle \boldsymbol{\gamma}\boldsymbol{\gamma}^\top \rangle = \begin{pmatrix} \langle \boldsymbol{\psi}\boldsymbol{\psi}^\top \rangle & \langle \boldsymbol{\psi}\bar{\boldsymbol{\psi}}^\top \rangle \\ \langle \bar{\boldsymbol{\psi}}\boldsymbol{\psi}^\top \rangle & \langle \bar{\boldsymbol{\psi}}\bar{\boldsymbol{\psi}}^\top \rangle \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\psi}}/N & \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}/N \\ \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}^\top/N & \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}}/N \end{pmatrix} \equiv \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}, \tag{A14}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \in \mathbb{R}^{(2N)\times(2N)}$ is its covariance. With this notation, the partition function simplifies to

$$Z^n = \int [d\tilde{\mu}(\mathbf{w}^a)] \big[d\mathcal{H}^a_\mu\big] \big[d\hat{H}^a_\mu\big]\, e^{i\sum_{a,\mu} \hat{H}^a_\mu H^a_\mu} e^{-i\sum_{a,\mu} \hat{H}^a_\mu \boldsymbol{\mathcal{W}}^a\cdot\boldsymbol{\gamma}^\mu}, \tag{A15}$$

and with the assumption that each sample $\boldsymbol{\gamma}^\mu$ is drawn i.i.d. from the normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}})$, its dataset average over $\boldsymbol{\gamma}^\mu$ yields

$$\langle Z^n \rangle = \int [d\tilde{\mu}(\mathbf{w}^a)] \left[\int [d\mathcal{H}^a][d\hat{H}^a]\, e^{i\sum_a \hat{H}^a H^a} \langle e^{-i\sum_a \hat{H}^a \boldsymbol{\mathcal{W}}^a\cdot\boldsymbol{\gamma}} \rangle_{\boldsymbol{\gamma}} \right]^P. \tag{A16}$$

We define the order parameters $q^a = \boldsymbol{\mathcal{W}}^a \cdot \boldsymbol{\gamma}$ and consider their first and second moments:

$$\langle q^a \rangle = 0, \quad C^{ab} = \langle q^a q^b \rangle = \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b. \tag{A17}$$

Then the quenched average over $\boldsymbol{\gamma} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}})$ gives

$$\begin{aligned} \langle Z^n \rangle &= \int [dC^{ab}][d\tilde{\mu}(\mathbf{w}^a)] \delta(C^{ab} - \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b) \left[\int [d\mathcal{H}^a][d\hat{H}^a]\, e^{i\sum_a \hat{H}^a H^a} e^{-\frac{1}{2}\sum_{a,b} \hat{H}^a C^{ab} \hat{H}^a} \right]^P \\ &= \int [dC^{ab}][d\tilde{\mu}(\mathbf{w}^a)] \delta(C^{ab} - \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b) \left[\int [d\mathcal{H}^a]\, e^{-\frac{1}{2}\sum_{a,b} H^a (C^{ab})^{-1} H^b - \frac{1}{2}\log\det(C^{ab})} \right]^P, \end{aligned} \tag{A18}$$

where in the last line we performed the integral over $[d\hat{H}^a]$. Furthermore, we integrate in new variables $\hat{C}^{ab}$ to replace the delta function:

$$\begin{aligned} \langle Z^n \rangle &= \int [dC^{ab}][d\hat{C}^{ab}][d\tilde{\mu}(\mathbf{w}^a)] e^{i\sum_{ab} \hat{C}^{ab}\left(C^{ab} - \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b\right)} \left[\int [d\mathcal{H}^a]\, e^{-\frac{1}{2}\sum_{a,b} H^a \left(C^{ab}\right)^{-1} H^b - \frac{1}{2}\log\det(C^{ab})} \right]^P \\ &\equiv \int [dC^{ab}][d\hat{C}^{ab}] e^{-n\frac{\beta N}{2}(G_0 + \alpha G_1)}, \end{aligned} \tag{A19}$$

where

$$\begin{aligned} G_0 &= -\frac{2}{n\beta N} \log\left(\int [d\tilde{\mu}(\mathbf{w}^a)] e^{i\sum_{ab} \hat{C}^{ab}\left(C^{ab} - \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b\right)} \right), \\ [d\tilde{\mu}(\mathbf{w}^a)] &= \prod_a d\mu(\mathbf{w}^a) e^{-\beta \mathbf{w}^{a\top} \mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}} \bar{\mathbf{w}}}, \\ \boldsymbol{\mathcal{W}}^a &= \begin{pmatrix} \mathbf{w}^a \\ -\bar{\mathbf{w}} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \equiv \frac{1}{N} \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\psi}} & \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}^\top & \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} \end{pmatrix}, \end{aligned} \tag{A20}$$

and

$$\begin{aligned} G_1 &= -\frac{2}{n\beta} \log\left(\int [d\mathcal{H}^a]\, e^{-\frac{1}{2}\sum_{a,b} H^a (C^{ab})^{-1} H^b - \frac{1}{2}\log\det(C^{ab})} \right), \\ [d\mathcal{H}^a] &= \prod_a dH^a\, e^{-\frac{\beta}{\lambda}\ell(H^a)}. \end{aligned} \tag{A21}$$

The coefficients were chosen so that G_0 and G_1 remain $\mathcal{O}(1)$ in the thermodynamic limit $N \to \infty$ while $\alpha \equiv P/N \sim \mathcal{O}(1)$. Next, we compute these two terms under the replica symmetric ansatz:

$$\begin{aligned} \mathbf{C} &= \frac{1}{\beta}(Q\mathbf{I}_n + \beta C \mathbf{1}_n \mathbf{1}_n^\top), \\ \hat{\mathbf{C}} &= \frac{\beta N}{2i}(\hat{Q}\mathbf{I}_n + \beta \hat{C} \mathbf{1}_n \mathbf{1}_n^\top). \end{aligned} \tag{A22}$$

The G_0 term: Representing $\mathbf{C}$ and $\hat{\mathbf{C}}$ as $n \times n$-matrices with entries C^{ab} and $\hat{C}^{ab}$, the G_0 term can be written as

$$G_0 = -\frac{2i}{n\beta N} \mathrm{Tr}\hat{\mathbf{C}}\mathbf{C} - \frac{2}{n\beta N} \log\left(\int [d\tilde{\mu}(\mathbf{w}^a)] e^{-i\sum_{ab} \hat{C}^{ab} \boldsymbol{\mathcal{W}}^a \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}} \boldsymbol{\mathcal{W}}^b} \right). \tag{A23}$$

Choosing $d\mu(\mathbf{w}^a) = (\frac{\beta}{2\pi})^{N/2} e^{-\frac{\beta}{2}\|\mathbf{w}^a\|^2}$, the measure simplifies to

$$[d\tilde{\mu}(\mathbf{w}^a)] = \prod_a d\mu(\mathbf{w}^a) e^{-\beta \mathbf{w}^{a\top} \mathbf{J}_{\psi\bar{\psi}} \bar{\mathbf{w}}} = \left(\frac{\beta}{2\pi}\right)^{nN/2} \prod_a d\mathbf{w}^a e^{-\frac{\beta}{2}\left(\|\mathbf{w}^a\|^2 + 2\mathbf{w}^{a\top} \mathbf{J}_{\psi\bar{\psi}} \bar{\mathbf{w}}\right)}. \tag{A24}$$

Taking the constants out

$$\begin{aligned} G_0 &= -\frac{2i}{n\beta N} \mathrm{Tr} \hat{\mathbf{C}}\mathbf{C} - \frac{2}{n\beta N} \log I_0, \\ I_0 &\equiv \left(\frac{\beta}{2\pi}\right)^{nN/2} \int [d\mathbf{w}^a] e^{-\frac{\beta}{2}\sum_a \left(\|\mathbf{w}^a\|^2 + 2\mathbf{w}^{a\top} \mathbf{J}_{\psi\bar{\psi}} \bar{\mathbf{w}}\right) - i \sum_{ab} \hat{C}^{ab} \mathcal{W}^a \mathbf{\Sigma}_\Gamma \mathcal{W}^b}. \end{aligned} \tag{A25}$$

The integral I_0 must be evaluated based on the structure of $\mathbf{\Sigma}_\Gamma$ and $\mathcal{W}^a$ which are given above. Then the term $\mathcal{W}^a \mathbf{\Sigma}_\Gamma \mathcal{W}^b$ can be expanded as

$$\mathcal{W}^a \mathbf{\Sigma}_\Gamma \mathcal{W}^b = \mathbf{w}^a \mathbf{\Sigma}_\psi \mathbf{w}^b + \bar{\mathbf{w}} \mathbf{\Sigma}_{\bar{\psi}} \bar{\mathbf{w}} - (\mathbf{w}^a + \mathbf{w}^b) \mathbf{\Sigma}_{\psi\bar{\psi}} \bar{\mathbf{w}}. \tag{A26}$$

Then, the exponent in I_0 can be written in a compact form by defining

$$\begin{aligned} &\mathbf{W} = (\mathbf{w}^1 \quad \mathbf{w}^2 \quad \dots \quad \mathbf{w}^a \quad \dots \quad \mathbf{w}^n), \\ &\bar{\mathbf{W}} = (\bar{\mathbf{w}} \quad \bar{\mathbf{w}} \quad \dots \quad \bar{\mathbf{w}} \quad \dots \quad \bar{\mathbf{w}}) = \bar{\mathbf{w}} \otimes \mathbf{1}_n, \\ &\mathbf{X} = \frac{2i}{\beta} \mathbf{\Sigma}_\psi \otimes \hat{\mathbf{C}}, \quad \mathbf{Y} = \frac{2i}{\beta} \mathbf{\Sigma}_{\psi\bar{\psi}} \otimes \hat{\mathbf{C}}, \quad \mathbf{Z} = \frac{2i}{\beta} \mathbf{\Sigma}_{\bar{\psi}} \otimes \hat{\mathbf{C}}, \quad \mathbf{J} = \mathbf{J}_{\psi\bar{\psi}} \otimes \mathbf{I}_n, \end{aligned} \tag{A27}$$

yielding

$$-i \sum_{a,b} \hat{C}^{ab} \mathcal{W}^a \mathbf{\Sigma}_\Gamma \mathcal{W}^b = -\frac{\beta}{2} (\mathbf{W}^\top \mathbf{X} \mathbf{W} - 2\mathbf{W}^\top \mathbf{Y} \bar{\mathbf{W}} + \bar{\mathbf{W}}^\top \mathbf{Z} \bar{\mathbf{W}}). \tag{A28}$$

Plugging this in the integral, we get

$$I_0 = \left(\frac{\beta}{2\pi}\right)^{nN/2} \int d\mathbf{W} e^{-\frac{\beta}{2}\left(\mathbf{W}^\top (\mathbf{I}+\mathbf{X}) \mathbf{W} + 2\mathbf{W}^\top (\mathbf{J}-\mathbf{Y}) \bar{\mathbf{W}} + \bar{\mathbf{W}}^\top \mathbf{Z} \bar{\mathbf{W}}\right)}. \tag{A29}$$

The measure also becomes $[d\mathbf{w}^a] = d\mathbf{W}$ and the integral over nN-dimensional space can be simply evaluated to

$$\log I_0 = -\tfrac{1}{2} \log\det(\mathbf{I} + \mathbf{X}) - \frac{\beta}{2} \bar{\mathbf{W}}^\top [\mathbf{Z} - (\mathbf{J} - \mathbf{Y})^\top (\mathbf{I} + \mathbf{X})^{-1} (\mathbf{J} - \mathbf{Y})] \bar{\mathbf{W}}. \tag{A30}$$

Plugging in the replica symmetric ansatz $\hat{\mathbf{C}} = \frac{\beta N}{2i} (\hat{Q} \mathbf{I}_n + \beta \hat{C} \mathbf{1}_n \mathbf{1}_n^\top)$ Eq. (A22), matrices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ simplify to

$$\begin{aligned} \mathbf{X} &= \hat{Q} \mathbf{\Sigma}_\psi \otimes \mathbf{I}_n + \beta \hat{C} \mathbf{\Sigma}_\psi \otimes \mathbf{1}_n \mathbf{1}_n^\top, \\ \mathbf{Y} &= \hat{Q} \mathbf{\Sigma}_{\psi\bar{\psi}} \otimes \mathbf{I}_n + \beta \hat{C} \mathbf{\Sigma}_{\psi\bar{\psi}} \otimes \mathbf{1}_n \mathbf{1}_n^\top, \\ \mathbf{Z} &= \hat{Q} \mathbf{\Sigma}_{\bar{\psi}} \otimes \mathbf{I}_n + \beta \hat{C} \mathbf{\Sigma}_{\bar{\psi}} \otimes \mathbf{1}_n \mathbf{1}_n^\top. \end{aligned} \tag{A31}$$

The $\log\det(\mathbf{I} + \mathbf{X})$ simplifies to

$$\begin{aligned} &\log\det[(\mathbf{I}_N + \hat{Q} \mathbf{\Sigma}_\psi) \otimes \mathbf{I}_n] + \log\det[\mathbf{I} + \beta \hat{C} \mathbf{\Sigma}_\psi (\mathbf{I}_N + \hat{Q} \mathbf{\Sigma}_\psi)^{-1} \otimes \mathbf{1}_n \mathbf{1}_n^\top] \\ &= n \log\det \mathbf{G} + \mathrm{Tr} \log(\mathbf{I} + \beta \hat{C} \mathbf{\Sigma}_\psi \mathbf{G}^{-1} \otimes \mathbf{1}_n \mathbf{1}_n^\top) \\ &= n\beta N \left(\frac{1}{\beta N} \log\det \mathbf{G} + \hat{C} \frac{1}{N} \mathrm{Tr} \mathbf{\Sigma}_\psi \mathbf{G}^{-1} \right) + \mathcal{O}(n^2), \end{aligned} \tag{A32}$$

where we defined $\mathbf{G} \equiv \mathbf{I}_N + \hat{Q} \mathbf{\Sigma}_\psi$, used the identity $\log\det \mathbf{A} = \mathrm{Tr} \log \mathbf{A}$ (second line), and Taylor expanded the log in small n (last line).

