A geometric theory of few-shot learning: review and the few-shot to many-shot crossover

Shoki Kishida

Harvard School of Engineering and Applied Sciences, Applied Physics, Cambridge, MA 02138

Abstract. Sorscher *et al.* [1] introduced a geometric theory of *few-shot* concept learning—recognizing a new category from only a handful of labeled examples—that predicts the generalization error of a prototype (nearest-mean) classifier from a small set of intrinsic quantities of neural object manifolds: a manifold radius R, an effective dimensionality D, and the displacement and noise-subspace alignment between class centroids. We review the derivation given in Sec. 3 of the Supplementary Information and recast it in the language of statistical mechanics, identifying the dimensionless centroid displacement Δx_0 and its noise-subspace overlap $\|\Delta x_0 \cdot U_a\|^2$ as natural order parameters via a quenched-average argument modelled on the Sherrington–Kirkpatrick free energy. We then analyze the crossover between the few-shot ($m = 1$) and many-shot ($m \to \infty$) regimes. The crossover identifies a *dimension-limited* phase, in which prototype sampling noise dominates, and a *signal-noise-overlap-limited* phase, in which an irreducible alignment of the centroid difference with the manifold's noise subspace dominates. A single scaling variable $\xi = m/m_*$ with $m_* = N/(D^2\|\Delta x_0\|^2)$ collapses the m-shot learning curves of several (N, D) configurations onto a common envelope in the dimension-limited regime, with finite-D corrections visible in the many-shot saturation. We close by sketching how the same geometric quantities should constrain Chung–Lee–Sompolinsky manifold capacity [2] and neural scaling-law exponents [3].

1 Introduction

A central question in the statistical mechanics of learning is how the *geometry* of internal representations controls the generalization behavior of downstream classifiers. Humans can recognize a new visual category after seeing only one or two examples, and a long-standing goal of machine learning is to build classifiers with the same data efficiency [1]; the regime of m labeled examples per class with $m \lesssim 10$ is referred to as m-*shot* or *few-shot* learning. For deep networks and biological sensory cortex alike, the population activity evoked by a fixed object class, say a cat photographed under different poses, illuminations, and scales, does not concentrate at a single point in firing-rate space $\mathbb{R}^N$ but on a low-dimensional *object manifold* with center $\boldsymbol{\mu}_a$ and within-class covariance Σ_a [1, 2].

Two geometric programs have emerged. The Chung–Lee–Sompolinsky (CLS) *manifold capacity* theory [2] asks how many object manifolds can be linearly separated per feature dimension in $\mathbb{R}^N$, generalizing the celebrated Gardner calculation [4]—which used the replica method to compute the capacity $\alpha_c = P/N$ of a perceptron storing P random *point* patterns in N dimensions—from points to manifolds. Sorscher *et al.* [1] complement this by predicting the few-shot generalization error of a prototype classifier from the same family of geometric quantities, and verify the theory in macaque IT, human fMRI, and pretrained deep networks. We focus exclusively on the analytical structure here; the empirical validation in Ref. [1] on real neural and DNN data is taken as given.

This work has two goals. First, we give a self-contained review of the derivation in Sec. 3 of the Sorscher *et al.* Supplementary Information, with the statistical-mechanics structure made explicit: the dimensionless centroid difference Δx_0 and the signal-noise overlap $\|\Delta x_0 \cdot U_a\|^2$ emerge as natural order parameters of a quenched-disordered ensemble of class manifolds, in direct lineage with the Gardner programme. Second, we analyze the few-shot to many-shot crossover of the resulting expression and identify a dimension-limited regime and a signal-noise-overlap-limited regime, separated by a finite shot count m_*. A numerical experiment on synthetic manifolds illustrates the collapse onto a single scaling variable $\xi = m/m_*$ and reveals finite-D corrections to the simplified Sorscher prediction in the saturation regime.

2 Geometric theory of few-shot learning: review

2.1 Object manifolds and the prototype classifier

Object manifold. Fix a category a ("cat"). Presenting many exemplars of a under nuisance variations (pose, illumination, scale) elicits a cloud of feature vectors $\{\boldsymbol{x}_{a,\mu}\} \subset \mathbb{R}^N$. Empirically and theoretically, this cloud is well approximated by a low-dimensional manifold with center $\boldsymbol{\mu}_a = \mathbb{E}[\boldsymbol{x}_{a,\mu}]$ and within-class covariance Σ_a, whose spectrum encodes the manifold's radius and effective dimension.

