The Role of Cortical Varibility in Supporting Few-shot Generalization: Theory and Empirical Evidence

Praveen Venkatesh[*,1,2], **Jiaqi Shang**[3], **Corbett Bennett**[1], **Sam Gale**[1], **Greggory R. Heller**[4], **Tamina K. Ramirez**[5], **Séverine Durand**[1], **Eric T. Shea-Brown**[2], **Shawn Olsen**[1], **Stefan Mihalas**[†,1]

[1]Allen Institute; [2]University of Washington; [3]Harvard University; [4]Massachusetts Institute of Technology; [5]Columbia University

[*]praveen.venkatesh@alleninstitute.org; [†]stefanm@alleninstitute.org;

Abstract

Cortical neurons exhibit a high degree of trial-to-trial variability in response to repeated presentations of the same stimulus. We examine a theory of how such variability can be *helpful* for generalizing from a small number of examples. We extract three predictions from a simplified Gaussian model of this theory: (1) to minimize generalization error, the optimal neural variability must have a covariance proportional to that of the data points within a class; (2) when considering just two classes, the magnitude of variability must shrink perpendicular to the decision boundary; and (3) the magnitude of variability must shrink in all directions with more examples to generalize from. We then provide evidence from experimental neural data in support of each of these hypotheses. We observe, in the visual cortex of mice, that variability is aligned with in-class variance; that the magnitude of variability shrinks in a task-specific direction with task engagement; and that the magnitude of variability shrinks in all directions with increased stimulus familiarity. Finally, we demonstrate that injecting noise with the appropriate correlation structure into the intermediate layers of a convolutional neural network can promote generalization over rotations of the input. Taken together, the data and simulations provide evidence consistent with the theory that cortical variability supports few-shot generalization.

1 Introduction

Cortical neurons portray extremely high variability in response to repeated stimulus presentations, even in highly standardized recordings [1, 2]. Previous studies in the literature have mainly focused on whether and how neural variability limits the encoding of information [3–5]. A few recent papers grounded in empirical evidence argue that, while variability *does* ultimately limit encoding [6–8], its correlation structure is such that a dominant fraction of variability is *not* information-limiting [9].

Incidentally, neurons *have the capacity* to be very reliable, such as neurons in the peripheral somatosensory system [10]. If most variability is harmless, and perhaps biologically evitable, one must wonder: does the brain *maintain* a certain degree of variability, and could this variability serve a computational purpose?

A different long-standing question in neuroscience has been that of how the brain is able to generalize from a small number of examples. For instance, Tenenbaum et al. [11] ask how 2-year old children learn new words like "horse" or "hairbrush" from just a few examples. More recently, it has been argued that few-shot generalization is made possible by a specific representational geometry in the high-dimensional space of neural activity [12].

We unite these two lines of research, and posit that cortical variability supports the brain's ability to generalize from a small number of examples. We believe variability performs this role by encoding

1^{st} Workshop on NeuroAI @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

invariances that are inherent in perception, such as object permanence, geometrical invariances arising from translations or rotations of 3D objects, etc. A similar idea has been floated before [13, 14], wherein variability is theorized to encode *perceptual uncertainty* by providing samples from a probabilistic representation of a stimulus. The notion that noise can be beneficial has also been long understood in the field of machine learning: for example, noise in stochastic gradient descent helps avoid local minima [15], and noise in the form of dropout makes models more robust [16].

Our contributions are threefold: (i) we mathematically formalize a theory of how neural variability can support few-shot generalization; (ii) we then find evidence from experimental neural data to support some of its predictions, and (iii) we show that injecting structured variability into a convolutional neural network imbues it with the ability to generalize over a new invariance space.

2 A Theory of Variability and Few-shot Generalization

In this section, we elucidate a theory of how variability might be helpful for few-shot generalization under a simplified Gaussian model. Suppose we are interested in learning to distinguish between two classes in $\mathbb{R}^d$, with distributions $P_0 = \mathcal{N}(\mu_0, \Sigma_{IC})$ and $P_1 = \mathcal{N}(\mu_1, \Sigma_{IC})$, where Σ_{IC} is represents an "in-class" variance common to both classes, while the means, μ_0 and μ_1 of the two distributions are themselves drawn i.i.d. from $P_\mu = \mathcal{N}(0, \Sigma_\mu)$. In the few-shot setting, let x_0^k and x_1^k be i.i.d. samples from P_0 and P_1 respectively, with samples $k \in \{1, ..., K\}$. We will rely on the fact that linear discriminant analysis (LDA) is the Bayes-optimal classifier for the binary classification problem described above [17, Sec. 17.4.1].

Now, to model *variability* in visual perception, suppose we are allowed to observe these samples x_i^k as many times as we want, so that we observe not just x_j^k, but a large number of *trials* of x_j^k with trial-to-trial variability.[1] Let this variability be characterized by a Gaussian distribution with mean centered around x_j^k and some covariance Σ. Then, we can ask: what is the optimal Σ that minimizes the generalization error of classifying between P_0 and P_1, given samples $\{x_0^k\}$ and $\{x_1^k\}$?