Next, we compute $(\mathbf{I} + \mathbf{X})^{-1}$ in the limit $n \to 0$:

$$\begin{aligned} (\mathbf{I} + \mathbf{X})^{-1} &= ((\mathbf{I}_N + \hat{Q} \mathbf{\Sigma}_\psi)^{-1} \otimes \mathbf{I}_n)(\mathbf{I} + \beta \hat{C} \mathbf{\Sigma}_\psi \mathbf{G}^{-1} \otimes \mathbf{1}_n \mathbf{1}_n^\top)^{-1} \\ &= (\mathbf{G}^{-1} \otimes \mathbf{I}_n)(\mathbf{I} - \beta \hat{C} \mathbf{\Sigma}_\psi \mathbf{G}^{-1} \otimes \mathbf{1}_n \mathbf{1}_n^\top + \mathcal{O}(n)) \\ &= \mathbf{G}^{-1} \otimes \mathbf{I}_n - \beta \hat{C} \mathbf{G}^{-1} \mathbf{\Sigma}_\psi \mathbf{G}^{-1} \otimes \mathbf{1}_n \mathbf{1}_n^\top + \mathcal{O}(n). \end{aligned} \tag{A33}$$

Note that the second and third terms in $\log I_0$ are of the form $\bar{\mathbf{w}} \otimes \mathbf{1}_n(\dots)\bar{\mathbf{w}} \otimes \mathbf{1}_n$, meaning that only terms like $\mathbf{A} \otimes \mathbf{I}_n$ in parentheses survive in the leading order. Therefore, the second term in $\log I_0$ becomes

$$-\frac{n\beta}{2}\bar{\mathbf{w}}^\top[\hat{Q}\mathbf{\Sigma}_{\bar{\psi}} - (\mathbf{J}_{\psi\bar{\psi}} - \hat{Q}\mathbf{\Sigma}_{\psi\bar{\psi}})^\top\mathbf{G}^{-1}(\mathbf{J}_{\psi\bar{\psi}} - \hat{Q}\mathbf{\Sigma}_{\psi\bar{\psi}})]\bar{\mathbf{w}} + \mathcal{O}(n^2). \tag{A34}$$

Then, $\log I_0$ becomes

$$-\frac{2}{n\beta N}\log I_0 = \frac{1}{\beta N}\log\det\mathbf{G} + \hat{C}\mathrm{tr}\mathbf{\Sigma}_{\psi}\mathbf{G}^{-1} + \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top[\hat{Q}\mathbf{\Sigma}_{\bar{\psi}} - (\mathbf{J}_{\psi\bar{\psi}} - \hat{Q}\mathbf{\Sigma}_{\psi\bar{\psi}})^\top\mathbf{G}^{-1}(\mathbf{J}_{\psi\bar{\psi}} - \hat{Q}\mathbf{\Sigma}_{\psi\bar{\psi}})], \tag{A35}$$

where we introduced the normalized trace operation

$$\mathrm{tr}\mathbf{M} = \frac{1}{\dim\mathbf{M}}\mathrm{Tr}\mathbf{M}. \tag{A36}$$

Finally, noting that the term $-\frac{2i}{n\beta N}\mathrm{Tr}\hat{\mathbf{C}}\mathbf{C}$ reduces to $-\frac{1}{\beta}Q\hat{Q} - \hat{C}Q - \hat{Q}C$, and keeping terms up to $\mathcal{O}(\mathbf{J}_{\psi\bar{\psi}})$, G_0 becomes

$$\begin{aligned} G_0 =& \frac{1}{\beta}\left(\frac{1}{N}\log\det\mathbf{G} - Q\hat{Q}\right) - \hat{C}(Q - \mathrm{tr}\mathbf{\Sigma}_{\psi}\mathbf{G}^{-1}) - \hat{Q}(C - \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top[\mathbf{\Sigma}_{\bar{\psi}} - \mathbf{\Sigma}_{\psi\bar{\psi}}^\top(\hat{Q}\mathbf{G}^{-1})\mathbf{\Sigma}_{\psi\bar{\psi}}]) + 2\mathrm{tr}\mathbf{J}_{\psi\bar{\psi}}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\mathbf{\Sigma}_{\psi\bar{\psi}}^\top(\hat{Q}\mathbf{G}^{-1}), \\ \mathbf{G} =& \mathbf{I} + \hat{Q}\mathbf{\Sigma}_{\psi}. \end{aligned} \tag{A37}$$

At this point, we can make further simplifications by assuming that $\mathbf{\Sigma}_{\psi}$ is full-rank and using the identity

$$\hat{Q}\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{G}^{-1})\mathbf{\Sigma}_{\psi}^{-1} = \mathbf{\Sigma}_{\psi}^{-1}(\mathbf{I} - \mathbf{G}^{-1}). \tag{A38}$$

Then, we get

$$G_0 = -\hat{C}(Q - \mathrm{tr}\mathbf{\Sigma}_{\psi}\mathbf{G}^{-1}) - \hat{Q}(C - \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top(\mathbf{\Sigma}_{\bar{\psi}} - \mathbf{\Sigma}_{\psi\bar{\psi}}^\top\mathbf{\Sigma}_{\psi}^{-1}\mathbf{\Sigma}_{\psi\bar{\psi}})) + \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top(\mathbf{\Sigma}_{\psi}^{-1}\mathbf{\Sigma}_{\psi\bar{\psi}})^\top(\mathbf{I} - \mathbf{G}^{-1})(\mathbf{\Sigma}_{\psi}^{-1}\mathbf{\Sigma}_{\psi\bar{\psi}} + 2\mathbf{J}_{\psi\bar{\psi}}). \tag{A39}$$

Finally, we define the irreducible error

$$\begin{aligned} E_\infty &= \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\mathbf{\Sigma}_{\bar{\psi}} - \mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top\mathbf{\Sigma}_{\psi}, \\ \mathbf{w}_* &= \mathbf{\Sigma}_{\psi}^{-1}\mathbf{\Sigma}_{\psi\bar{\psi}}\bar{\mathbf{w}}, \end{aligned} \tag{A40}$$

where $\mathbf{w}_*$ is the target weights projected on the learnable subspace and $\mathbf{w}_*{}^\top\mathbf{\Sigma}_{\psi}\mathbf{w}_*$ is the maximum explainable variance by the model. With these definitions, the equation for G_0 finally simplifies to

$$\begin{aligned} G_0 &= -\hat{C}(Q - \mathrm{tr}\mathbf{\Sigma}_{\psi}\mathbf{G}^{-1}) - \hat{Q}(C - E_\infty) + \mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top(\mathbf{I} - \mathbf{G}^{-1}) + 2\mathrm{tr}\mathbf{J}_{\psi\bar{\psi}}\bar{\mathbf{w}}\mathbf{w}_*{}^\top(\mathbf{I} - \mathbf{G}^{-1}), \\ \mathbf{G} &= \mathbf{I} + \hat{Q}\mathbf{\Sigma}_{\psi}, \\ \mathbf{w}_* &= \mathbf{\Sigma}_{\psi}^{-1}\mathbf{\Sigma}_{\psi\bar{\psi}}\bar{\mathbf{w}}, \\ E_\infty &= \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\mathbf{\Sigma}_{\bar{\psi}} - \mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top\mathbf{\Sigma}_{\psi}. \end{aligned} \tag{A41}$$

The G_1 term: The G_1 term is given by

$$\begin{aligned} G_1 &= -\frac{2}{n\beta}\left(-\frac{1}{2}\log\det(\mathbf{C}) + \log I_1\right), \\ I_1 &= \int[d\mathcal{H}^a]e^{-\frac{1}{2}\sum_{a,b}H^a(C^{ab})^{-1}H^b}, \quad [d\mathcal{H}^a] = \prod_a dH^a\, e^{-\frac{\beta}{\lambda}\ell(H^a)}. \end{aligned} \tag{A42}$$

Under the replica symmetric ansatz $\mathbf{C} = \frac{1}{\beta}(Q\mathbf{I}_n + \beta C\mathbf{1}_n\mathbf{1}_n^\top)$:

$$\begin{aligned} \log\det(\mathbf{C}) &= n\left(\log(Q/\beta) + \frac{\beta C}{Q}\right) + \mathcal{O}(n), \\ \mathbf{C}^{-1} &= \frac{\beta}{Q}\mathbf{I}_n - \frac{\beta^2 C}{Q^2}\mathbf{1}_n\mathbf{1}_n^\top + \mathcal{O}(n), \end{aligned} \tag{A43}$$

where we used the identity $\log\det\mathbf{A} = \mathrm{Tr}\log\mathbf{A}$ for the first expression and used the Sherman-Morrison formula for the second expression. Hence, I_1 becomes

$$I_1 = \int[d\mathcal{H}^a]e^{-\frac{\beta}{2Q}\sum_a(H^a)^2 + \frac{\beta^2 C}{2Q^2}\sum_{a,b}H^aH^b}. \tag{A44}$$

Using the Hubbard-Stratonovich transformation, this integral can be written as

$$\begin{aligned} I_1(C, C_0) &= \int DT \left[\int d\mathcal{H} e^{-\frac{\beta}{2Q}H^2 + \frac{\beta\sqrt{C}}{Q}HT} \right]^n \\ &= \int DT \left[e^{\frac{\beta C}{2Q}T^2} \int d\mathcal{H} e^{-\frac{\beta}{2Q}(H-\sqrt{C}T)^2} \right]^n \\ &= \int DT [e^{\frac{\beta C}{2Q}T^2 + \frac{1}{2}\log(Q/\beta) + \log z(T)}]^n, \end{aligned} \tag{A45}$$

where $DT = \frac{dT}{\sqrt{2\pi}} e^{-T^2/2}$ is the standard Gaussian measure and we defined

$$z(T) = \int d\mathcal{H} \sqrt{\frac{\beta}{Q}} e^{-\frac{\beta}{2Q}(H-\sqrt{C}T)^2} = \int dH \sqrt{\frac{\beta C}{Q}} e^{-\frac{\beta C}{2Q}(H-T)^2 - \frac{\beta}{\lambda}\ell(H\sqrt{C})}, \tag{A46}$$

where we redefined $H \to H\sqrt{C}$. Using $\int DT A^n \approx \int DT (1 + n \log A) = 1 + n \int DT \log A$ in the limit $n \to 0$, we obtain

$$\log I_1 = n\frac{\beta C}{2Q} + \frac{n}{2}\log(Q/\beta) + n \int DT \log z(T). \tag{A47}$$

Inserting this expression in G_1, we simplify

$$\begin{aligned} G_1 &= -\frac{2}{n\beta}\left(-\frac{1}{2}\log\det(\mathbf{C}) + \log I_1 \right), \\ &= \frac{1}{\beta}\log(Q/\beta) + \frac{C}{Q} - \left(\frac{C}{Q} + \frac{1}{\beta}\log(Q/\beta) + \frac{2}{\beta}\int DT \log z(T) \right) \\ &= -\frac{2}{\beta}\int DT \log z(T). \end{aligned} \tag{A48}$$

To evaluate this integral, we first express $z(T)$ as

$$z(T) = \int dH \sqrt{\frac{\beta C}{Q}} e^{-\beta L(H)}, \quad L(H) = \frac{C}{2Q}\left[(H-T)^2 + \frac{2Q}{\lambda C}\ell(H\sqrt{C}) \right]. \tag{A49}$$

Note that as $\beta \to \infty$, the integral concentrates around the saddle point $H^* = \arg\min_H L(H)$ of $L(H)$. Plugging this saddle point back, G_1 becomes

$$G_1 = -\frac{1}{\beta}\log\left(\frac{\beta C}{Q}\right) + 2\langle L^*(T)\rangle_T, \tag{A50}$$

where

$$L^*(T) \equiv L(H^*(T)), \quad H^*(T) = \arg\min_H \frac{C}{2Q}\left[(H-T)^2 + \frac{2Q}{\lambda C}\ell(H\sqrt{C}) \right]. \tag{A51}$$

We will consider the ε-insensitive loss function:

$$\ell(x) = \frac{1}{n!}\max(0, |x| - \varepsilon)^n. \tag{A52}$$

Then, we have

$$L(H) = \frac{C}{2Q}\left[(H-T)^2 + \zeta\frac{2}{n!}\max(0, |H| - \tilde{\varepsilon})^n \right], \tag{A53}$$

where we defined the following effective quantities:

$$\tilde{\varepsilon} \equiv \frac{\varepsilon}{\sqrt{C}}, \quad \zeta \equiv \frac{Q}{\lambda}C^{\frac{n-2}{2}}. \tag{A54}$$

Next, we solve for the saddle point H^* for general n.

(1) *When* $|\boldsymbol{H}| \leqslant \tilde{\boldsymbol{\varepsilon}}$, the loss function does not contribute and the solution is simply

$$H^*(T) = T, \quad |H| \leqslant \tilde{\varepsilon}. \tag{A55}$$

(2) *When* $|\boldsymbol{H}| > \tilde{\boldsymbol{\varepsilon}}$, sign of T determines the sign of H. We replace $H = \mathrm{sign}(T)|H|$ and differentiate Eq. (A53) with respect to $|H|$ to get

$$|H^*| - |T| + \zeta \frac{1}{(n-1)!}(|H^*| - \tilde{\varepsilon})^{n-1} = 0, \quad n \geqslant 1. \tag{A56}$$

Since $|H^*| > \varepsilon$, we instead consider $|H^*| = \tilde{\varepsilon} + t$ for $t > 0$. Then the solution becomes

$$H^*(T) = \mathrm{sign}(T)(\tilde{\varepsilon} + t), \quad |T| > \tilde{\varepsilon}, \tag{A57}$$

where t is the solution of

$$t + \frac{\zeta}{(n-1)!} t^{n-1} = |T| - \tilde{\varepsilon}, \quad t \geqslant 0. \tag{A58}$$

Here, $n = 0$ is a special case. Recalling $(-1)! = \Gamma(0) = \infty$, its solution is simply $t = |T| - \tilde{\varepsilon}$.