Prototype (nearest-mean) classifier. Given m labeled training examples per class, form the empirical mean ("prototype") $\hat{\boldsymbol{\mu}}_a^{(m)}$ of each class and assign a test point $\boldsymbol{x}$ to the class with the closest prototype: $\hat{a}(\boldsymbol{x}) = \arg\min_a \|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_a^{(m)}\|$. This is a supervised, m-shot rule with no further training; the few-shot error of this rule is the quantity we wish to predict from $(\boldsymbol{\mu}_a, \Sigma_a)$ alone.

2.2 Setup

Following Sorscher *et al.* [1], we parameterize points on class manifold a as

$$\boldsymbol{x}_{a,\mu} = \boldsymbol{x}_0^a + \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \, \boldsymbol{u}_i^a \, s_i^{a,\mu}, \tag{1}$$

where $\boldsymbol{x}_0^a$ is the centroid, $\{R_i^a, \boldsymbol{u}_i^a\}$ are the manifold's principal radii and orthonormal axes, and the random coefficients $s^{a,\mu} \sim \mathrm{Unif}(S^{D_a^{\mathrm{tot}}-1})$ are drawn uniformly on the unit sphere of the latent space. The *manifold radius* and *effective dimen-*

sionality are

$$R_a^2 = \frac{1}{D_a^{\rm tot}} \sum_{i=1}^{D_a^{\rm tot}} (R_i^a)^2, \quad D_a = \frac{\left(\sum_i (R_i^a)^2\right)^2}{\sum_i (R_i^a)^4}, \tag{2}$$

i.e. a per-direction mean-square radius and a participation ratio of the radii. The participation ratio D_a is the *effective rank* of Σ_a: $D_a \to D_a^{\rm tot}$ for isotropic radii and $D_a \to 1$ for a ridge dominated by one direction. The empirical prototype from m training examples is

$$\hat{\boldsymbol{\mu}}_a^{(m)} = \boldsymbol{x}_0^a + \frac{1}{m} \sum_{\mu=1}^{m} \sum_i R_i^a \, \boldsymbol{u}_i^a \, s_i^{a,\mu}. \tag{3}$$

It is convenient to introduce two derived quantities now and motivate their role below. The *dimensionless centroid difference* is

$$\Delta \boldsymbol{x}_0 \equiv \frac{\boldsymbol{x}_0^a - \boldsymbol{x}_0^b}{\sqrt{R_a^2}}, \tag{4}$$

and the *rescaled noise basis* of manifold a is the $N \times D_a^{\rm tot}$ matrix

$$U_a \equiv \frac{1}{\sqrt{R_a^2}} \left[R_1^a \boldsymbol{u}_1^a, \, \ldots, \, R_{D_a^{\rm tot}}^a \boldsymbol{u}_{D_a^{\rm tot}}^a \right], \tag{5}$$

and similarly U_b for manifold b. We will see below that $\Delta \boldsymbol{x}_0$ alone controls the bias of the discriminant, while the projections $\Delta \boldsymbol{x}_0 \cdot U_a$ and $\Delta \boldsymbol{x}_0 \cdot U_b$ control its variance.

2.3 Generalization error

We classify a test point $\boldsymbol{\xi}^a$ drawn from class a by the *bare discriminant*

$$h \equiv \tfrac{1}{2} \left\| \boldsymbol{\xi}^a - \hat{\boldsymbol{\mu}}_b^{(m)} \right\|^2 - \tfrac{1}{2} \left\| \boldsymbol{\xi}^a - \hat{\boldsymbol{\mu}}_a^{(m)} \right\|^2, \tag{6}$$

assigning $\boldsymbol{\xi}^a$ to class a when $h \geq 0$. Both $\|\boldsymbol{\xi} - \hat{\boldsymbol{\mu}}\|^2$ terms carry the dimension of length squared, set by the per-direction noise scale R_a^2. We therefore rescale to the dimensionless variable

$$\tilde{h} \equiv \frac{h}{R_a^2}, \tag{7}$$

which is $\mathcal{O}(1)$ in the high-dimensional limit $N \to \infty$ at fixed $D_a^{\rm tot}/N \to 0$ and admits a central-limit description.