Proposition 1. *Let $K = 1$, and suppose we have infinitely many trials. Then, the optimal trial-to-trial variability, Σ, which minimizes the generalization error of LDA, is proportional to the in-class variance, i.e., $\Sigma^* = \alpha\Sigma_{IC}$, for some scaling constant α.*

A sketch of the proof is provided in Appendix A. Proposition 1 formalizes an intuitive argument: to minimize generalization error, the trial-to-trial variability should smear the representation of a sample along directions that would mimic *other samples* of the same class. In essence, variability should *augment* the dataset. For simplicity, we present the restricted case, where Σ_μ (the covariance corresponding to P_μ) is taken to be equal to Σ_{IC}. We leave a more complete analysis of this theorem to future work. We provide two more predictions based on this Gaussian model, along with proof sketches in the appendix:

Proposition 2. *Suppose there are only two classes, i.e., μ_0 and μ_1 are arbitrary constants, and not drawn from a distribution P_μ. Then, the optimal noise Σ is a degenerate Gaussian, with zero variance in the direction orthogonal to the true decision boundary.*

Proposition 3. *Suppose $K > 1$, and that we estimate the covariance to be used for LDA from the samples $\{x_i^k\}$. Then, the scaling factor α decreases with increasing K, falling roughly as $1/\sqrt{K}$.*

3 Experimental Evidence from Mouse Visual Cortex

We used two datasets collected by us at the Allen Institute [18], consisting of spiking neural data collected from the visual cortex of mice using neuropixels probes. In **Dataset 1**, mice were made to passively watch 10 movie clips, each repeated a total of 200 times in random order. In **Dataset 2**, mice were presented one of eight images in 250ms flashes separated by 500ms gray screens. The image *changed* after a variable number of flashes; mice had to lick to receive a reward when the image changed. Apart from a "Familiar" session with eight images that were seen during training, a "Novel" session with six new images and two familiar images was also recorded. In each recording session, after one hour of "active" task engagement, the mice were replayed the same sequence of stimuli for "passive" viewing. In both datasets, neural activity was recorded using six neuropixels

[1] Note the distinction between a *sample*, x_0^k, representing one of a small number of examples we are given for few-shot learning, and a *trial*, which gives rise to trial-to-trial variability and of which we have many.

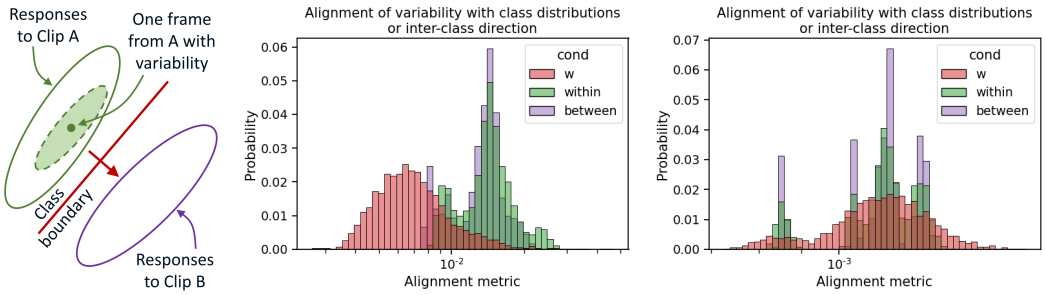

Figure 1: **(Left)** We measure the alignment between the distribution of variability of spike count responses corresponding to frames of a given movie clip, and the distribution of responses to all frames of that clip (green). We also measure the alignment between the variability and the in-class variance of frames from a *different* clip (purple), and to the between-clip direction (red). Alignments are measured using the weighted average of q-values, denoted $\bar{q}$ (see Appendix B). **(Center)** Histograms of $\bar{q}$-values, taken across all appropriate frame-movie pairs: comparing variability of frames with *in-class* variance in green; comparing variability of frames with *between-class* variance in purple; and comparing variability of frames with the between-class direction, denoted "w". The $\bar{q}$-values are *more* aligned to the in-class variance (as well as to the variance of *another* class) than to the between-class direction, confirming that variability smears representations along directions of invariance. **(Right)** Null distribution histograms of $\bar{q}$-values for randomized controls: each frame's variability was randomly rotated using a random orthogonal matrix before computing the alignment. This measures the extent to which the alignments shown in the Center figure are produced by chance. Observe that the histograms are all clustered around a $\bar{q}$-value of $\sim 10^{-3}$ in the null distribution, whereas they are clustered around a $\bar{q}$-value of $\sim 10^{-2}$ for the true histograms, indicating that the true alignment is much higher than chance.

probes targeting visual cortical regions. We consider the variability in *neural representations*, which are defined by spike counts of visual cortical neurons in short time windows, in response to movie frames in Dataset 1, or image flashes in Dataset 2. We provide evidence from these experimental data supporting each of the three theoretical results presented in the previous section:

1. In Dataset 1, we find that the distribution of trial-to-trial variability is aligned with the distribution of in-class variance, as predicted by Prop. 1.
2. In Dataset 2, we find that the magnitude of variability shrinks in a task-specific direction when the mouse is actively engaged in a discrimination task, compared to a setting where it passively observes the stimuli. This is in line with Prop. 2, which states that the optimal variability is infinitesimal in the direction orthogonal to the decision boundary.
3. In Dataset 2, we also find that the magnitude of variability is observed to shrink in all directions with increased stimulus familiarity, consistent with Prop. 3.