Hence, the saddle point for general n is given by

$$H^*(T) = \begin{cases} T, & |T| \leqslant \tilde{\varepsilon}, \\ \mathrm{sign}(T)(\tilde{\varepsilon} + t), & |T| > \tilde{\varepsilon}, \end{cases} \quad L(H^*) = \begin{cases} 0, & |T| \leqslant \tilde{\varepsilon}, \\ \frac{C}{2Q}\left[(\tilde{\varepsilon} + t - |T|)^2 + 2\zeta \frac{t^n}{n!}\right], & |T| > \tilde{\varepsilon}, \end{cases} \tag{A59}$$

where t is the solution of

$$t + \frac{\zeta}{(n-1)!} t^{n-1} = |T| - \tilde{\varepsilon}, \quad t \geqslant 0. \tag{A60}$$

Below, we report its solution for $n = 0, 1, 2$.

(1) $n = 0$:

$$t = \begin{cases} 0, & \tilde{\varepsilon} < |T| \leqslant \tilde{\varepsilon} + \sqrt{2\zeta} \\ |T| - \tilde{\varepsilon}, & |T| > \tilde{\varepsilon} + \sqrt{2\zeta} \end{cases}, \quad L(H^*) = \begin{cases} 0, & |T| \leqslant \tilde{\varepsilon} \\ \frac{C}{2Q}(|T| - \tilde{\varepsilon})^2, & \tilde{\varepsilon} < |T| \leqslant \tilde{\varepsilon} + \sqrt{2\zeta}. \\ \frac{C\zeta}{Q}, & |T| > \tilde{\varepsilon} + \sqrt{2\zeta} \end{cases} \tag{A61}$$

(2) $n = 1$:

$$t = \begin{cases} 0, & \tilde{\varepsilon} < |T| \leqslant \tilde{\varepsilon} + \zeta \\ |T| - \tilde{\varepsilon} - \zeta, & |T| > \tilde{\varepsilon} + \zeta \end{cases}, \quad L(H^*) = \begin{cases} 0, & |T| \leqslant \tilde{\varepsilon} \\ \frac{C}{2Q}(|T| - \tilde{\varepsilon})^2, & \tilde{\varepsilon} < |T| \leqslant \tilde{\varepsilon} + \zeta. \\ \frac{C}{2Q}[\zeta^2 + 2\zeta(|T| - \tilde{\varepsilon} - \zeta)], & |T| > \tilde{\varepsilon} + \zeta \end{cases} \tag{A62}$$

(3) $n = 2$:

$$t = \frac{|T| - \tilde{\varepsilon}}{1 + \zeta}, \quad L(H^*) = \begin{cases} 0, & |T| \leqslant \tilde{\varepsilon} \\ \frac{C}{2Q}\frac{\zeta}{1+\zeta}(|T| - \tilde{\varepsilon})^2, & |T| > \tilde{\varepsilon} \end{cases}. \tag{A63}$$

(4) $n \geqslant 3$: These cases require solving higher-order polynomial equations. However, $\zeta \sim \lambda^{-1}$ is typically large for small regularization $\lambda \approx 0$. In this limit ($\zeta \to \infty$), we have

$$t \approx \left(\frac{(n-1)!(|T| - \tilde{\varepsilon})}{\zeta}\right)^{\frac{1}{n-1}}, \tag{A64}$$

hence, for $|T| > \tilde{\varepsilon}$, we have

$$L(H^*) \approx \frac{C}{2Q}(|T| - \tilde{\varepsilon})^2 + \mathcal{O}(\zeta^{-\frac{1}{n-1}}(|T| - \tilde{\varepsilon})^{\frac{n}{n-1}}). \tag{A65}$$

This implies that the leading order behavior for $n \geqslant 3$ should be similar to the $n = 2$ case.

Now, we can evaluate the average $\langle L(H^*)\rangle_T$. First, we introduce the notation

$$\langle f(T)\rangle|_a^b \equiv \int_a^b f(T)\frac{e^{-T^2/2}}{\sqrt{2\pi}}dT. \tag{A66}$$

Hence, $\langle L(H^*)\rangle_T = \langle L(H^*)\rangle|_{-\infty}^{\infty}$. Next, we define functions $f(x)$ and $g(x)$ which we will commonly refer to

$$f(x) = \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right), \quad g(x) = 1 + x^2 - x\frac{\sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2}}}{\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right)}. \tag{A67}$$

We also need to evaluate the following integrals:

$$\begin{aligned}\left\langle (T-x)^2\right\rangle\Big|_x^\infty &= \frac{1}{2}\sqrt{\frac{2}{\pi}}\int_x^\infty (T-x)^2 e^{-T^2/2}dT = \frac{1}{2}f(x)g(x),\\ \langle (T-x)\rangle\Big|_x^\infty &= \frac{1}{2}\sqrt{\frac{2}{\pi}}\int_x^\infty (T-x)\, e^{-T^2/2}dT = \frac{1}{2}f(x)\frac{1-g(x)}{x}.\end{aligned} \tag{A68}$$

Finally, we evaluate $\langle L(H^*)\rangle_T$. While we study only the $n = 2$ case, here we compute it for all $n = 0, 1, 2$ for reference.

(1) $n = 0$:

$$\begin{aligned}\langle L(H^*)\rangle_T &= \frac{C}{Q}\left[\left\langle (T-\tilde{\varepsilon})^2\right\rangle\Big|_{\tilde{\varepsilon}}^{\tilde{\varepsilon}+\sqrt{2\zeta}} + \langle 2\zeta\rangle\Big|_{\tilde{\varepsilon}+\sqrt{2\zeta}}^{\infty}\right] = \frac{C}{Q}\left[\left\langle (T-\tilde{\varepsilon})^2\right\rangle\Big|_{\tilde{\varepsilon}}^{\infty} - \left\langle (T-\tilde{\varepsilon})^2 - 2\zeta\right\rangle\Big|_{\tilde{\varepsilon}+\sqrt{2\zeta}}^{\infty}\right]\\ &= \frac{C}{Q}\left[\left\langle (T-\tilde{\varepsilon})^2\right\rangle\Big|_{\tilde{\varepsilon}}^{\infty} - \langle (T-\tilde{\varepsilon}-\sqrt{2\zeta})^2\rangle\Big|_{\tilde{\varepsilon}+\sqrt{2\zeta}}^{\infty} - \sqrt{8\zeta}\,\langle T-\tilde{\varepsilon}-\sqrt{2\zeta}\rangle\Big|_{\tilde{\varepsilon}+\sqrt{2\zeta}}^{\infty}\right]\\ &= \frac{C}{2Q}\left[f(\tilde{\varepsilon})g(\tilde{\varepsilon}) - f(\tilde{\varepsilon}+\sqrt{2\zeta})g(\tilde{\varepsilon}+\sqrt{2\zeta}) - \sqrt{8\zeta}\, f(\tilde{\varepsilon}+\sqrt{2\zeta})\frac{1-g(\tilde{\varepsilon}+\sqrt{2\zeta})}{\tilde{\varepsilon}+\sqrt{2\zeta}}\right]\\ &= \frac{C}{2Q}\left[f(\tilde{\varepsilon})g(\tilde{\varepsilon}) - f(\tilde{\varepsilon}+\sqrt{2\zeta})g(\tilde{\varepsilon}+\sqrt{2\zeta})\left(1 + \frac{\sqrt{8\zeta}}{\tilde{\varepsilon}+\sqrt{2\zeta}}\frac{1-g(\tilde{\varepsilon}+\sqrt{2\zeta})}{g(\tilde{\varepsilon}+\sqrt{2\zeta})}\right)\right].\end{aligned} \tag{A69}$$

(2) $n = 1$:

$$\begin{aligned}\langle L(H^*)\rangle_T &= \frac{C}{Q}\left[\left\langle (T-\tilde{\varepsilon})^2\right\rangle\Big|_{\tilde{\varepsilon}}^{\tilde{\varepsilon}+\zeta} + \left\langle \zeta^2 + 2\zeta(|T| - \tilde{\varepsilon} - \zeta)\right\rangle\Big|_{\tilde{\varepsilon}+\zeta}^{\infty}\right]\\ &= \frac{C}{Q}\left[\left\langle (T-\tilde{\varepsilon})^2\right\rangle\Big|_{\tilde{\varepsilon}}^{\infty} - \left\langle (T-\tilde{\varepsilon}-\zeta)^2\right\rangle\Big|_{\tilde{\varepsilon}+\zeta}^{\infty}\right]\\ &= \frac{C}{2Q}[f(\tilde{\varepsilon})g(\tilde{\varepsilon}) - f(\tilde{\varepsilon}+\zeta)g(\tilde{\varepsilon}+\zeta)].\end{aligned} \tag{A70}$$

(3) $n = 2$:

$$\begin{aligned}\langle L(H^*)\rangle_T &= \frac{C}{Q}\frac{\zeta}{1+\zeta}\langle (T-\tilde{\varepsilon})^2\rangle\Big|_{\tilde{\varepsilon}}^{\infty}\\ &= \frac{C}{2Q}\frac{\zeta}{1+\zeta}f(\tilde{\varepsilon})g(\tilde{\varepsilon}).\end{aligned} \tag{A71}$$

Note that the standard ε-SVR corresponds to $n = 1$ case and can be analyzed in our framework. Here, we focus on $n = 2$ case for simplicity. Since we mainly consider very small ridge parameters for SVR, our calculation above demonstrates that all cases behave similarly in the limit $\zeta \to \infty$.

Plugging the result for $n = 2$ case in the expression Eq. (A50) for G_1, we get

$$G_1 = \frac{C}{Q+\lambda}f(\tilde{\varepsilon})g(\tilde{\varepsilon}) - \frac{1}{\beta}\log\left(\frac{\beta C}{Q}\right), \tag{A72}$$

where we used the definition from Eq. (A54) to get $\zeta = Q/\lambda$ for $n = 2$.

2. Combining all terms

In the thermodynamic limit $P \to \infty$, the final partition function after the replica symmetric ansatz becomes

$$\langle \log Z \rangle = \lim_{n\to 0} \frac{\langle Z^n \rangle - 1}{n} = -\frac{\beta N}{2} S^*, \tag{A73}$$

where S^* is obtained by solving the following saddle-point equation

$$\begin{aligned} S^* &= \underset{C,\hat{C},Q,\hat{Q}}{\text{extr}} \left[G_0 + \alpha G_1 + \frac{1}{\beta}\left(\frac{1}{N} \log\det \mathbf{G} - Q\hat{Q} - \alpha \log\left(\frac{\beta C}{Q} \right) \right) \right], \\ G_0 &= -\hat{C}(Q - \text{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}) - \hat{Q}(C - E_\infty) + \text{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top(\mathbf{I} - \mathbf{G}^{-1}), \\ G_1 &= \frac{C}{Q+\lambda} f(\tilde{\varepsilon}) g(\tilde{\varepsilon}), \end{aligned} \tag{A74}$$

where we have the following definitions

$$\begin{aligned} E_\infty &= \text{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} - \text{tr}\mathbf{w}_*\mathbf{w}_*{}^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \\ \mathbf{G} &= \mathbf{I} + \hat{Q}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \tilde{\varepsilon} = \frac{\varepsilon}{\sqrt{C}}, \quad \mathbf{w}_* = \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}}. \end{aligned} \tag{A75}$$

3. Saddle-point equations

We next obtain the saddle-point equations for S. Let us first minimize for $\hat{C}$:

(1) Minimizing with respect to $\hat{C}$ gives $\boxed{Q = \text{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}}$.

(2) Minimizing with respect to $\hat{Q}$ gives

$$\boxed{C = -\hat{C}\text{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1} + E_\infty + \text{tr}\big(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}}\big)\mathbf{w}_*{}^\top\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}}. \tag{A76}$$

(3) Minimizing with respect to C gives

$$\boxed{\hat{Q} = \frac{\alpha f(\tilde{\varepsilon})}{Q+\lambda}}. \tag{A77}$$

(4) Minimizing with respect to Q gives

$$\boxed{\hat{C} = -\frac{\alpha f(\tilde{\varepsilon})}{(Q+\lambda)^2} C g(\tilde{\varepsilon})}. \tag{A78}$$

(5) As a reminder, we had the following definitions:

$$\begin{aligned} f(\tilde{\varepsilon}) &= \text{erfc}\left(\frac{\tilde{\varepsilon}}{\sqrt{2}} \right), \quad g(\tilde{\varepsilon}) = 1 + \tilde{\varepsilon}^2 - \tilde{\varepsilon}\frac{\sqrt{\frac{2}{\pi}} e^{-\frac{\tilde{\varepsilon}^2}{2}}}{\text{erfc}\left(\frac{\tilde{\varepsilon}}{\sqrt{2}}\right)}, \quad \tilde{\varepsilon} = \frac{\varepsilon}{\sqrt{C}}, \\ \mathbf{G} &= \mathbf{I} + \hat{Q}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \mathbf{w}_* = \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}}, \\ E_\infty &= \text{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} - \text{tr}\mathbf{w}_*\mathbf{w}_*{}^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}. \end{aligned} \tag{A79}$$

(6) Notice that both $\hat{C}$ and $\hat{Q}$ are functions of C and Q. Hence, plugging these terms in, we get two coupled self-consistent equations as above. In general, it is challenging to solve these equations analytically.