Substituting Eqs. (1)–(3) into $\tilde{h}$ generates seven independent contributions (Eq. SI.18 of Ref. [1]). Listed in order, with their mean and variance role: (1) a deterministic *signal* $\frac{1}{2}\|\Delta \boldsymbol{x}_0\|^2$ contributing only to the mean; (2) a *dimension-noise* term with zero mean and variance D_a^{-1}/m; (3),(4) two *bias* terms quadratic in the training coordinates, with non-zero means $\pm(R_b^2/R_a^2 - 1)/(2m)$ and variances of order $1/m^2$; (5),(6) two *signal-noise overlaps*, between the centroid displacement and the test (a) and training (b) noise bases, with zero means and variances $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$ and $\|\Delta \boldsymbol{x}_0 \cdot U_b\|^2/m$; (7) a *noise-noise overlap* $\propto \boldsymbol{u}_i^a \cdot \boldsymbol{u}_j^b$ with zero mean and variance $\|U_a^\top U_b\|_F^2/m$. The CLT collapses the sum onto an approximately Gaussian variable, so $\varepsilon_{ab} = \Pr[\tilde{h} \leq 0] = H(\mu/\sigma)$ with

$$H(x) \equiv \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt. \tag{8}$$

Combining the seven contributions, the mean and variance are

$$\mu = \tfrac{1}{2}\|\Delta \boldsymbol{x}_0\|^2 + \tfrac{1}{2m}\left(R_b^2/R_a^2 - 1\right), \tag{9}$$

$$\sigma^2 = \underbrace{\frac{D_a^{-1}}{m}}_{\text{(I) dim noise}} + \underbrace{\frac{D_a^{-1}}{2m^2}\left(1 - \frac{D_a}{m D_a^{\rm tot}}\right)}_{\text{(II) bias var. } a} + \underbrace{\frac{D_b^{-1}(R_b^2/R_a^2)^2}{2m^2}\left(1 - \frac{D_b}{m D_b^{\rm tot}}\right)}_{\text{(III) bias var. } b}$$
$$+ \underbrace{\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2}_{\text{(IV) sig.-noise overlap (test)}} + \underbrace{\frac{\|\Delta \boldsymbol{x}_0 \cdot U_b\|^2}{m}}_{\text{(V) sig.-noise overlap (train)}} + \underbrace{\frac{\|U_a^\top U_b\|_F^2}{m}}_{\text{(VI) noise-noise overlap}}. \tag{10}$$

This is Sorscher's Eq. SI.34, reorganized as a sum of six variance contributions plus a two-term mean.

Which terms can be dropped, and when. Terms (II), (III), and (VI) are subleading in the high-dimensional limit and can be discarded under explicit conditions, leaving the three retained terms (I), (IV), (V).

- *Bias variances (II)+(III).* Both scale as $1/m^2$, faster than the dimension noise (I)$\sim 1/m$. The Sorscher SI shows that for $m \geq 3$ they contribute less than a few percent to ε (Supp. Fig. 13b,c of Ref. [1]); we adopt the stricter $m \geq 10$ to be safely in this regime in what follows.
- *Noise-noise overlap (VI).* For random independent noise subspaces in $\mathbb{R}^N$, $\|U_a^\top U_b\|_F^2$ averages to $D_a D_b/N$, so its variance contribution is $D_a D_b/(mN)$. Comparing to the kept terms, dropping (VI) requires

$$\frac{D_a D_b}{m \, N \, \|\Delta \boldsymbol{x}_0\|^2} \ll 1. \tag{11}$$

  For the synthetic ensemble we use in Sec. 3 ($N = 1000$, $D_a = D_b = 10$, $\|\Delta \boldsymbol{x}_0\|^2 \geq 0.08$, $m \geq 10$), the left-hand side is ≤ 0.125 and decreases with m, so the approximation is controlled. For real DNN/IT manifolds the structure of U_a, U_b is non-random and Sorscher reports that (VI) is even smaller [1].