Metrics to Measure Variability. We quantify the extent of variability in different directions using three different metrics: (i) the noise projection, which measures the extent of variability in a given direction; (ii) the q-value, a measure of the *relative* variability in a given direction, compared to all other directions; and (iii) the signal-to-noise ratio (SNR), which measures the (linear) distinguishability between two distributions (refer Appendix B for details, incl. a diagram in Fig. 7).

Evidence from repeated stimulus presentations. To test whether variability has the appropriate correlation structure as described by Prop. 1, we measure the alignment of variability with in-class variance in Dataset 1 (details in Appendix C). Here, the variability is defined by the distribution of 200 different responses to individual frames from the movie stimuli; the "in-class variance" refers to the distribution of all responses across all frames within the same movie.[2] We computed the weighted averages of q-values (denoted $\bar{q}$-values; see Appendix B for details) measuring the alignment between the distributions of trial-to-trial variability for a particular frame and the overall distributions from all frames and trials for an entire movie clip. We observed that the histogram of $\bar{q}$-values (over all frame-movie pairs) is greater than what would be produced by chance, indicating that the variability

[2]The "classes" in this dataset are dissimilar from conventional ML, because the mice have not been *trained* to distinguish between them. Rather, we *interpret* them as classes, since mice might need to distinguish between them for *ethological* reasons.

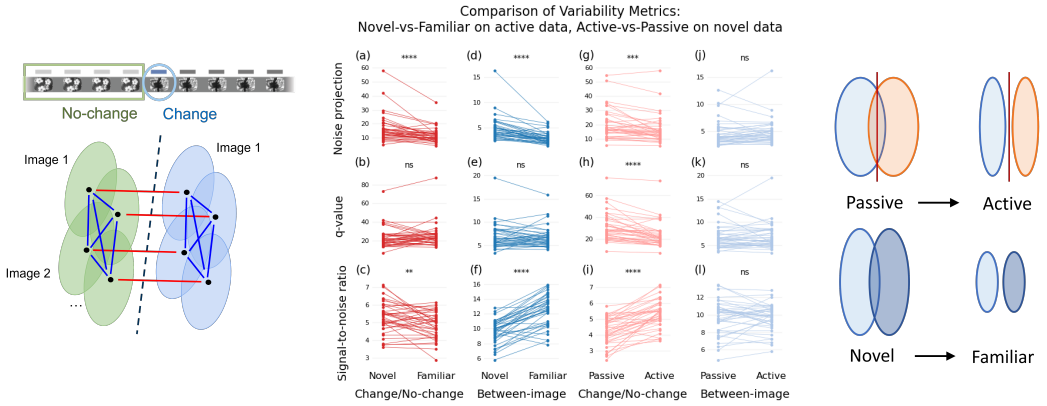

Figure 2: **(Left)** A depiction of the responses to different images under change and no-change conditions, in a 2D projection of high-dimensional neural activity space. The ellipses correspond to the distributions of responses due to trial-to-trial variability; the directions along which variability is measured are shown in red, for the change/no-change direction, and blue, for the between-image direction. **(a-l)** Different variability metrics (on rows) computed for the change/no-change direction (a-c,g-i) and the between-image direction (d-f,j-l). Lines correspond to different mice. Comparing novel vs. familiar conditions (a-f), variability shrinks in all directions, with increased distinguishability of images. Comparing active vs. passive conditions (g-l), variability shrinks in only the (task-specific) change/no-change direction, increasing distinguishability along the task axis. **(Right)** Summary cartoons depicting how variability changes with task engagement (top) and familiarity (bottom). The red line depicts the change/no-change class boundary.

of individual frames is aligned with the variance across frames within a clip (i.e., aligned with what we interpret as in-class variance, consistent with Prop. 1; see Fig. 1). Interestingly, the variability of individual frames was also aligned with the distribution of frames from a *different* movie clip (i.e., a different class), possibly indicating that variability smears the representations of different classes in the same way in this dataset, capturing invariances common to all classes. As an additional control, we observed that the variability was *less* aligned with the directions along which different movies were *separated*.

Evidence from task engagement. To examine the effect of task engagement on the geometry of neural variability, we measure the variability of neural activity for passive and active conditions in Dataset 2. We consider two directions: (i) that between the trial-averages under change and non-change conditions for each image (red lines in Fig. 2Left); (ii) the direction between every pair of images *within* change and non-change conditions (blue lines in Fig. 2Left). The former is a task-relevant direction, while the latter is ethologically relevant.

We compute the noise projection, the q-value and the SNR, for each image (in the change/no-change direction) and for every pair of images (in the between-image direction), and average over all images or image pairs (details in Appendix D; results in Fig. 2g–l). We find a small but statistically significant reduction in variability in the change/no-change direction, going from passive to active task engagement (14.0% reduction in median across mice, one-sided signed-rank test; Fig. 2g). Moreover, this decrease is highly specific, as indicated by a decrease in the q-value (24.2% decrease in median across mice; Fig. 2h). The decreased noise is also accompanied by increased distinguishability between non-change and change stimuli (26.3% increase; Fig. 2i). In contrast, we are unable to detect statistically significant changes in any metric in the direction between images (median decreases of -11.4%, 3.3% and 2.8% in Figs. 2j–l respectively). This is consistent with Prop. 2: i.e., when focusing on two classes (in the active context), variability shrinks in a *specific* direction orthogonal to the class boundary (as depicted in Fig. 2Right,top).