(7) At the saddle point, we have

$$\begin{aligned} S^* &= \underset{C,\hat{C},Q,\hat{Q}}{\text{extr}} [G_0 + \alpha G_1], \\ G_0 &= \hat{Q}(E_\infty - C) + \text{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top(\mathbf{I} - \mathbf{G}^{-1}) \\ &= \hat{Q}(E_\infty + \text{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1} - C), \\ G_1 &= \frac{1}{\alpha}\hat{Q}C g(\tilde{\varepsilon}). \end{aligned} \tag{A80}$$

4. Simplifications

Let us define the following quantities which appear in the solution:

$$\tilde{\lambda} = \lambda + \mathrm{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \quad \tilde{\varepsilon} = \frac{\varepsilon}{\sqrt{C}}, \quad \tilde{\alpha} = \alpha f(\tilde{\varepsilon}),$$
$$\mathbf{G} = \mathbf{I} + \frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \gamma = \frac{\tilde{\alpha}}{\tilde{\lambda}^2}\mathrm{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}. \tag{A81}$$

Then the saddle-point equations are compactly written as

$$Q = \tilde{\lambda} - \lambda, \quad \hat{Q} = \frac{\tilde{\alpha}}{\tilde{\lambda}}, \quad \hat{C} = -\frac{\tilde{\alpha}}{\tilde{\lambda}^2}Cg(\tilde{\varepsilon}), \quad C = \frac{1}{1 - g(\tilde{\varepsilon})\gamma}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda}),$$
$$\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda}) = E_\infty + \mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}. \tag{A82}$$

In terms of these, the action becomes

$$\begin{aligned}
\langle \log Z \rangle &= -\frac{\beta N}{2}S, \\
S &= \frac{\tilde{\alpha}}{\tilde{\lambda}}[C(g(\tilde{\varepsilon}) - 1) + E_\infty + \mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}] \\
&= \frac{\tilde{\alpha}}{\tilde{\lambda}}\left[Cg(\tilde{\varepsilon})(1 - \gamma) + \frac{\tilde{\alpha}}{\tilde{\lambda}}\mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\right] \\
&= \frac{\tilde{\alpha}}{\tilde{\lambda}}Cg(\tilde{\varepsilon})(1 - \gamma) + \mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top(\mathbf{I} - \mathbf{G}^{-1})^2.
\end{aligned} \tag{A83}$$

Note that we can compute the following average quantities:

$$\begin{aligned}
-\frac{2}{N}\frac{\partial\langle \log Z\rangle}{\partial\beta} &= \left\langle \frac{1}{N}\|\mathbf{w}\|^2 + \frac{2\alpha}{\lambda}E_{\mathrm{tr}}\right\rangle = S, \\
\frac{2\lambda}{\beta N}\frac{\partial\langle \log Z\rangle}{\partial\lambda} &= \left\langle \frac{2\alpha}{\lambda}E_{\mathrm{tr}}\right\rangle = -\lambda\frac{\partial S}{\partial\lambda},
\end{aligned} \tag{A84}$$

where the average training error is defined as $\frac{1}{P}\sum_{\mu=1}^{P}\ell(\mathbf{w}\cdot\boldsymbol{\psi}^\mu - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}}^\mu)$.

5. Computing variations

We have two coupled self-consistent equations in the form of

$$\begin{aligned}
\tilde{\lambda} &= \mathcal{F}(\tilde{\lambda}, \tilde{\varepsilon}, x) \equiv \lambda + \mathrm{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \\
\tilde{\varepsilon} &= \mathcal{G}(\tilde{\lambda}, \tilde{\varepsilon}, x) \equiv \varepsilon\sqrt{1 - g(\tilde{\varepsilon})\gamma}\,\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})^{-1/2}, \\
S &= \mathcal{H}(\tilde{\lambda}, \tilde{\varepsilon}, x) \equiv \frac{\tilde{\alpha}}{\tilde{\lambda}}\left[\frac{\varepsilon^2}{\tilde{\varepsilon}^2}(g(\tilde{\varepsilon}) - 1) + E_\infty + \mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\right], \\
\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda}) &= E_\infty + \mathrm{tr}(\mathbf{w}_* + 2\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}})\mathbf{w}_*{}^\top\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1},
\end{aligned} \tag{A85}$$

where x collectively denotes any external variable. Their variations with respect to x is given by

$$\begin{aligned}
\delta\tilde{\lambda} &= \frac{1}{1 - \partial_{\tilde{\lambda}}\mathcal{F}}(\partial_{\tilde{\varepsilon}}\mathcal{F}\delta\tilde{\varepsilon} + \partial_x\mathcal{F}\delta x), \\
\delta\tilde{\varepsilon} &= \frac{1}{1 - \partial_{\tilde{\varepsilon}}\mathcal{G}}(\partial_{\tilde{\lambda}}\mathcal{G}\delta\tilde{\lambda} + \partial_x\mathcal{G}\delta x), \\
\delta S &= \partial_{\tilde{\lambda}}\mathcal{H}\delta\tilde{\lambda} + \partial_{\tilde{\varepsilon}}\mathcal{H}\delta\tilde{\varepsilon} + \partial_x\mathcal{H}\delta x.
\end{aligned} \tag{A86}$$

Solving these equations, we get

$$\begin{aligned}
\delta\tilde{\lambda} &= \frac{(1 - \partial_{\tilde{\varepsilon}}\mathcal{G})\partial_x\mathcal{F} + \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_x\mathcal{G}}{(1 - \partial_{\tilde{\lambda}}\mathcal{F})(1 - \partial_{\tilde{\varepsilon}}\mathcal{G}) - \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}\delta x, \\
\delta\tilde{\varepsilon} &= \frac{(1 - \partial_{\tilde{\lambda}}\mathcal{F})\partial_x\mathcal{G} + \partial_{\tilde{\lambda}}\mathcal{G}\partial_x\mathcal{F}}{(1 - \partial_{\tilde{\lambda}}\mathcal{F})(1 - \partial_{\tilde{\varepsilon}}\mathcal{G}) - \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}\delta x.
\end{aligned} \tag{A87}$$

We need to compute $\partial_{\tilde{\lambda}}\mathcal{F}$, $\partial_{\tilde{\varepsilon}}\mathcal{F}$, $\partial_{\tilde{\lambda}}\mathcal{G}$, $\partial_{\tilde{\varepsilon}}\mathcal{G}$. First, we compute the derivatives of $\mathcal{F}$ and $\mathcal{H}$:

$$\partial_{\tilde{\lambda}}\mathcal{F} = \gamma, \quad \partial_{\tilde{\varepsilon}}\mathcal{F} = -\frac{f'}{f}\tilde{\lambda}\gamma, \quad \partial_{\tilde{\lambda}}\mathcal{H} = -\frac{f}{\tilde{\lambda}^2}\frac{g(1-\gamma)}{1-g\gamma}\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda}), \quad \partial_{\tilde{\varepsilon}}\mathcal{H} = -\frac{f'}{\tilde{\lambda}}\frac{g\gamma}{1-g\gamma}\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda}). \tag{A88}$$

Note that

$$\frac{\partial_{\tilde{\varepsilon}}\mathcal{H}}{\partial_{\tilde{\lambda}}\mathcal{H}} = -\frac{\partial_{\tilde{\varepsilon}}\mathcal{F}}{1-\partial_{\tilde{\lambda}}\mathcal{F}} = \frac{f'\tilde{\lambda}}{f}\frac{\gamma}{1-\gamma}. \tag{A89}$$

Inserting the variation of $\delta\tilde{\lambda}$, we get the simplification

$$\begin{aligned}\delta S &= \partial_{\tilde{\lambda}}\mathcal{H}\left(\delta\tilde{\lambda} - \frac{\partial_{\tilde{\varepsilon}}\mathcal{F}}{1-\partial_{\tilde{\lambda}}\mathcal{F}}\delta\tilde{\varepsilon}\right) + \partial_x\mathcal{H}\delta x = \left(\frac{\partial_{\tilde{\lambda}}\mathcal{H}}{1-\gamma}\partial_x\mathcal{F} + \partial_x\mathcal{H}\right)\delta x \\ &= \left(-\frac{fg}{\tilde{\lambda}^2}\frac{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}{1-g\gamma}\partial_x\mathcal{F} + \partial_x\mathcal{H}\right)\delta x = \left(-f(\tilde{\varepsilon})g(\tilde{\varepsilon})\frac{C}{\tilde{\lambda}^2}\partial_x\mathcal{F} + \partial_x\mathcal{H}\right)\delta x.\end{aligned} \tag{A90}$$

The observables from the action hence only depend on $\mathcal{F}$. This provides a huge simplification for computing first-order derivatives of $\langle \log Z\rangle$.

6. Observables

Here, we compute the expected values of the observables from the partition function using our calculations above. First, we are interested in the following observables from the partition function:

$$\begin{aligned}\frac{2\alpha}{\lambda}E_{\rm tr} &= \frac{2\lambda}{\beta N}\frac{\partial}{\partial\lambda}\langle\log Z\rangle = -\lambda\frac{\partial S}{\partial\lambda} = \frac{\tilde{\alpha}}{\tilde{\lambda}}Cg(\tilde{\varepsilon})\frac{\lambda}{\tilde{\lambda}}, \\ \frac{1}{N}\langle\bar{\mathbf{w}}\mathbf{w}^\top\rangle &= -\frac{1}{\beta N}\frac{\partial}{\partial\mathbf{J}_{\psi\bar{\psi}}}\langle\log Z\rangle = \frac{1}{2}\frac{\partial S}{\partial\mathbf{J}_{\psi\bar{\psi}}} = \frac{\tilde{\alpha}}{\tilde{\lambda}}\bar{\mathbf{w}}\mathbf{w}_*{}^\top\boldsymbol{\Sigma}_{\psi}\mathbf{G}^{-1} = \frac{1}{N}(\mathbf{I}-\mathbf{G}^{-1})\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\bar{\mathbf{w}}\bar{\mathbf{w}}^\top.\end{aligned} \tag{A91}$$

From these observables we also obtain

$$\begin{aligned}\langle\mathbf{w}\rangle &= (\mathbf{I}-\mathbf{G}^{-1})\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\bar{\mathbf{w}} = N\frac{\tilde{\alpha}}{\tilde{\lambda}}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\bar{\mathbf{w}}, \\ \frac{1}{N}\langle\|\mathbf{w}\|^2\rangle &= S - \frac{2\alpha}{\lambda}E_{\rm tr} = \frac{\tilde{\alpha}}{\tilde{\lambda}}Cg(\tilde{\varepsilon})\left(1-\gamma-\frac{\lambda}{\tilde{\lambda}}\right) + \frac{1}{N}\bar{\mathbf{w}}^\top\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)^\top(\mathbf{I}-\mathbf{G}^{-1})^2\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\bar{\mathbf{w}}, \\ \frac{1}{N}\langle\mathbf{w}\mathbf{w}^\top\rangle - \frac{1}{N}\langle\mathbf{w}\rangle\langle\mathbf{w}\rangle^\top &= Cg(\tilde{\varepsilon})\frac{\tilde{\alpha}}{\tilde{\lambda}^2}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\psi}\mathbf{G}^{-1},\end{aligned} \tag{A92}$$

where we used the identity $\delta \equiv 1-\gamma-\frac{\lambda}{\tilde{\lambda}} = \frac{\mathrm{Tr}\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\psi}\mathbf{G}^{-1}}{\tilde{\lambda}}$ at the last line.

Finally, we compute the generalization error as

$$\begin{aligned}E_g &= \langle\mathbb{E}(\mathbf{w}\cdot\boldsymbol{\psi} - \bar{\mathbf{w}}\cdot\bar{\boldsymbol{\psi}})^2\rangle = \langle\mathbf{w}^\top\boldsymbol{\Sigma}_{\psi}\mathbf{w}\rangle + \langle\bar{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\bar{\psi}}\bar{\mathbf{w}}\rangle - 2\langle\mathbf{w}^\top\boldsymbol{\Sigma}_{\psi\bar{\psi}}\bar{\mathbf{w}}\rangle \\ &= Cg(\tilde{\varepsilon})\gamma + \bar{\mathbf{w}}^\top\big[\boldsymbol{\Sigma}_{\bar{\psi}} - 2\boldsymbol{\Sigma}_{\psi\bar{\psi}}^\top(\mathbf{I}-\mathbf{G}^{-1})\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big) + \big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)^\top\boldsymbol{\Sigma}_{\psi}(\mathbf{I}-\mathbf{G}^{-1})^2\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\big]\bar{\mathbf{w}} \\ &= Cg(\tilde{\varepsilon})\gamma + \bar{\mathbf{w}}^\top\big[\boldsymbol{\Sigma}_{\bar{\psi}} - \boldsymbol{\Sigma}_{\psi\bar{\psi}}^\top\big(2\mathbf{I} - \boldsymbol{\Sigma}_{\psi}^{-1}(\mathbf{I}-\mathbf{G}^{-1})\boldsymbol{\Sigma}_{\psi}\big)(\mathbf{I}-\mathbf{G}^{-1})\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\big]\bar{\mathbf{w}} \\ &= Cg(\tilde{\varepsilon})\gamma + \bar{\mathbf{w}}^\top\big[\boldsymbol{\Sigma}_{\bar{\psi}} - \boldsymbol{\Sigma}_{\psi\bar{\psi}}^\top(\mathbf{I}-\mathbf{G}^{-2})\big(\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}}\big)\big]\bar{\mathbf{w}} \\ &= Cg(\tilde{\varepsilon})\gamma + \mathcal{W}(\tilde{\varepsilon},\tilde{\lambda}) = C.\end{aligned} \tag{A93}$$

7. Alternative formulation in terms of center-variation covariance

We consider the data model where $\boldsymbol{\psi} = \bar{\boldsymbol{\psi}} + \boldsymbol{\delta}$ and hence

$$\begin{aligned}\boldsymbol{\Sigma}_{\psi} &= \boldsymbol{\Sigma}_{\bar{\psi}} + \boldsymbol{\Sigma}_{\delta} + \boldsymbol{\Sigma}_{\bar{\psi}\delta} + \boldsymbol{\Sigma}_{\bar{\psi}\delta}^\top \\ \boldsymbol{\Sigma}_{\psi\boldsymbol{\delta}} &= \boldsymbol{\Sigma}_{\delta} + \boldsymbol{\Sigma}_{\bar{\psi}\delta} \\ \boldsymbol{\Sigma}_{\psi\bar{\psi}} &= \boldsymbol{\Sigma}_{\bar{\psi}} + \boldsymbol{\Sigma}_{\bar{\psi}\delta} = \boldsymbol{\Sigma}_{\psi} - \boldsymbol{\Sigma}_{\psi\boldsymbol{\delta}}.\end{aligned} \tag{A94}$$