With (II), (III), (VI) dropped, Eq. (10) simplifies to the form we use below:

$$\sigma_{\rm red}^2 = \frac{1}{m D_a} + \|\Delta \boldsymbol{x}_0 \cdot U_a\|^2 + \frac{\|\Delta \boldsymbol{x}_0 \cdot U_b\|^2}{m}. \tag{12}$$

2.4 Order parameters and the quenched average

Equations (9)–(12) expose a striking feature: although the underlying disorder—the orthonormal axes $\{\boldsymbol{u}_i^a\}$, $\{\boldsymbol{u}_i^b\}$, the centroids, the radii—is specified by $\mathcal{O}(N D^{\rm tot})$ random parameters, the error ε_{ab} depends only on the few scalars $\|\Delta \boldsymbol{x}_0\|^2$, $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$, $\|\Delta \boldsymbol{x}_0 \cdot U_b\|^2$, D_a, D_b, and the radius ratio R_b^2/R_a^2. Performing the quenched disorder average over manifold realizations,

$$\bar{\varepsilon}_{ab} = F\Big(\overline{\|\Delta \boldsymbol{x}_0\|^2}, \, \overline{\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2}, \, \overline{\|\Delta \boldsymbol{x}_0 \cdot U_b\|^2}, \, D_a, D_b, \ldots \Big), \tag{13}$$

yields a *universal function* of these few scalars. The structure is the same as the Sherrington–Kirkpatrick spin glass, in which the quenched free energy is a function of the replica overlap $q = \overline{\langle \sigma_i^\alpha \sigma_i^\beta \rangle}$, and the mean-field Ising model, in which the free energy is a function of the magnetization $m = \overline{\langle \sigma_i \rangle}$. The mapping is direct (Table 1). The Gaussian limit $\tilde{h} \to \mathcal{N}(\mu, \sigma^2)$ plays the role of the saddle-point evaluation, and the high-dimensional limit $D^{\mathrm{tot}}/N \to 0$ provides the self-averaging that makes ε_{ab} deterministic in the order parameters. This is the sense in which Eqs. (9)–(12) extends the Gardner programme from points to manifolds.

3 Few-shot to many-shot crossover

Equation (12) splits into two physically distinct geometric contributions. The dimension term $1/(mD_a)$ is the *prototype sampling noise*: the empirical mean of m training examples fluctuates around the true centroid, with variance suppressed by the number of effective directions D_a over which the fluctuation spreads. It decreases as $1/m$ and would vanish at perfect training. The signal-noise overlap $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$ is the *irreducible test noise* in the signal direction: it quantifies how much the manifold's noise subspace overlaps with the centroid displacement and is *independent of* m—it survives even at infinite training data. The competition between these terms is the engine of the crossover analyzed below.

Matched isotropic limit. Fix two classes with matched geometry, $R_a = R_b \equiv R$, $D_a = D_b \equiv D$, and a generic centroid direction $\Delta \boldsymbol{x}_0$ in the high-dimensional limit $D/N \to 0$. The radius-mismatch bias in Eq. (9) vanishes, and the signal-noise overlap averages to its random-projection value $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2 \to \|\Delta \boldsymbol{x}_0\|^2 D/N$, yielding

$$\mu = \tfrac{1}{2}\|\Delta \boldsymbol{x}_0\|^2, \quad \sigma^2 = \frac{1}{mD} + \frac{\|\Delta \boldsymbol{x}_0\|^2 D}{N}\Big(1 + \frac{1}{m}\Big). \tag{14}$$

The two variance contributions are equal at the *crossover shot count*

$$m_* = \frac{N}{D^2 \|\Delta \boldsymbol{x}_0\|^2}. \tag{15}$$

Introducing the dimensionless shot variable $\xi \equiv m/m_*$ and the asymptotic signal-to-noise ratio $\mathrm{SNR}_\infty \equiv \|\Delta \boldsymbol{x}_0\| \sqrt{N/D}/2$, the generalization error becomes

$$\varepsilon(\xi) \approx H\left(\mathrm{SNR}_\infty \sqrt{\frac{\xi}{\xi + 1}}\right), \tag{16}$$

a one-parameter master curve in ξ modulo the $1/m$ correction in σ^2, which is subleading for moderate m.

Two regimes. Equation (16) interpolates between:

- *Dimension-limited (few-shot) phase*, $\xi \ll 1$: $\sigma^2 \approx 1/(mD)$ and $\varepsilon \approx H\big(\frac{1}{2}\|\Delta \boldsymbol{x}_0\|^2 \sqrt{mD}\big)$ depends on the manifold dimension only through m/D. Doubling D is equivalent to halving the shot count; the bottleneck is the *prototype sampling noise*.
- *Overlap-limited (many-shot) phase*, $\xi \gg 1$: $\sigma^2 \approx \|\Delta \boldsymbol{x}_0\|^2 D/N$ and $\varepsilon \to \varepsilon_\infty = H(\mathrm{SNR}_\infty)$. Further samples no longer help; the bottleneck is the *signal-noise overlap*, controlled by D/N and by the orientation of $\Delta \boldsymbol{x}_0$ relative to the manifold's noise subspace.