Evidence from stimulus familiarity. To examine how variability changes with familiarity, we compute the same metrics as before, in the same two directions, between novel and familiar stimuli in Dataset 2 (details in Appendix D). Going from 'Novel' to 'Familiar', we observe a reduction in the noise projection in both the change/no-change direction (24.9% decrease in median across mice) *and* in the between-image direction (34.8% decrease in median across mice; see Fig. 2a,d). However, we are unable to detect a statistically significant decrease in the q-value (-9.9% median decrease in the

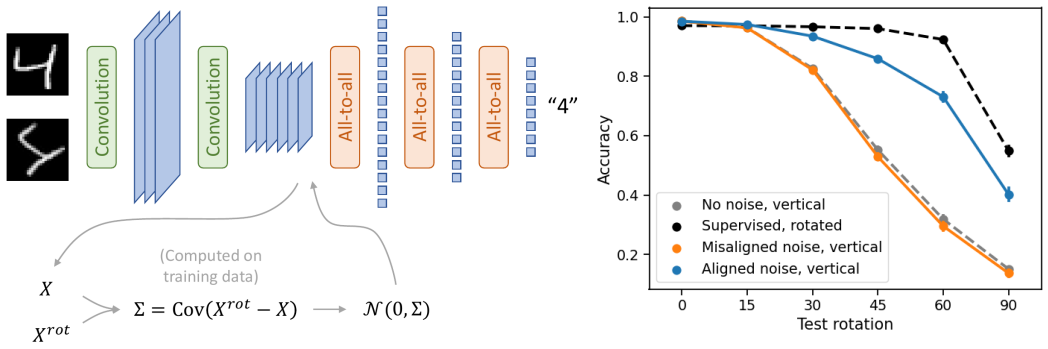

Figure 3: **(Left)** A depiction of the convolutional neural network trained on the MNIST dataset, along with the noise injection process. A single covariance matrix is computed for all digits, which is computed from the *change* in an intermediate representation induced by rotations. Noise with this covariance is added back to the network and subsequent layers are re-trained. **(Right)** The result of retraining the network with noise injection, using the correctly aligned covariance matrix (shown in blue), relative to a control using a misaligned diagonal covariance matrix (shown in orange). Upper and lower baselines, from supervised training with rotations and no noise injection, respectively, are shown as black and grey dashed lines.

change/no-change direction, -2.5% decrease in the between-image direction; one-sided signed-rank test; Fig. 2a–f), suggesting an *overall* shrinkage in neural variability, consistent with Prop. 3. This shrinkage is accompanied by increased SNR in the direction between images (27.9% increase in median across mice; Fig. 2f), suggesting that familiarity increases distinguishability of ethologically relevant stimuli (depicted in Fig. 2Right).

4 Empirical Evidence from Artificial Neural Networks

Prop. 2 suggests that the optimal trial-to-trial variability smears the internal representations along directions of invariance for the classification problem at hand. We next ask if we can imbue artificial neural networks with certain invariances by injecting variability into their internal representations.

As a simple test, we consider the task of identifying handwritten digits on the MNIST dataset [19], and train a small convolutional neural network modeled after LeNet-5 [20] to solve it. Although convolutional neural networks are invariant to translations by design, they are not inherently invariant to rotations of the inputs. We infer the covariance structure that rotations would impart to representation at an intermediate layer of the CNN in an unsupervised manner (see Fig. 3Left; details in Appendix E). We then test whether injecting variability with this correlation structure while retraining subsequent layers makes the CNN invariant to rotations of its inputs.

We test the network trained with noise injection on input digits rotated at various angles. To serve as baselines, we train a second CNN on only vertical digits (for a lower bound), and a third CNN whose subsequent layers are retrained with rotated digits in a supervised manner (for an upper bound). We also compare our result with a fourth CNN that is trained with noise injection, but with a misaligned covariance matrix that has all off-diagonal elements zeroed-out, to test the importance of the *structure* of correlations.

Our results show that injecting noise with the right covariance structure imbues the CNN with a significant amount of rotational invariance (see Fig. 3Right). Using noise with the right alignment is closer to supervised training, than to the lower baseline trained only on vertical digits. The control that uses misaligned noise is identical to, or worse than, the lower baseline.

5 Discussion

Our work theorizes that variability supports few-shot generalization, with three predictions that are supported by empirical evidence from neural data and artificial neural networks. Although it may be surprising that noise can be *beneficial*, we believe its usefulness lies in destroying *unnecessary* information by smearing representations along directions that the brain needs to generalize over.

Importantly, the source of cortical variability is irrelevant to our theory of its computational usefulness. Neural spiking may be purely deterministic (sans thermodynamic noise), with cortical variability being entirely a product of variability in external stimuli and efference copies of the animal's own behavior [5]. Alternatively, variability may be a product of the inherent unreliability of biological building blocks. In either case, the question is only whether the brain harnesses variability to its advantage (as we theorize), or if it merely copes with it.