We also note that

$$\boldsymbol{\Sigma}_{\psi\bar{\psi}}^\top\boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\bar{\psi}} = \big(\mathbf{I} - \boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\boldsymbol{\delta}}\big)^\top\boldsymbol{\Sigma}_{\psi}\big(\mathbf{I} - \boldsymbol{\Sigma}_{\psi}^{-1}\boldsymbol{\Sigma}_{\psi\boldsymbol{\delta}}\big). \tag{A95}$$

Replacing this back in generalization error

$$\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda}) = \bar{\mathbf{W}}^\top \mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}\bar{\mathbf{W}} + \bar{\mathbf{w}}^\top \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}} - \bar{\mathbf{W}}^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\bar{\mathbf{W}},$$
$$\bar{\mathbf{W}} = \big(\mathbf{I} - \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\boldsymbol{\delta}}\big)\bar{\mathbf{w}}. \tag{A96}$$

8. Derivation by Loureiro *et al.* [17]

Here, we show that our results can also be obtained by using the generic formula derived in Ref. [17]. For our loss function $\ell(x) = \frac{1}{2}\max(0,|x|-\varepsilon)^2$, we denote the proximal operator as

$$h_y(x) = \arg\min_z \left\{ \ell(z-y) + \frac{1}{2V}(x-z)^2 \right\}. \tag{A97}$$

Solving the optimization problem as before, we explicitly get

$$h_y(x) = \begin{cases} x, & |x-y| \leqslant \varepsilon \\ \frac{x+Vy}{1+V} + \operatorname{sign}(x-y)\frac{V\varepsilon}{1+V}, & |x-y| > \varepsilon \end{cases}, \quad h_y'(x) = \begin{cases} 1, & |x-y| \leqslant \varepsilon \\ \frac{1}{1+V}, & |x-y| > \varepsilon \end{cases}$$

for $h_y(x)$ and its first derivative.

Let $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$ denote the covariance matrix of d-dimensional student features and p-dimensional teacher features, respectively, and $\boldsymbol{\Phi} \in \mathbb{R}^{p\times d}$ their cross-covariance. $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ denotes the teacher weights. We define the ratios $\alpha \equiv n/d$ and $\eta \equiv p/d$ where n is the number of training samples. The self-consistent equations reported by Loureiro *et al.* [17] [see Eq. (2.8)] are

$$V = \operatorname{tr}\boldsymbol{\Omega}(\lambda\mathbf{I} + \hat{V}\boldsymbol{\Omega})^{-1}, \qquad \hat{V} = \tfrac{\alpha}{V}\mathbb{E}_{s,h}[1 - F'(s,h)],$$
$$m = \tfrac{\hat{m}}{\sqrt{\eta}}\operatorname{tr}(\boldsymbol{\Phi}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Phi})(\lambda\mathbf{I} + \hat{V}\boldsymbol{\Omega})^{-1}, \qquad \hat{m} = \tfrac{1}{\sqrt{\rho\eta}}\tfrac{\alpha}{V}\mathbb{E}_{s,h}\big[sF(s,h) - \tfrac{m}{\sqrt{\rho}}F'(s,h)\big],$$
$$q = \operatorname{tr}\boldsymbol{\Omega}(\lambda\mathbf{I} + \hat{V}\boldsymbol{\Omega})^{-2}(\hat{m}^2\boldsymbol{\Phi}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Phi} + \hat{q}\boldsymbol{\Omega}), \quad \hat{q} = \tfrac{\alpha}{V^2}\mathbb{E}_{s,h}\big[\big(\tfrac{m}{\sqrt{\rho}}s + \sqrt{q - \tfrac{m^2}{\rho}}h - F(s,h)\big)^2\big],$$

where we defined the random functions $F(s,h) = h_y(x)$ and $F'(s,h) = h_y'(x)$ with

$$x \equiv \frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h, \quad y \equiv \sqrt{\rho}s, \quad s,h \sim \mathcal{N}(0,1). \tag{A98}$$

For $\hat{V}$, the argument for the expectation value is independent of s,h and can be expressed as an average over the random variable $z \equiv x - y \sim \mathcal{N}(0, \rho + q - 2m)$. The result is

$$\hat{V} = \frac{\alpha}{1+V}\langle\Theta(|z|-\varepsilon)\rangle_z = \frac{\alpha}{1+V}f(\tilde{\varepsilon}), \quad \tilde{\varepsilon} = \frac{\varepsilon}{\sqrt{C}}, \quad C \equiv \rho + q - 2m, \tag{A99}$$

where f was defined in Eq. (A67), and $\tilde{\varepsilon}$ coincides with the definition of effective tube size.

For $\hat{m}$, let us first observe that $\mathbb{E}_{s,h}[sF(s,h)] = \mathbb{E}_{s,h}[\frac{d}{ds}F(s,h)]$ using integration by parts. Then, we get

$$\hat{m} = \frac{1}{\sqrt{\eta}}\frac{\alpha}{1+V}\langle\Theta(|z|-\varepsilon)\rangle_z = \frac{\hat{V}}{\sqrt{\eta}}. \tag{A100}$$

For $\hat{q}$, using the definition of function g in Eq. (A67), we get

$$\hat{q} = \frac{\alpha}{(1+V)^2}\big\langle(|z|-\varepsilon)^2\Theta(|z|-\varepsilon)\big\rangle_z = \frac{\alpha}{(1+V)^2}f(\tilde{\varepsilon})g(\tilde{\varepsilon})C. \tag{A101}$$

Before moving on, we define the effective load as $\tilde{\alpha} = \alpha f(\tilde{\varepsilon})$ and effective regularization as $\tilde{\lambda} \equiv \lambda(1+V)$. Then, the self-consistent equation for V becomes equivalent to

$$\tilde{\lambda} = \lambda + \operatorname{tr}\boldsymbol{\Omega}\mathbf{G}^{-1}, \quad \mathbf{G} = \mathbf{I} + \frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Omega}. \tag{A102}$$

In terms of these quantities, self-consistent equations become

$$\hat{m} = \frac{\lambda}{\sqrt{\eta}}\frac{\tilde{\alpha}}{\tilde{\lambda}}, \quad \hat{q} = C\lambda^2\frac{\tilde{\alpha}}{\tilde{\lambda}^2}g(\tilde{\varepsilon}), \quad m = \frac{1}{\eta}\frac{\tilde{\alpha}}{\tilde{\lambda}}\operatorname{tr}\mathbf{G}^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Phi}),$$
$$q = \frac{1}{\eta}\frac{\tilde{\alpha}^2}{\tilde{\lambda}^2}\operatorname{tr}\boldsymbol{\Omega}\mathbf{G}^{-2}(\boldsymbol{\Phi}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Phi}) + Cg(\tilde{\varepsilon})\frac{\tilde{\alpha}}{\tilde{\lambda}^2}\operatorname{tr}\boldsymbol{\Omega}^2\mathbf{G}^{-2}. \tag{A103}$$

Next, we define

$$\gamma \equiv \frac{\tilde{\alpha}}{\tilde{\lambda}^2}\mathrm{tr}\boldsymbol{\Omega}^2\mathbf{G}^{-2}, \quad \mathbf{w}_* \equiv \frac{1}{\sqrt{\eta}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\theta}_0. \tag{A104}$$

Using the identity $\frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Omega}\mathbf{G}^{-1} = \mathbf{I} - \mathbf{G}^{-1}$, we simplify m and q to

$$\begin{aligned} m &= \mathrm{tr}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top - \mathrm{tr}\mathbf{G}^{-1}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top, \\ q &= \mathrm{tr}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top - 2\mathrm{tr}\mathbf{G}^{-1}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top + \mathrm{tr}\mathbf{G}^{-1}\boldsymbol{\Omega}\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*^\top + Cg(\tilde{\varepsilon})\gamma. \end{aligned} \tag{A105}$$

Noting that $\rho = \mathrm{tr}\boldsymbol{\Psi}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top$, we can finally evaluate $C = \rho + q - 2m$ and recover our result

$$\begin{aligned} C &= Cg(\tilde{\varepsilon})\gamma + (\mathrm{tr}\boldsymbol{\Psi}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top - \mathrm{tr}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top) + \mathrm{tr}\mathbf{G}^{-1}\boldsymbol{\Omega}\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*^\top \\ &= \frac{1}{1 - g(\tilde{\varepsilon})\gamma}[E_\infty + \mathrm{tr}\mathbf{G}^{-1}\boldsymbol{\Omega}\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*^\top], \end{aligned} \tag{A106}$$

where $E_\infty \equiv \mathrm{tr}\boldsymbol{\Psi}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top - \mathrm{tr}\boldsymbol{\Omega}\mathbf{w}_*\mathbf{w}_*^\top$.

APPENDIX B: OPTIMAL LEARNING RATES

In this Appendix, we compute the optimal learning rates both for ridge regression by solving for $\partial_\lambda E_g = 0$, and for ε-SVR by solving $\partial_\varepsilon E_g = 0$ using the analytical expression we derived. Since E_g is defined through two self-consistent equations, its derivatives also have to be solved self-consistently. Hence, to compute optimal hyperparameters λ and ε, we need to compute the variations of $\tilde{\lambda}$ and $\tilde{\varepsilon}$.

1. Computing variations

Again we start with the two coupled self-consistent equations in the form of

$$\begin{aligned} \tilde{\lambda} &= \mathcal{F}(\tilde{\lambda}, \tilde{\varepsilon}, x) \equiv \lambda + \mathrm{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \\ \tilde{\varepsilon} &= \mathcal{G}(\tilde{\lambda}, \tilde{\varepsilon}, x) \equiv \varepsilon\sqrt{1 - g(\tilde{\varepsilon})\gamma}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})^{-1/2}, \\ \mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda}) &= E_\infty + \mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \end{aligned} \tag{B1}$$

where x collectively denotes any explicit variable, and we omitted the terms with $\mathbf{J}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}$. We found earlier that the variations of $\tilde{\lambda}$ and $\tilde{\varepsilon}$ are given by

$$\begin{aligned} \delta\tilde{\lambda} &= \frac{(1 - \partial_{\tilde{\varepsilon}}\mathcal{G})\partial_x\mathcal{F} + \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_x\mathcal{G}}{(1 - \partial_{\tilde{\lambda}}\mathcal{F})(1 - \partial_{\tilde{\varepsilon}}\mathcal{G}) - \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}\delta x, \\ \delta\tilde{\varepsilon} &= \frac{(1 - \partial_{\tilde{\lambda}}\mathcal{F})\partial_x\mathcal{G} + \partial_{\tilde{\lambda}}\mathcal{G}\partial_x\mathcal{F}}{(1 - \partial_{\tilde{\lambda}}\mathcal{F})(1 - \partial_{\tilde{\varepsilon}}\mathcal{G}) - \partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}\delta x, \end{aligned} \tag{B2}$$

and that the derivatives $\partial_{\tilde{\lambda}}\mathcal{F}$ and $\partial_{\tilde{\varepsilon}}\mathcal{F}$ are given by

$$\partial_{\tilde{\lambda}}\mathcal{F} = \gamma, \quad \partial_{\tilde{\varepsilon}}\mathcal{F} = -\frac{f'}{f}\tilde{\lambda}\gamma. \tag{B3}$$

Next, we compute the derivative $\partial_{\tilde{\varepsilon}}\mathcal{G}$

$$\begin{aligned} \partial_{\tilde{\varepsilon}}\log\mathcal{G} &= -\frac{\partial_{\tilde{\varepsilon}}(g(\tilde{\varepsilon})\gamma)}{2(1 - g(\tilde{\varepsilon})\gamma)} - \frac{1}{2}\frac{\partial_{\tilde{\varepsilon}}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})}{\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})} = -\frac{1}{2(1 - g(\tilde{\varepsilon})\gamma)}\left[\partial_{\tilde{\varepsilon}}(g(\tilde{\varepsilon})\gamma) + \frac{\partial_{\tilde{\varepsilon}}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})}{C}\right] \\ &= -\frac{1}{2(1 - g(\tilde{\varepsilon})\gamma)}\left[\gamma\frac{\partial_{\tilde{\varepsilon}}(f(\tilde{\varepsilon})g(\tilde{\varepsilon}))}{f(\tilde{\varepsilon})} + f(\tilde{\varepsilon})g(\tilde{\varepsilon})\partial_{\tilde{\varepsilon}}\frac{\gamma}{f(\tilde{\varepsilon})} + \frac{\partial_{\tilde{\varepsilon}}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda})}{C}\right]. \end{aligned} \tag{B4}$$