Numerical verification. Figure 1 confirms the collapse on synthetic sphere-normalized manifolds with $N \in \{500, 1000, 2000\}$, $D \in \{5, 10, 20, 40\}$, and two values of $\mathrm{SNR}_\infty \in \{1.0, 2.0\}$ chosen along the family $\|\Delta \boldsymbol{x}_0\|^2 = 4\,\mathrm{SNR}_\infty^2\, D/N$. Each cell uses 800 Monte-Carlo trials with independent random subspaces (U_a, U_b) per trial, and the well-separation condition (11) is satisfied at all sampled $m \geq 10$. Panel (a) shows the raw m-shot learning curves spanning three orders of magnitude in ε; panel (b) rescales the horizontal axis to $\xi = m/m_*$ and reveals that the six curves of each SNR_∞ family collapse onto a common envelope in the dimension-limited regime $\xi \lesssim 1$. The dashed master curves from Eq. (16) agree with the collapse in that regime but overestimate ε in the overlap-limited phase: the simulated error continues to drop below the predicted saturation. This systematic deviation is a *finite-D correction* not captured by the leading high-dimensional CLT: in our rank-D sphere construction, the $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$ term carries a sub-leading $1/D$ factor that sharpens the asymptotic decay relative to Sorscher's main-text prediction [1]. The qualitative picture—a dimension-limited to overlap-limited crossover at $m_* = N/(D^2 \|\Delta \boldsymbol{x}_0\|^2)$—is robust.

4 Discussion

4.1 Relation to manifold capacity

The CLS *manifold capacity* $\alpha_M(R, D)$ [2] is the maximal number of object manifolds per feature dimension that an N-dimensional perceptron can linearly separate; it reduces to the Gardner capacity [4] in the point limit $R \to 0$. The manifold capacity and the few-shot SNR of Eq. (14) are two faces of the same geometric saddle point: both are controlled by the dimensionless ratio $R\sqrt{D}$ to leading order, and what the capacity calculation calls a "subspace correlation" is precisely the signal-noise overlap $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$ entering our σ^2. A precise dictionary would map the inverse capacity $1/\alpha_M$ to the integrated learning-curve area $\int_0^\infty \varepsilon^{(m)}\, d \log m$; we leave this to future work.

4.2 Relation to neural scaling laws

The participation ratio D_a in Eq. (2) depends sharply on the eigenvalue tail of Σ_a: for a power-law spectrum $\lambda_k \propto k^{-\alpha}$ with cutoff K, $D_a \sim K$ for $\alpha < 1/2$ but $D_a \sim \mathcal{O}(1)$ for $\alpha > 1$. This dovetails with the data-pruning analysis of Ref. [3], which argued that the exponent of neural scaling laws is set by the geometry of the data distribution rather than by model capacity per se. A quantitative connection between D_a and the scaling exponent remains open.

Table 1: Mapping of the few-shot prototype-classifier problem onto standard mean-field setups in statistical mechanics. In each row the quenched-averaged free energy (or, equivalently here, the generalization error) reduces to a function of a small number of scalar order parameters. The few-shot prototype problem extends the Gardner programme from points to manifolds, with $(\|\Delta \boldsymbol{x}_0\|^2,\ \|\Delta \boldsymbol{x}_0 \cdot U_a\|^2,\ D_a)$ playing the role analogous to the magnetization m in Ising and the replica overlap q in SK.