Although we do not formally address the mechanisms that underlie the geometry of neural variability, we conjecture that simple Hebbian plasticity is sufficient to develop correlations with the right structure. For example, associations of the different views of a 3D object can be naturally learned over the course of development, and inhered in synaptic connections through Hebbian plasticity. Thus, during inference, when one view of an object is perceived, variability smears the representation of this object along the strongest synaptic connections, organically suggesting what that object might look like from other angles. We leave a more formal investigation of mechanisms to future work.

Questions about how the brain computes in the presence of variability, as well as about how the brain is able to generalize from a small number of examples have long been discussed in neuroscientific literature. Our paper offers a theoretical formalism connecting these concepts, and provides evidence from experimental neural data supporting this theory. We also demonstrate the practical feasibility of this idea through simulations on artificial neural networks. The theory presented here could be further validated through targeted experiments that explore the relationship between the structure of variability and generalization in diverse settings.

References

[1] Saskia EJ de Vries, Jerome A Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23 (1):138–151, 2020.

[2] Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

[3] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

[4] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

[5] Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015.

[6] Ramon Bartolo, Richard C Saunders, Andrew R Mitz, and Bruno B Averbeck. Information-limiting correlations in large neural populations. *Journal of Neuroscience*, 40(8):1668–1678, 2020.

[7] Omer Hazon, Victor H Minces, David P Tomàs, Surya Ganguli, Mark J Schnitzer, and Pablo E Jercog. Noise correlations in neural ensemble activity limit the accuracy of hippocampal spatial representations. *Nature communications*, 13(1):4276, 2022.

[8] Xaq Pitkow, Sheng Liu, Dora E Angelaki, Gregory C DeAngelis, and Alexandre Pouget. How can single sensory neurons predict behavior? *Neuron*, 87(2):411–423, 2015.

[9] Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105, 2020.

[10] Yi Dong, Stefan Mihalas, Sung Soo Kim, Takashi Yoshioka, Sliman Bensmaia, and Ernst Niebur. A simple model of mechanotransduction in primate glabrous skin. *Journal of neurophysiology*, 109(5):1350–1359, 2013.

[11] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[12] 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.

[13] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[14] Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

[15] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pages 2698–2707. PMLR, 2018.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

[17] Robert E Kass, Uri T Eden, Emery N Brown, et al. *Analysis of neural data*, volume 491. Springer, 2014.

[18] Allen Institute. Visual behavior neuropixels dataset overview, 2022. URL https://portal.brain-map.org/explore/circuits/visual-behavior-neuropixels.

[19] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] George K Karagiannidis and Athanasios S Lioumpas. An improved approximation for the gaussian q-function. *IEEE Communications Letters*, 11(8):644–646, 2007.

[22] Abhranil Das and Wilson S Geisler. Methods to integrate multinormals and compute classification measures. *arXiv e-prints*, pages arXiv–2012, 2020.

A Proof Sketches of Theoretical Results

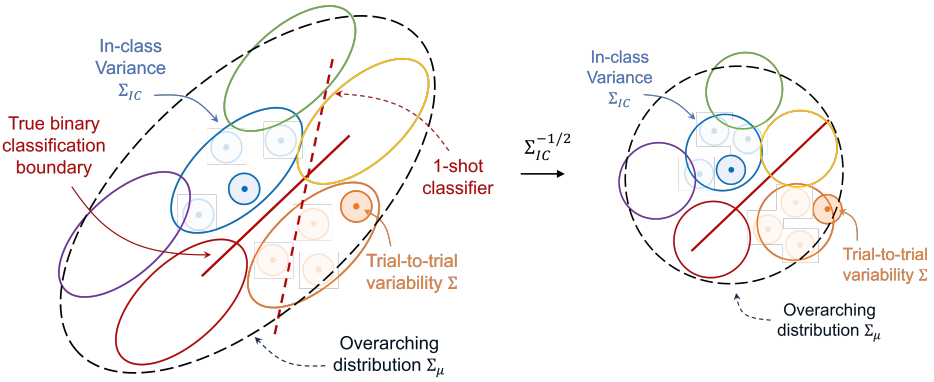

Figure 4: Diagram to provide an intuition for the proof of Prop. 1. **(Left)** A depiction of the problem setup for optimizing the trial-to-trial variability Σ to minimize generalization error. The ellipses of different colors refer to different possible centers μ_i, which are drawn from an overarching distribution with covariance Σ_μ, depicted by the large dashed ellipse in black. The two filled ellipses in blue and orange refer to the two classes that are actually drawn, with centers μ_0 and μ_1. The shapes of the blue and orange ellipses are described by the in-class variance Σ_{IC}. The two samples x_0 and x_1 drawn from each of these classes are depicted as points, with their trial-to-trial variability Σ shown by the smaller circles. The true decision boundary between the two selected classes is given by the solid red line, while the 1-shot estimated classifier is shown in dashed red. **(Right)** The same picture shown after transforming the space by $\Sigma_{IC}^{-1/2}$, for the case where Σ_μ is assumed to be identity post-transformation.

In this section, we provide sketches of the Propositions presented in Section 2.