Here,

$$\begin{aligned} &\frac{\partial_{\tilde{\varepsilon}}(f(\tilde{\varepsilon})g(\tilde{\varepsilon}))}{f(\tilde{\varepsilon})} = -\frac{2(1 - g(\tilde{\varepsilon}))}{\tilde{\varepsilon}}, \quad \partial_{\tilde{\varepsilon}}\frac{\gamma}{f(\tilde{\varepsilon})} = -\frac{2\gamma}{\tilde{\varepsilon}}\frac{f'\tilde{\varepsilon}}{f^2}\frac{\frac{\tilde{\alpha}}{\tilde{\lambda}}\mathrm{Tr}\mathbf{M}^3}{\mathrm{Tr}\mathbf{M}^2}, \\ &\partial_{\tilde{\varepsilon}}\mathcal{W}(\tilde{\varepsilon}, \tilde{\lambda}) = -2\mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top\mathbf{G}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}(\partial_{\tilde{\varepsilon}}\mathbf{G})\mathbf{G}^{-1}, \quad \partial_{\tilde{\varepsilon}}\mathbf{G} = \frac{f'}{f}\frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \end{aligned} \tag{B5}$$

where we defined $\mathbf{M} \equiv \mathbf{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}$ to shorten the notation. Inserting these back,

$$\begin{aligned}
\tilde{\varepsilon}\partial_{\tilde{\varepsilon}}\log\mathcal{G} &= -\frac{\tilde{\varepsilon}}{2(1-g(\tilde{\varepsilon})\gamma)}\left[-\frac{2\gamma(1-g(\tilde{\varepsilon}))}{\tilde{\varepsilon}} - \frac{2\gamma g(\tilde{\varepsilon})}{\tilde{\varepsilon}}\frac{f'\tilde{\varepsilon}}{f}\frac{\frac{\tilde{\alpha}}{\tilde{\lambda}}\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} - \frac{2}{\tilde{\varepsilon}}\frac{f'\tilde{\varepsilon}}{f}\frac{\frac{\tilde{\alpha}}{\tilde{\lambda}}\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{C}\right] \\
&= \frac{1}{(1-g(\tilde{\varepsilon})\gamma)}\left[\gamma(1-g(\tilde{\varepsilon})) + \frac{f'\tilde{\varepsilon}}{f}\frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\gamma g(\tilde{\varepsilon})\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{C}\right)\right] \\
&= \frac{\gamma(1-g(\tilde{\varepsilon}))}{(1-g(\tilde{\varepsilon})\gamma)} + \frac{f'\tilde{\varepsilon}}{f}\frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{\gamma g(\tilde{\varepsilon})}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{(1-g(\tilde{\varepsilon})\gamma)C}\right) \\
&= 1 - \frac{1-\gamma}{(1-g(\tilde{\varepsilon})\gamma)} + \frac{f'\tilde{\varepsilon}}{f}\frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{\gamma g(\tilde{\varepsilon})}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right).
\end{aligned}$$

Since $\partial_{\tilde{\varepsilon}}\log\mathcal{G} = \frac{\partial_{\tilde{\varepsilon}}\mathcal{G}}{\mathcal{G}} = \frac{\partial_{\tilde{\varepsilon}}\mathcal{G}}{\tilde{\varepsilon}}$, we get

$$\boxed{\partial_{\tilde{\varepsilon}}\mathcal{G} = \frac{\gamma(1-g(\tilde{\varepsilon}))}{(1-g(\tilde{\varepsilon})\gamma)} + \frac{f'\tilde{\varepsilon}}{f}\frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{\gamma g(\tilde{\varepsilon})}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right)}. \tag{B6}$$

Next, we compute the derivative $\partial_{\tilde{\lambda}}\mathcal{G}$

$$\partial_{\tilde{\lambda}}\log\mathcal{G} = -\frac{\partial_{\tilde{\lambda}}(g(\tilde{\varepsilon})\gamma)}{2(1-g(\tilde{\varepsilon})\gamma)} - \frac{1}{2}\frac{\partial_{\tilde{\lambda}}\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})} = -\frac{1}{2(1-g(\tilde{\varepsilon})\gamma)}\left[g(\tilde{\varepsilon})\partial_{\tilde{\lambda}}\gamma + \frac{\partial_{\tilde{\lambda}}\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}{C}\right]. \tag{B7}$$

Here, we get

$$\partial_{\tilde{\lambda}}\gamma = -\frac{2\gamma}{\tilde{\lambda}}\left(1 - \frac{\tilde{\alpha}}{\tilde{\lambda}}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2}\right), \quad \partial_{\tilde{\lambda}}\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda}) = -2\mathrm{tr}\mathbf{w}_*\mathbf{w}_*{}^\top\mathbf{G}^{-1}\mathbf{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}(\partial_{\tilde{\lambda}}\mathbf{G})\mathbf{G}^{-1}, \quad \partial_{\tilde{\lambda}}\mathbf{G} = -\frac{\tilde{\alpha}}{\tilde{\lambda}^2}\mathbf{\Sigma}_{\boldsymbol{\psi}}.$$

Inserting these

$$\begin{aligned}
\tilde{\lambda}\partial_{\tilde{\lambda}}\log\mathcal{G} &= -\frac{\tilde{\lambda}}{2(1-g(\tilde{\varepsilon})\gamma)}\left[-\frac{2g(\tilde{\varepsilon})\gamma}{\tilde{\lambda}}\left(1 - \frac{\tilde{\alpha}}{\tilde{\lambda}}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2}\right) + \frac{2\tilde{\alpha}}{\tilde{\lambda}^2}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{C}\right] \\
&= -\frac{1}{(1-g(\tilde{\varepsilon})\gamma)}\left[-g(\tilde{\varepsilon})\gamma + \frac{\tilde{\alpha}}{\tilde{\lambda}}\left(g(\tilde{\varepsilon})\gamma\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{C}\right)\right] \\
&= \frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)} - \frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right).
\end{aligned}$$

Hence, we have

$$\boxed{\partial_{\tilde{\lambda}}\mathcal{G} = \frac{\tilde{\varepsilon}}{\tilde{\lambda}}\left(\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)} - \frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right)\right)}. \tag{B8}$$

Let us focus on the following term:

$$\begin{aligned}
\frac{\tilde{\alpha}}{\tilde{\lambda}}\left(\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^3}{\mathrm{tr}\mathbf{M}^2} + \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right) &= \frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^2(\mathbf{I}-\mathbf{G}^{-1})}{\mathrm{tr}\mathbf{M}^2} + \frac{\tilde{\alpha}}{\tilde{\lambda}}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})} \\
&= \frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)} - \left(\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\mathrm{tr}\mathbf{M}^2} - \frac{\tilde{\alpha}}{\tilde{\lambda}}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathcal{W}(\tilde{\varepsilon},\tilde{\lambda})}\right) \\
&= \frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)} - \frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}),
\end{aligned} \tag{B9}$$

where we defined

$$\begin{aligned}
\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}) &= g(\tilde{\varepsilon})\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\mathrm{tr}\mathbf{M}^2} - \frac{\tilde{\lambda}}{C}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathrm{tr}\mathbf{M}^2} \\
&= \frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\mathrm{tr}\mathbf{M}^2}\left(g(\tilde{\varepsilon}) - \frac{\tilde{\lambda}}{C}\frac{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*{}^\top}{\mathrm{tr}\mathbf{M}^2\mathbf{G}^{-1}}\right).
\end{aligned} \tag{B10}$$

This conveniently simplifies $\partial_{\tilde{\lambda}}\mathcal{G}$ to

$$\boxed{\partial_{\tilde{\lambda}}\mathcal{G} = \frac{\tilde{\varepsilon}}{\tilde{\lambda}}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}. \tag{B11}$$

Similarly, $\partial_{\tilde{\varepsilon}}\mathcal{G}$ becomes

$$\begin{aligned}\partial_{\tilde{\varepsilon}}\mathcal{G} &= \frac{\gamma(1-g(\tilde{\varepsilon}))}{(1-g(\tilde{\varepsilon})\gamma)} + \frac{f'\tilde{\varepsilon}}{f}\frac{g(\tilde{\varepsilon})\gamma}{(1-g(\tilde{\varepsilon})\gamma)} \\ &\quad - \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}) \\ &= \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\left(\frac{f}{f'\tilde{\varepsilon}}(1-g(\tilde{\varepsilon})) + g(\tilde{\varepsilon}) - \mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})\right).\end{aligned} \tag{B12}$$

Using the identity

$$g = 1 + \tilde{\varepsilon}^2 + \frac{f'\tilde{\varepsilon}}{f}, \tag{B13}$$

this simplifies $\partial_{\tilde{\varepsilon}}\mathcal{G}$ to

$$\boxed{\partial_{\tilde{\varepsilon}}\mathcal{G} = \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}[h(\tilde{\varepsilon}) - \mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})]}, \tag{B14}$$

where we defined the function

$$h(\tilde{\varepsilon}) \equiv \tilde{\varepsilon}\left(\tilde{\varepsilon} + \frac{f'}{f} - \frac{f}{f'}\right) = \frac{(g(\tilde{\varepsilon})-1)^2 - g(\tilde{\varepsilon})\tilde{\varepsilon}^2}{g(\tilde{\varepsilon}) - 1 - \tilde{\varepsilon}^2}. \tag{B15}$$

We obtain the denominator in Eq. (B2) as

$$\begin{aligned}&1 - \partial_{\tilde{\varepsilon}}\mathcal{G} - \frac{\partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}{1-\partial_{\tilde{\lambda}}\mathcal{F}} \\ &\quad = 1 - \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\left[h(\tilde{\varepsilon}) - \frac{\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}{1-\gamma}\right].\end{aligned} \tag{B16}$$

Finally, we get the variation of $\tilde{\varepsilon}$ as

$$\delta\tilde{\varepsilon} = \frac{\partial_x\mathcal{G}}{1-\partial_{\tilde{\varepsilon}}\mathcal{G} - \frac{\partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}{1-\partial_{\tilde{\lambda}}\mathcal{F}}}\delta x + \frac{\frac{\partial_{\tilde{\lambda}}\mathcal{G}}{1-\partial_{\tilde{\lambda}}\mathcal{F}}\partial_x\mathcal{F}}{1-\partial_{\tilde{\varepsilon}}\mathcal{G} - \frac{\partial_{\tilde{\varepsilon}}\mathcal{F}\partial_{\tilde{\lambda}}\mathcal{G}}{1-\partial_{\tilde{\lambda}}\mathcal{F}}}\delta x. \tag{B17}$$

2. Optimal ridge paremeter

To compute the optimal ridge parameter, we need to compute the derivative of generalization error C with respect to λ when the tube size is zero. Since

$$\partial_\lambda C = -2C\partial_\lambda \log\tilde{\varepsilon}, \tag{B18}$$

it suffices to compute $\partial_\lambda\tilde{\varepsilon}$ using Eq. (B17)

$$\begin{aligned}\partial_\lambda \log\tilde{\varepsilon} &= \frac{1}{1 - \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\left[h(\tilde{\varepsilon}) - \frac{\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}{1-\gamma}\right]}\frac{1}{\tilde{\lambda}} \\ &\quad \times \frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}) \\ &= \frac{1}{\tilde{\lambda}}\frac{\gamma}{(1-\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}),\end{aligned} \tag{B19}$$

where we used the fact that $\varepsilon = 0$ in the last equality. Hence, the condition for optimal ridge becomes

$$\partial_\lambda C = -\frac{2C}{\tilde{\lambda}}\frac{\gamma}{(1-\gamma)}\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda}) = 0, \tag{B20}$$

which implies another self-consistent equation for effective ridge parameter

$$\tilde{\lambda}_{\text{opt}} = C\frac{\text{tr}\mathbf{M}^2\mathbf{G}^{-1}}{\text{tr}\mathbf{M}^2\mathbf{G}^{-1}\mathbf{w}_*\mathbf{w}_*^\top}, \tag{B21}$$

replacing the original equation $\tilde{\lambda} = \lambda + \text{tr}\mathbf{M}$.

3. Optimal tube size

For optimal SVR, we set $\lambda = 0$ and consider the derivative of generalization error with respect to ε:

$$\partial_\varepsilon C = \frac{2C}{\varepsilon} - 2C\partial_\varepsilon \log\tilde{\varepsilon}, \tag{B22}$$

where $\partial_\varepsilon \log\tilde{\varepsilon}$ is computed using Eq. (B17)

$$\partial_\varepsilon \log\tilde{\varepsilon} = \frac{1}{\varepsilon}\frac{1}{1 - \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\left[h(\tilde{\varepsilon}) - \frac{\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}{1-\gamma}\right]}, \tag{B23}$$

yielding

$$\begin{aligned}\partial_\varepsilon C &= -2\sqrt{C}\frac{f'}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\frac{\left[h(\tilde{\varepsilon}) - \frac{\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}{1-\gamma}\right]}{1 - \frac{f'\tilde{\varepsilon}}{f}\frac{\gamma}{(1-g(\tilde{\varepsilon})\gamma)}\left[h(\tilde{\varepsilon}) - \frac{\mathcal{A}(\tilde{\varepsilon},\tilde{\lambda})}{1-\gamma}\right]} \\ &= 0.\end{aligned} \tag{B24}$$

This implies the self-consistency condition for effective tube size

$$h(\tilde{\varepsilon}_{\text{opt}}) = \frac{\mathcal{A}(\tilde{\varepsilon}_{\text{opt}},\tilde{\lambda})}{1-\gamma}, \tag{B25}$$

complemented with the self-consistent equation $\tilde{\lambda} = \text{tr}\mathbf{M}$.

4. Optimal parameters in Fig. 4

The optimal values of λ and ε change as a function of α, hence each point in Fig. 4 has a different optimal ridge parameter or tube size. Here, we show how the optimal parameters depend on the number of samples.