System	Order parameter	Free energy / error
Ising (MF)	$m = \langle \sigma_i \rangle$	$f(m, T, h)$
SK spin glass	$q = \overline{\langle \sigma_i^\alpha \sigma_i^\beta \rangle}$	$f(q, T)$
Few-shot prototype	$\big(\|\Delta \boldsymbol{x}_0\|^2,\ \|\Delta \boldsymbol{x}_0 \cdot U_a\|^2,\ D_a\big)$	$F(\cdot)$ in Eq. (13)

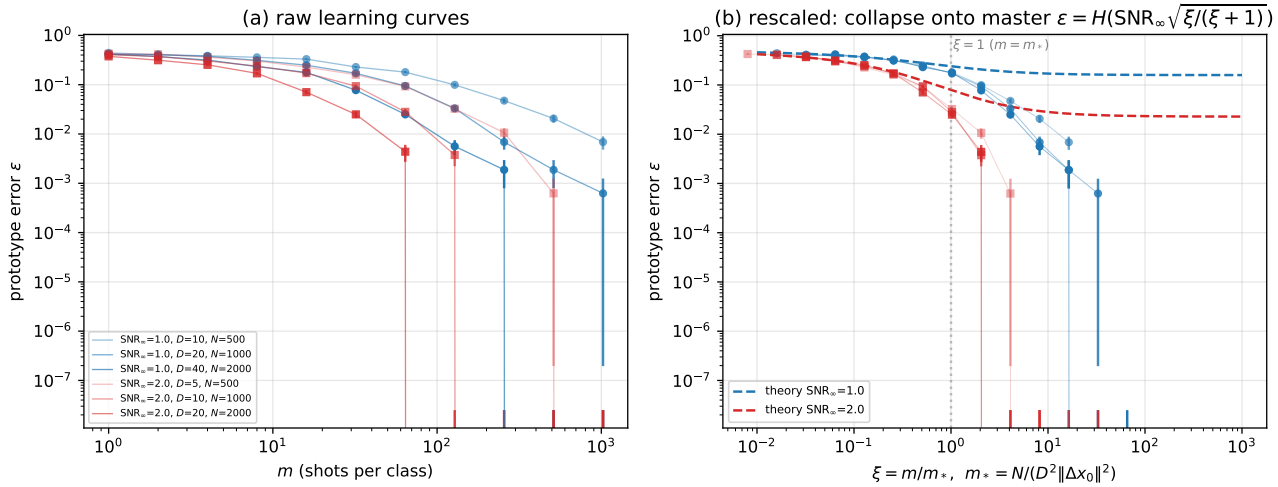

Figure 1: Few-shot to many-shot crossover for the matched isotropic prototype classifier. (a) Raw m-shot learning curves for six (N, D) configurations along two SNR_∞ families. (b) Same data plotted against $\xi = m/m_*$ with $m_* = N/(D^2\|\Delta \boldsymbol{x}_0\|^2)$ from Eq. (15). Curves within each family collapse for $\xi \lesssim 1$ (dimension-limited phase); the simulated ε continues to decay below the predicted saturation in the overlap-limited phase ($\xi \gtrsim 1$), a finite-D correction to Sorscher's main-text Eq. 1. Markers: MC estimates with $\pm 1\sigma$ error bars (800 trials per cell); dashed curves: master $H(\mathrm{SNR}_\infty \sqrt{\xi/(\xi+1)})$; dotted vertical: $\xi = 1$.

5 Conclusion

We reviewed the Sorscher *et al.* geometric theory of few-shot learning, recovered the full mean and variance of the renormalized discriminant in Eqs. (9)–(10) from a seven-term central-limit decomposition, and identified the dimensionless centroid difference $\Delta \boldsymbol{x}_0$ and its noise-subspace projections $\|\Delta \boldsymbol{x}_0 \cdot U_a\|^2$ as natural order parameters by analogy with the quenched-average construction of the Sherrington–Kirkpatrick free energy. The main analytical result is the few-shot to many-shot crossover: a single scaling variable $\xi = m/m_*$ with $m_* = N/(D^2\|\Delta \boldsymbol{x}_0\|^2)$ separates a dimension-limited phase, in which the prototype sampling noise $\sim 1/(mD)$ bottlenecks generalization, from an overlap-limited phase, in which the irreducible signal-noise overlap $\sim \|\Delta \boldsymbol{x}_0\|^2 D/N$ does. Monte-Carlo experiments on synthetic manifolds confirm the collapse in the dimension-limited regime and reveal a finite-D correction in the saturation regime, where the empirical error decays below Sorscher's simplified prediction. The crossover identifies the geometric bottleneck in each regime, offering a falsifiable prediction for synthetic and biological data.

Acknowledgments

The author thanks Mehran Kardar for the 8.334 final-project framing, and Haim Sompolinsky for ongoing discussions on representation geometry.

References

[1] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119, 2022.

[2] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

[3] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[4] Elizabeth Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, 1988.