Proof Sketch for Proposition 1. We assume $K = 1$, so we have one sample x_0 and x_1 from each class P_0 and P_1. These samples have variability Σ, which we are trying to optimize to minimize generalization error. We first write out the closed-form expression of the LDA classifier, and state the generalization error of the classifier.

The LDA classifier is a function $\phi : \mathbb{R}^d \to \{0, 1\}$, mapping any new data point $x \in \mathbb{R}^d$ to the class label.

$$\phi(x) = \begin{cases} 0, & w^\mathsf{T} x \leq c \\ 1, & w^\mathsf{T} x > c \end{cases} \tag{1}$$

where

$$w = \Sigma^{-1}(x_1 - x_0) \tag{2}$$

$$c = w^\mathsf{T}(x_0 + x_1)/2 \tag{3}$$

The generalization error for the LDA classifier is given by the Gaussian tail probability, which we write using the Q-function [21]:

$$Q(z) \coloneqq \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz \tag{4}$$

The generalization error is then given by the tail probability of each Gaussian distribution P_0 and P_1, averaged under a prior of $\text{Ber}(1/2)$:

$$GE_\phi(\Sigma) = \mathbb{E}_{\mu_0,\mu_1}\mathbb{E}_{x_0,x_1}\left[\frac{1}{2}Q\left(\frac{c - w^\mathsf{T}\mu_0}{\sqrt{w^\mathsf{T}\Sigma_{IC}w}}\right) + \frac{1}{2}Q\left(\frac{w^\mathsf{T}\mu_1 - c}{\sqrt{w^\mathsf{T}\Sigma_{IC}w}}\right)\right], \tag{5}$$

where Σ is implicit in w and c; the expectation of x_0 and x_1 is taken over the independent product of P_0 and P_1; and the expectation of μ_0 and μ_1 is taken over $P_\mu \times P_\mu$. We then wish to find the optimal Σ that minimizes the generalization error:

$$\Sigma^* = \min_\Sigma GE_\phi(\Sigma) \tag{6}$$

Without loss of generality, we can take $\Sigma_{IC} = I$, by simply transforming all vectors in the space by $\Sigma_{IC}^{-1/2}$. This simplifies the generalization error:

$$GE_\phi(\Sigma) = \mathbb{E}_{\mu_0,\mu_1}\mathbb{E}_{x_0,x_1}\left[\frac{1}{2}Q\left(\frac{w^\mathsf{T}(\frac{x_0+x_1}{2} - \mu_0)}{\|w\|}\right) + \frac{1}{2}Q\left(\frac{w^\mathsf{T}(\mu_1 - \frac{x_0+x_1}{2})}{\|w\|}\right)\right] \tag{7}$$

$$= \mathbb{E}_{\mu_0,\mu_1}\mathbb{E}_{x_0,x_1}\left[\frac{1}{2}Q\left(\hat{w}^\mathsf{T}\left(\frac{x_0 + x_1}{2} - \mu_0\right)\right) + \frac{1}{2}Q\left(\hat{w}^\mathsf{T}\left(\mu_1 - \frac{x_0 + x_1}{2}\right)\right)\right], \tag{8}$$

where $\hat{w} = w/\|w\|$. In this setup, it should now be clear that Σ^* is determined only up to a scaling factor α, since Σ only appears within $\hat{w}$, which is a unit vector. Thus, Σ only controls the direction of the classifier $\hat{w}$.

Now, if $\Sigma_\mu = I$, then there are no special directions in this space, since both inner and outer expectations are carried out over standard, isotropic, Gaussian distributions (see Fig. 4). Therefore, by symmetry, we must have $\Sigma = I$ so as to not impose any inductive biases, which will only worsen the generalization.

If $\Sigma = I$ in the space transformed by $\Sigma_{IC}^{-1/2}$, then $\Sigma = \Sigma_{IC}$ (modulo a scaling factor) in the original space, prior to transformation. Thus, the optimal trial-to-trial variability that minimizes generalization error is proportional to the in-class variance. □

Proof Sketch for Proposition 2. Continuing from the intuition provided in the proof above, if there are only two classes, then there is a clear usefulness for having an inductive bias, as depicted in Fig. 5. Intuitively, the generalization error is minimized when the estimated 1-shot classifier is parallel to the true decision boundary.

This can be enforced by a Σ that has some finite extent in all directions, except for the single axis orthogonal to the decision plane, along which it has an eigenvalue of zero. In other words, generalization error is minimized in the binary classification setting by a trial-to-trial variability that is degenerate, with an infinitesimal thickness that is parallel to the decision plane. □