APPENDIX C: TOY DATA MODEL

To analyze our result, we consider a toy data model where the center and noise model are given by

$$\begin{aligned}&\boldsymbol{\psi} = \bar{\boldsymbol{\psi}} + \sigma\boldsymbol{\delta}, \\ &\bar{\boldsymbol{\psi}} \sim \mathcal{N}(0,\mathbf{I}), \quad \boldsymbol{\delta} \sim \mathcal{N}(0,\boldsymbol{\Sigma}_\delta), \\ &\boldsymbol{\Sigma}_\delta = (1-\beta)\mathbf{P} + \beta(\mathbf{I}-\mathbf{P}) = (1-\beta)\mathbf{I} + (2\beta-1)(\mathbf{I}-\mathbf{P}),\end{aligned} \tag{C1}$$

where $\mathbf{P} = \mathbf{I} - \frac{1}{N}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top$ is the projection matrix to the subspace orthogonal to the coding direction. Here, we assumed that the target weights are normalized such that

$$\text{Tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top = \text{Tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} = \frac{1}{1+\beta\sigma^2}\text{Tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = N. \tag{C2}$$

In this model, the noise features $\boldsymbol{\delta}$ are independent of centers $\bar{\boldsymbol{\psi}}$, but their covariance is allowed to correlate with the coding

direction $\bar{\mathbf{w}}$. In this case, we get the following covariance matrices:

$$\begin{aligned}\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} &= \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}} = \mathbf{I},\\ \boldsymbol{\Sigma}_{\boldsymbol{\psi}} &= \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} + \sigma^2\boldsymbol{\Sigma}_{\delta} = (1+\sigma^2(1-\beta))\mathbf{I} + \sigma^2(2\beta-1)\frac{1}{N}\bar{\mathbf{w}}\bar{\mathbf{w}}^{\top}.\end{aligned} \tag{C3}$$

Cleaning the last expression, we get $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}$ and its inverse as

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\psi}} &= (1+\sigma^2(1-\beta))\mathbf{P} + (1+\beta\sigma^2)(\mathbf{I}-\mathbf{P}),\\ \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1} &= \frac{1}{1+\sigma^2(1-\beta)}\mathbf{P} + \frac{1}{1+\beta\sigma^2}(\mathbf{I}-\mathbf{P}).\end{aligned} \tag{C4}$$

Hence, the error at infinity becomes

$$E_{\infty} = \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^{\top}\left(\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} - \boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\right) = \frac{\beta\sigma^2}{1+\beta\sigma^2}. \tag{C5}$$

We also compute $\mathbf{G}$ and $\mathbf{G}^{-1}$ as

$$\begin{aligned}\mathbf{G} &= \left[1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))\right]\mathbf{P} + \left[1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\beta\sigma^2)\right](\mathbf{I}-\mathbf{P}),\\ \mathbf{G}^{-1} &= \frac{1}{1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))}\mathbf{P} + \frac{1}{1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\beta\sigma^2)}(\mathbf{I}-\mathbf{P}).\end{aligned} \tag{C6}$$

Plugging these into our formula we get

$$\begin{aligned}\mathcal{F}(\tilde{\lambda},\tilde{\varepsilon}) &= \lambda + \frac{1+\sigma^2(1-\beta)}{1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))} + \frac{1}{N}\left(\frac{1+\beta\sigma^2}{1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\beta\sigma^2)} - \frac{1+\sigma^2(1-\beta)}{1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))}\right),\\ \mathcal{G}(\tilde{\lambda},\tilde{\varepsilon}) &= \frac{1}{1-g(\tilde{\varepsilon})\gamma}\left(\frac{\beta\sigma^2}{1+\beta\sigma^2} + \frac{1}{1+\beta\sigma^2}\frac{1}{\left(1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\beta\sigma^2)\right)^2}\right),\\ \gamma &= \frac{\tilde{\alpha}}{\tilde{\lambda}^2}\frac{(1+\sigma^2(1-\beta))^2}{\left(1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))\right)^2} + \frac{1}{N}\frac{\tilde{\alpha}}{\tilde{\lambda}^2}\left(\frac{(1+\beta\sigma^2)^2}{\left(1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\beta\sigma^2)\right)^2} - \frac{(1+\sigma^2(1-\beta))^2}{\left(1+\frac{\tilde{\alpha}}{\tilde{\lambda}}(1+\sigma^2(1-\beta))\right)^2}\right).\end{aligned} \tag{C7}$$

For $N \gg 1$, we can ignore $\mathcal{O}(N^{-1})$ terms. Furthermore, we redefine $\tilde{\lambda} \to \frac{\tilde{\lambda}}{1+\sigma^2(1-\beta)}$ to simplify the self-consistent equations and obtain:

$$\begin{aligned}\mathcal{F}(\tilde{\lambda},\tilde{\varepsilon}) &= \frac{\lambda}{1+\sigma^2(1-\beta)} + \frac{\tilde{\lambda}}{\tilde{\alpha}+\tilde{\lambda}},\\ \mathcal{G}(\tilde{\lambda},\tilde{\varepsilon}) &= \frac{1}{1-g(\tilde{\varepsilon})\gamma}\left(\frac{\beta\sigma^2}{1+\beta\sigma^2} + \frac{1}{1+\beta\sigma^2}\frac{\tilde{\lambda}^2}{\left(\tilde{\lambda}+\tilde{\alpha}\frac{1+\sigma^2\beta}{1+\sigma^2(1-\beta)}\right)^2}\right),\\ \gamma &= \frac{\tilde{\alpha}}{(\tilde{\alpha}+\tilde{\lambda})^2}.\end{aligned} \tag{C8}$$

1. Ridgeless limit as exactly solvable model

The solution to the self-consistent equation for $\tilde{\lambda}$ in the limit $\lambda \to 0$ becomes

$$\tilde{\lambda} = \begin{cases}1-\tilde{\alpha}, & \tilde{\alpha} < 1\\ 0, & \tilde{\alpha} \geqslant 1\end{cases}, \tag{C9}$$

and the self-consistent equation for $\tilde{\varepsilon}$ simplifies to

$$\tilde{\varepsilon} = \begin{cases}\frac{\varepsilon}{\sqrt{E_{\infty}}}\sqrt{1-g(\tilde{\varepsilon})\tilde{\alpha}}\left(1+\frac{1-E_{\infty}}{E_{\infty}}\frac{1}{\left(1+\frac{\tilde{\alpha}}{1-\tilde{\alpha}}\frac{1+\sigma^2\beta}{1+\sigma^2(1-\beta)}\right)^2}\right)^{-1/2}, & \tilde{\alpha} < 1\\ \frac{\varepsilon}{\sqrt{E_{\infty}}}\sqrt{1-g(\tilde{\varepsilon})\tilde{\alpha}^{-1}}, & \tilde{\alpha} \geqslant 1\end{cases}. \tag{C10}$$

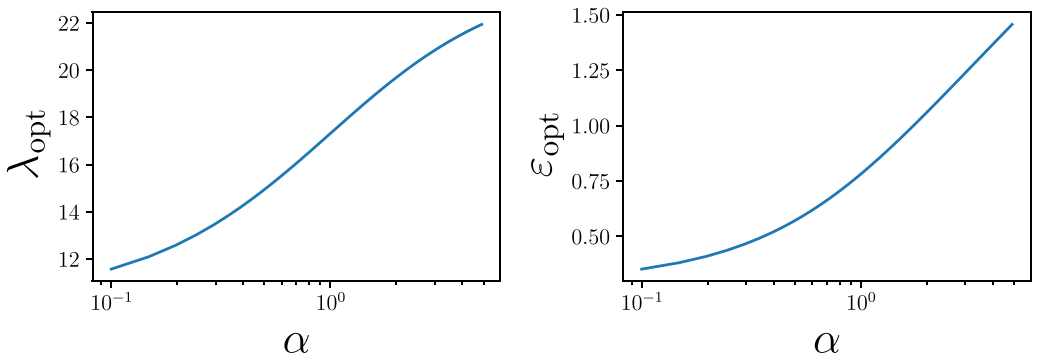

FIG. 6. Optimal hyperparameters in Fig. 4. Generally, optimal regularization increases with increasing sample sizes.

Assuming nonzero $E_\infty > 0$, this reduces to the following transcendental equation at the capacity $\tilde{\alpha} = 1$,

$$\tilde{\varepsilon} = d' \sqrt{1 - g(\tilde{\varepsilon})}, \tag{C11}$$

where we defined the discriminability $d' \equiv \frac{\varepsilon}{\sqrt{E_\infty}}$. We asymptotically extract its behavior for large and small d'. Note that the range of $\tilde{\varepsilon}$ is $[0, d')$, since the function $1 \geqslant g(x) > 0$.

For large $d' \gg 1$, $\tilde{\varepsilon}$ gets large and approaches to d'. Expanding the function g in this limit as $g(\tilde{\varepsilon}) \approx \frac{2}{\tilde{\varepsilon}^2}$, we get

$$\tilde{\varepsilon} \approx d'(1 - d'^{-2}), \quad d' \gg 1. \tag{C12}$$

Similarly, for small $d' \ll 1$, the argument approaches to zero and admits the expansion $g(\tilde{\varepsilon}) \approx 1 - \sqrt{\frac{2}{\pi}}\tilde{\varepsilon}$. Replacing this and squaring both sides, we obtain

$$\tilde{\varepsilon} \approx \sqrt{\frac{2}{\pi}} d'^2, \quad d' \ll 1. \tag{C13}$$

In Fig. 7, we confirm the validity of this approximation against its numerical solution.

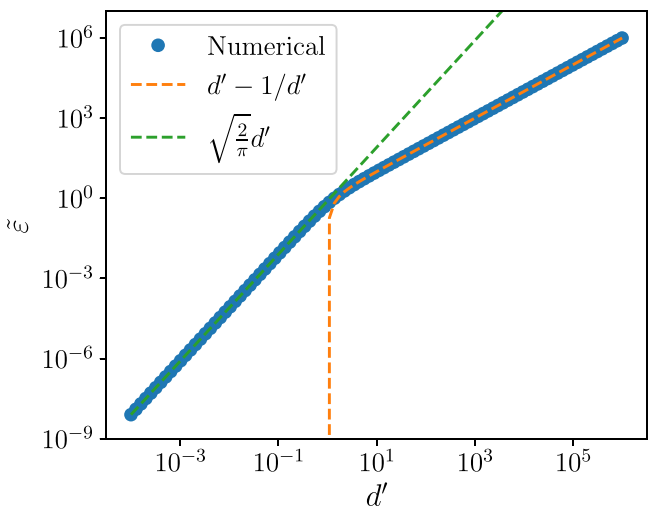

FIG. 7. Comparison of the numerical solution to Eq. (C11) to its asymptotic behavior.

APPENDIX D: SPECTRAL DECOMPOSITION

Previously, we had formulated our theory in terms of the covariance matrices of representations where we defined the functions

$$\mathcal{F}(\tilde{\lambda}, \tilde{\varepsilon}) = \mathrm{tr}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-1}, \quad \mathbf{G} = \mathbf{I} + \frac{\tilde{\alpha}}{\tilde{\lambda}}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \gamma = \partial_{\tilde{\lambda}}\mathcal{F}(\tilde{\lambda}, \tilde{\varepsilon}),$$
$$\mathcal{G}(\tilde{\lambda}, \tilde{\varepsilon}) = \frac{E_\infty + \mathrm{tr}\mathbf{w}_*\mathbf{w}_*^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{G}^{-2}}{1 - g(\tilde{\varepsilon})\gamma},$$
$$E_\infty = \mathrm{tr}\bar{\mathbf{w}}\bar{\mathbf{w}}^\top \boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} - \mathrm{tr}\mathbf{w}_*\mathbf{w}_*^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}, \quad \mathbf{w}_* = \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}}\bar{\mathbf{w}}, \tag{D1}$$

where $\tilde{\lambda}$ and $\tilde{\varepsilon}$ obey the coupled self-consistent equations given in Eq. (5). Here, we cast our formulas in terms of the spectral properties of representations. Following the kernel notation in Refs. [11,13], the covariance matrices above are given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \int \boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^\top d\mu(\mathbf{x}),$$
$$\boldsymbol{\Sigma}_{\bar{\boldsymbol{\psi}}} = \int \bar{\boldsymbol{\psi}}(\mathbf{x})\bar{\boldsymbol{\psi}}(\mathbf{x})^\top d\mu(\mathbf{x}),$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}\bar{\boldsymbol{\psi}}} = \int \boldsymbol{\psi}(\mathbf{x})\bar{\boldsymbol{\psi}}(\mathbf{x})^\top d\mu(\mathbf{x}), \tag{D2}$$

where $d\mu(\mathbf{x})$ is the distribution of the inputs and we treat representations $\boldsymbol{\psi}(\mathbf{x})$ and $\bar{\boldsymbol{\psi}}(\mathbf{x})$ as feature maps. We also express the target function and predictor as

$$\bar{y}(\mathbf{x}) = \bar{\mathbf{w}} \cdot \bar{\boldsymbol{\psi}}(\mathbf{x}),$$
$$y(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\psi}(\mathbf{x}). \tag{D3}$$

Note that the asymptotic error is simply the variance of the target function minus the variance of its projection to the learnable subspace:

$$E_\infty = \|\bar{y}(\mathbf{x})\|_2^2 - \|\bar{y}_\parallel(\mathbf{x})\|_2^2 = \|\bar{y}_\perp(\mathbf{x})\|_2^2, \tag{D4}$$

which is the unexplainable variance as $\alpha \to \infty$. Note that the explainable portion of the target function can also be obtained by

$$\bar{y}_\parallel(\mathbf{x}) = \mathbf{w}_* \cdot \boldsymbol{\psi}(\mathbf{x}) = \int d\mu(\mathbf{x}')\, \bar{y}(\mathbf{x}')\, (\boldsymbol{\psi}(\mathbf{x}')^\top \boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\psi}(\mathbf{x})), \tag{D5}$$

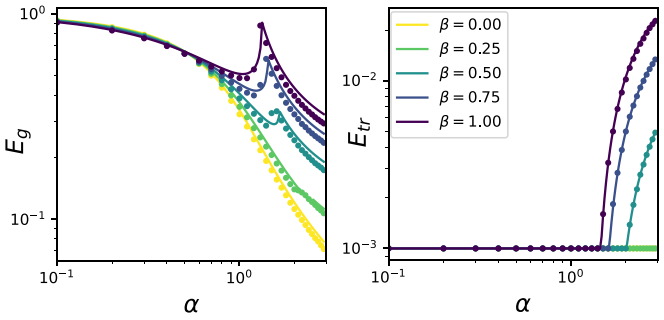

FIG. 8. Theoretical (lines) and experimental (dots) learning curves for the toy model with $\sigma = 0.5$ and $\epsilon = 0.5$. With the same noise strength, increasing the alignment between the noise and coding direction affects the generalization adversely.

where the term in the parenthesis can be considered as a projection kernel in the space of square-integrable functions. This allows us to diagonalize our formulas in the basis of $\mathbf{\Sigma}_{\boldsymbol{\psi}}$.