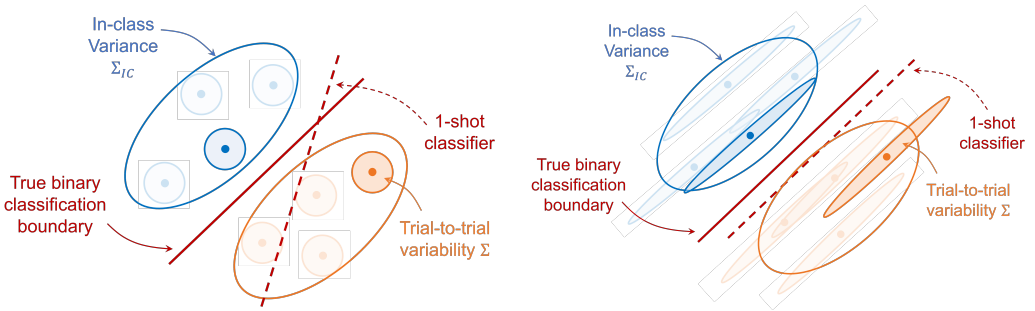

Figure 5: Diagrams to provide an intuition for the proof sketch of Prop. 2. **(Left)** Isotropic variability Σ can give rise to 1-shot classifiers that generalize poorly, depending on which sample is drawn from each class. **(Right)** If the variability is highly anisotropic, with infinitesimal variance in a direction orthogonal to the true decision plane, estimated 1-shot classifiers are guaranteed to be parallel to the true decision plane, thus minimizing generalization error.

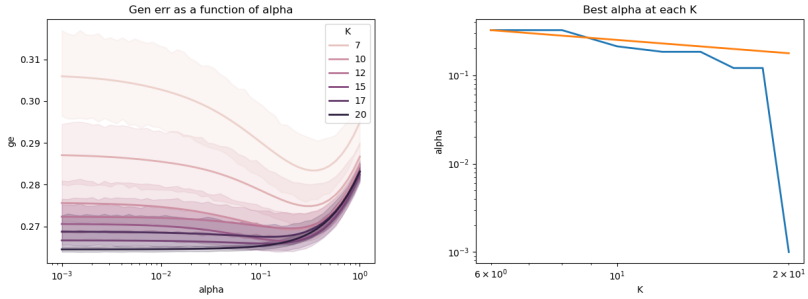

Figure 6: Figures for the numerical proof sketch of Prop. 3. **(Left)** Numerically computed generalization error as a function of α, for different values of the number of samples K. **(Right)** The optimal value of α that minimizes generalization error at each value of K (shown in blue). The line in orange shows what an $\alpha \propto 1/\sqrt{K}$ would look like. α decreases with increasing K, and the estimated trend closely matches the $1/\sqrt{K}$ line until estimation errors appear to cause the trend to diverge.

Proof Sketch for Proposition 3. Beginning with the expression for the generalization error of LDA given in Equation (8), we numerically evaluate how the optimal scaling α depends on the number of samples K.

We take K values between 3 and 10, and a range of α values between 10^-3 and 1. We consider a 2-dimensional case, setting $\mu_0 = [-1, 0]^\mathsf{T}$ and $\mu_1 = [1, 0]^\mathsf{T}$, with $\Sigma_{IC} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$. We simulate the generalization problem 500 times for each value of K and α, with results as shown in Fig. 6. We find that the optimal value of α decreases with increasing K, indicating that the optimal variability shrinks as more samples are available to generalize from. This scaling approximately matches the $1/\sqrt{K}$ trend, before estimation errors cause a divergence in the trend lines.

This is intuitive, since as more samples become available, the distribution of the samples provides a sufficient description of the shape of in-class variance for classification, and the inductive bias introduced by Σ becomes less important. □

B Mathematical Definitions of the Variability Metrics

We measure the geometry of variability by considering the distribution of neural activity over multiple trials. We assume that the neural activity of a single neuron is given by a real number, e.g., its spike count within some temporal window. Then, in the datasets, the overall distribution of neural activity is an empirical distribution given by a set of vectors $x_i \in \mathbb{R}^N$, $i \in \{1, ..., T\}$, where N is the number of neurons and T is the number of trials.

Let $\{x_i\}$ represent the N-dimensional neural activity vector across T trials, and $\hat{w}$ be a unit vector in some direction of interest. Let $\bar{x} \in \mathbb{R}^N$ be the trial-averaged neural activity. Then, two of the three variability metrics—the noise projection and the q-value—are given by:

$$\text{NoiseProj}(\{x_i\}, \hat{w}) = \frac{1}{T}\sum_{i=1}^{T}\left(\hat{w}^\mathsf{T}(x_i - \bar{x})\right)^2 = \text{Var}\big(\hat{w}^\mathsf{T}(x_i - \bar{x})\big) \tag{9}$$

$$\text{q}(\{x_i\}, \hat{w}) = \frac{\text{Var}(\hat{w}^\mathsf{T}(x_i - \bar{x}))}{\frac{1}{N}\sum_{j=1}^{N}\text{Var}(x_{ij} - \bar{x}_j)}. \tag{10}$$

In most cases in the results section, we refer to the noise projection and q-value between two *distributions*. In these cases, the noise projection and q-value are more precisely defined as:

$$\text{NoiseProj}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \text{Var}\big(\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\} \cup \{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}\big) \tag{11}$$

$$\text{q}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \frac{\text{Var}\big(\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\} \cup \{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}\big)}{\frac{1}{N}\sum_{j=1}^{N}\text{Var}\big(\{x_{ij}^{(1)} - \bar{x}_j^{(1)}\} \cup \{x_{ij}^{(2)} - \bar{x}_j^{(2)}\}\big)}, \tag{12}$$

where $\hat{v} = (\bar{x}^{(1)} - \bar{x}^{(2)})/\|\bar{x}^{(1)} - \bar{x}^{(2)}\|$ is the unit vector along the line joining the centers of the two distributions.