We want to study generalization error in regards to task-model alignment and spectral bias [13] and hence need to obtain the decomposition of the target in the space of square-integrable functions expressable by the feature map $\boldsymbol{\psi}(\mathbf{x})$ via regression. These functions belong to the function space associated with the inner-product kernel $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}')$ and can be diagonalized as

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}') = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}'),$$

$$\int \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) d\mu(\mathbf{x}) = \delta_{ij}, \tag{D6}$$

where $\{\lambda_i\}$ are nonnegative eigenvalues of the kernel and $\{\phi_i(\mathbf{x})\}$ is a set of complete eigenfunctions with respect to the measure $d\mu(\mathbf{x})$. While there are infinitely many eigenfunctions with respect to $d\mu(\mathbf{x})$, there at most $\dim(\boldsymbol{\psi})$ nonzero eigenvalues. We define the index set $\mathcal{I}^+ = \{i \mid \lambda_i > 0\}$ for positive eigenvalues and $\mathcal{I}^0 = \{i \mid \lambda_i = 0\}$ for vanishing eigenvalues. However, target function generically may have infinitely many components,

$$\bar{y}(\mathbf{x}) = \sum_i \bar{a}_i \phi_i(\mathbf{x}), \quad \bar{a}_i = \int \bar{y}(\mathbf{x}) \phi_i(\mathbf{x}) d\mu(\mathbf{x}), \tag{D7}$$

where $\{\bar{a}_i\}$ are target coefficients.

Then the equations in Eq. (D1) can be written as

$$\mathcal{F}(\tilde{\lambda}, \tilde{\varepsilon}) = \frac{\tilde{\lambda}}{N} \sum_{i \in \mathcal{I}^+} \frac{\lambda_i}{\tilde{\lambda} + \tilde{\alpha} \lambda_i}, \quad \gamma = \frac{1}{N} \sum_{i \in \mathcal{I}^+} \frac{\tilde{\alpha} \lambda_i^2}{(\tilde{\lambda} + \tilde{\alpha} \lambda_i)^2},$$

$$\mathcal{G}(\tilde{\lambda}, \tilde{\varepsilon}) = \frac{1}{1 - g(\tilde{\varepsilon})\gamma} \left(\sum_{i \in \mathcal{I}^0} \bar{a}_i^2 + \sum_{i \in \mathcal{I}^+} \frac{\tilde{\lambda}^2}{(\tilde{\lambda} + \tilde{\alpha} \lambda_i)^2} \bar{a}_i^2 \right), \tag{D8}$$

which exactly matches the equations from Ref. [13] for $\varepsilon = 0$ up to rescaling $\lambda_i \to N \lambda_i$.

This form enables us to analyze the generalization error from an error mode geometry perspective introduced in Ref. [38]. Error mode geometry relates linear predictivity to the geometric properties of representations based on the observation that the generalization error for ridge regression can be decomposed as a sum of mode errors associated with each spectral component [11,13,38]. Error modes are defined as

$$W_i(\tilde{\alpha}) \equiv \frac{\tilde{\lambda}^2}{1 - g(\tilde{\varepsilon})\gamma} \frac{\bar{a}_i^2}{(\tilde{\lambda} + \tilde{\alpha} \lambda_i)^2}, \quad E_g = \sum_i W_i(\tilde{\alpha}), \tag{D9}$$

and corresponds to the unexplained variance in the i^{th} eigendirection. Note that with increasing α, the contribution of each mode $\bar{a}_i^2$ decreases with a rate proportional to the inverse eigenvalue λ_i^{-1} associated with that mode, revealing the fact that variances associated with small eigenvalues require more samples to be reduced. Error mode geometry refers to the summarization of these error modes in terms of radius and dimensionality defined as:

$$R_{\text{em}} = \sqrt{\sum_i W_i(\tilde{\alpha})^2}, \quad D_{\text{em}} = \frac{\left(\sum_i W_i(\tilde{\alpha})\right)^2}{\sum_i W_i(\tilde{\alpha})^2} \tag{D10}$$

which explicitly relates to the prediction error $E_g = R_{\text{em}} \sqrt{D_{\text{em}}}$.

APPENDIX E: ADDITIONAL EXPERIMENTS

1. Effect of target-noise correlation

Even in the presence of noise in the representations, large correlations β between noise and the coding direction $\bar{\mathbf{w}}$ cause more uncertainty in the predictions and, therefore, affect the

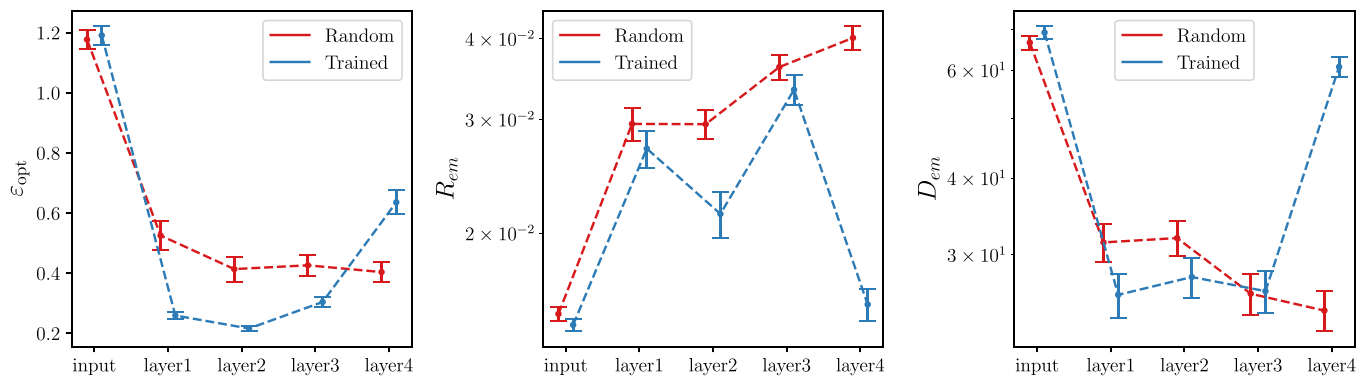

FIG. 9. Comparing optimal tube size to error mode geometry [38].

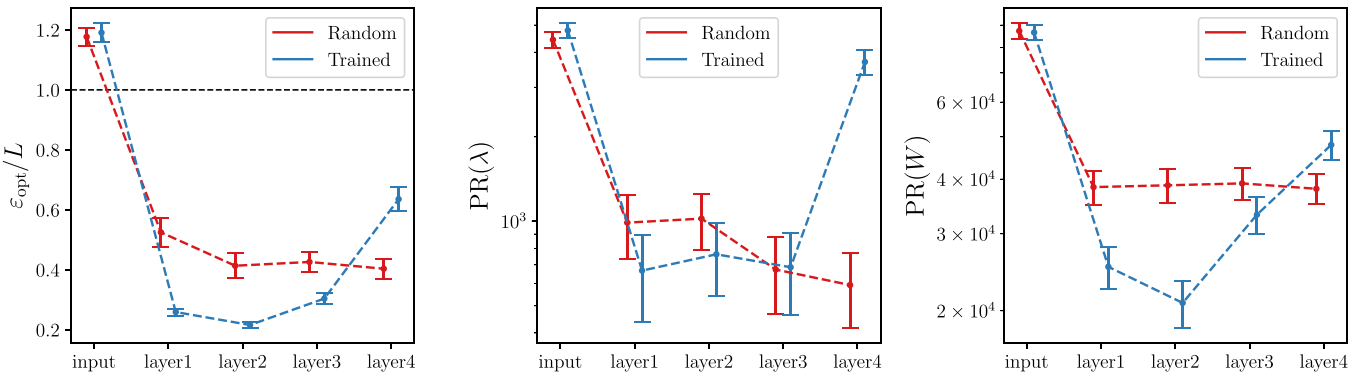

FIG. 10. Comparing the optimal tube size to participation ratios of eigenvalues and alignment coefficients.

precision. In Fig. 8, we show this effect for various β, where the noise geometry varies between completely orthogonal and parallel to the coding direction.

2. Error mode geometry of real data

Here, the metrics introduced in Ref. [38] and summarized in Appendix D are applied to the experiment presented in Fig. 5. We show how the error mode geometry decomposition compares to the optimal tube size in Fig. 9. From input representation to the first layer, dimensionality $D_{\rm em}$ drives explain lower values of $\varepsilon_{\rm opt}$ despite the increase in $R_{\rm em}$. For random networks, dimensionality across the layer hierarchy decreases, while for the trained networks dimensionality remains the same except for the peak in the last layer. This can be related to the fact that later layers of trained networks are more task-tailored and, hence, become invariant to nuisance variations. Furthermore, we compare the optimal tube size to the representational properties independent of α. We define the participation ratios for the eigenvalues and alignment coefficients as

$$\mathrm{PR}(\lambda) = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}, \quad \mathrm{PR}(W) = \frac{\left(\sum_i W_i(0)\right)^2}{\sum_i W_i(0)^2} = D_{\rm em}(0), \tag{E1}$$

which respectively characterize the dimensionality of the neural code and distribution of target components across the model's eigenbasis. In Fig. 10, we compare optimal tube size to these measures.

[1] T. A. Poggio and F. Anselmi, *Visual Cortex and Deep Networks: Learning Invariant Representations* (MIT Press, Cambridge, MA, 2016)

[2] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).

[3] I. Goodfellow, H. Lee, Q. Le, A. Saxe, and A. Ng, Adv. Neural Inf. Process. Syst. **22**, 646 (2009).

[4] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo, Nat. Neurosci. **19**, 613 (2016).

[5] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.

[6] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, Neural Comput. **12**, 1207 (2000).

[7] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Proc. Natl. Acad. Sci. USA **116**, 15849 (2019).

[8] 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, Singapore, 1987), Vol. 9.

[9] A. Engel, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).

[10] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, UK, 2009).

[11] B. Bordelon, A. Canatar, and C. Pehlevan, in *Proceedings of the International Conference on Machine Learning* (PMLR, Cambridge, MA, 2020), pp. 1024–1034.

[12] A. Jacot, B. Simsek, F. Spadaro, C. Hongler, and F. Gabriel, in *Proceedings of the International Conference on Machine Learning* (PMLR, Cambridge, MA, 2020), pp. 4631–4640.

[13] A. Canatar, B. Bordelon, and C. Pehlevan, Nat. Commun. **12**, 2914 (2021).

[14] J. B. Simon, M. Dickens, D. Karkada, and M. Deweese, Trans. Mach. Learn. Res. (2023).

[15] A. Atanasov, B. Bordelon, S. Sainathan, and C. Pehlevan, in *Proceedings of the 11th International Conference on Learning Representations* (2022).

[16] A. B. Atanasov, J. A. Zavatone-Veth, and C. Pehlevan, arXiv:2405.00592.

[17] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, Adv. Neural Inf. Process. Syst. **34**, 18137 (2021).

[18] Y. Bengio, A. Courville, and P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[19] A. Achille and S. Soatto, J. Mach. Learn. Res. **19**, 1 (2018).

[20] S. Chung and L. Abbott, Curr. Opin. Neurobiol. **70**, 137 (2021).

[21] S. Mei, T. Misiakiewicz, and A. Montanari, in *Proceedings of the Conference on Learning Theory* (PMLR, Cambridge, MA, 2021), pp. 3351–3418.

[22] A. Favero, F. Cagnetta, and M. Wyart, Adv. Neural Inf. Process. Syst. **34**, 9456 (2021).

[23] B. Elesedy, Adv. Neural Inf. Process. Syst. **34**, 17273 (2021).

[24] S. Y. Chung, D. D. Lee, and H. Sompolinsky, Phys. Rev. X **8**, 031003 (2018).

[25] A. J. Wakhloo, T. J. Sussman, and S. Y. Chung, Phys. Rev. Lett. **131**, 027301 (2023).
[26] M. Farrell, B. Bordelon, S. Trivedi, and C. Pehlevan, arXiv:2110.07472.
[27] B. Ruben and C. Pehlevan, Adv. Neural Inf. Process. Syst. **36**, 50041 (2023).
[28] B. Adlam and J. Pennington, in *Proceedings of the International Conference on Machine Learning* (PMLR, Cambridge, MA, 2020), pp. 74–84.
[29] C. Cheng and A. Montanari, Dimension free ridge regression, Ann. Stat. **52**, 2879 (2024).
[30] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).
[31] H. S. Seung and H. Sompolinsky, Proc. Natl. Acad. Sci. USA **90**, 10749 (1993).
[32] N. Brunel and J.-P. Nadal, Neural Comput. **10**, 1731 (1998).
[33] M. Belkin, S. Ma, and S. Mandal, in *Proceedings of the International Conference on Machine Learning* (PMLR, Cambridge, MA, 2018), pp. 541–549.
[34] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA, 2002).
[35] A. Maloney, D. A. Roberts, and J. Sully, arXiv:2210.16859.
[36] D. Wu and J. Xu, Adv. Neural Inf. Process. Syst. **33**, 10112 (2020).
[37] T. Hastie, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, Berlin, 2009).
[38] A. Canatar, J. Feather, A. Wakhloo, and S. Chung, Adv. Neural Inf. Process. Syst. **36**, 47052 (2023).
[39] C. Thrampoulidis, S. Oymak, and B. Hassibi, in *Proceedings of the Conference on Learning Theory* (PMLR, Cambridge, MA, 2015), pp. 1683–1709.
[40] E. Dobriban and S. Wager, Ann. Stat. **46**, 247 (2018).
[41] D. Kobak, J. Lomond, and B. Sanchez, J. Mach. Learn. Res. **21**, 1 (2020).
[42] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Ann. Stat. **50**, 949 (2022).
[43] Z. Zhang, Machine Learning: Science and Technology **6**, 025010 (2025).
[44] https://github.com/canatara/svr_code.
[45] S. Mei, T. Misiakiewicz, and A. Montanari, Appl. Comput. Harmon. Anal. **59**, 3 (2022).
[46] S. Mei and A. Montanari, Commun. Pure Appl. Math. **75**, 667 (2022).
[47] H. Hu and Y. M. Lu, arXiv:2205.06798.
[48] T. Misiakiewicz and B. Saeed, arXiv:2403.08938.