For **Dataset 1**, we use a slightly different version of the q-value, which does not normalize by $1/N$, and which takes a weighted average with respect to the principal component values of the first distribution. This boils down to:

$$\bar{q}(\Sigma^{(1)}, \Sigma^{(2)}) = \frac{Tr\{\Sigma^{(1)}\Sigma^{(2)}\}}{Tr\{\Sigma^{(1)}\}Tr\{\Sigma^{(2)}\}}, \tag{13}$$

where $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are respectively the N-dimensional covariance matrices corresponding to $\{x_i^{(1)}\}$ and $\{x_i^{(2)}\}$.

The signal-to-noise ratio (SNR) is a generalization of the d-prime score, which measures the distinguishability between two Gaussian distributions. For distributions with equal covariance matrices, this measure is given by:

$$\text{SNR}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \frac{\|\bar{x}^{(1)} - \bar{x}^{(2)}\|}{\text{Std}\big(\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\} \cup \{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}\big)}, \tag{14}$$

When $\{x_i^{(1)}\}$ and $\{x_i^{(2)}\}$ are sampled from two distributions with unequal covariances, this metric can be generalized by computing the generalization error of a Quadratic Discriminant Analysis classifier. We estimate the variances of the two 1D distributions given by $\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\}$ and $\{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}$ separately and compute the SNR numerically [22].

A diagrammatic representation of these metrics is provided in Fig. 7.

On the Stability of our Estimates. It should be noted that the expressions provided here involve estimating the variance or standard deviation of one-dimensional quantities. We avoid computing the covariance matrix itself, since the dimensionality of our data is larger than the number of available trials. A more careful analysis of the variance of our estimates of these metrics is left to future work.

On the Skewness of Variability. In constructing our metrics, we ignore the precise shape of the distribution of neural variability, and assume that it is approximately ellipsoidal. In practice, if the variability is highly anisotropic, with different degrees of skew along different directions, our metrics would not capture these effects, since they only consider up to the second moment. This could occur, for instance, with Poisson spike counts at low rates, wherein a Poisson distribution with a rate of λ has a positive skew of $1/\sqrt{\lambda}$. A more careful analysis of the impact of such effects are beyond the scope of the current study and could be taken up in future work.

C Analysis Pipeline for Dataset 1

1. We take three consecutive movie frames (for a total of 100 ms) to be a single frame for the purposes of our analysis. The "representation" of this (combined) single frame is given by the spike counts of the recorded neurons in this window.

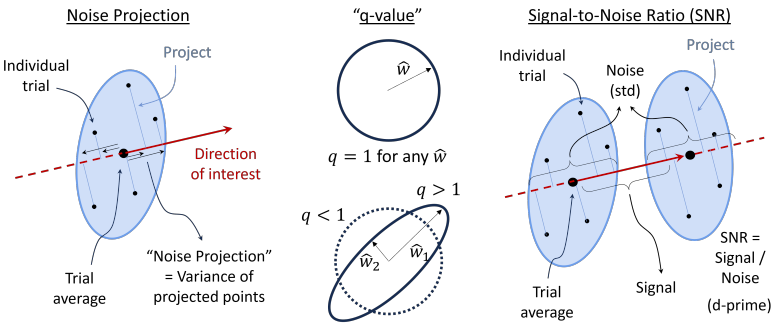

Figure 7: Depictions of metrics used to measure trial-to-trial variability. The noise projection measures the variance of the variability, projected along a direction of interest. The q-value computes the variance in a direction of interest, and normalizes it by the average variance over all directions. The signal-to-noise ratio measures the distinguishability of two distributions, and is a generalization of the d-prime measure depicted here, to two Gaussian distributions with unequal covariance matrices.

2. Only movie clips that are longer than 2.5s are included, to ensure a sufficient number of frames within each clip, resulting in 9 clips. In addition, we exclude the first three bins from each movie clip to avoid responses that result from transitions between clips.
3. Neural data are filtered to include only time points when the animals are not running (a speed below 5 cm/s) to control for changes in firing rates due to running.
4. Different frames from a single movie clip are treated as different samples from a single class. It should be understood that the dataset is entirely passive, and that mice are not trained to classify between clips, and thus this nomenclature of "classes" is only valid to the mice in the sense that it is ethologically relevant for them to distinguish different stimuli.
5. We compute q-values, as defined in Appendix B for the 200 trials of each sample. We use the first 10 principal components of the in-class variance, defined by the covariance of the trial-averaged responses to each frame within each clip. We then take a weighted average of these 10 q-values, using the percentage of variance explained by the corresponding principal component as the weight.
6. A histogram of these averaged q-values over all samples is presented in Figure 1.

D Analysis Pipeline for Dataset 2

1. We only considered mice with both familiar and novel sessions, and which had at least 20 neurons in each of the following visual cortical regions: VISp, VISl, VISal, VISam and VISpm. We sub-selected units that had a quality of 'good', with an SNR of at least 1, and with fewer than 1 inter-spike interval violations.
2. In each session, we computed the neural activity by counting the spikes of each unit in a 50–125ms time window after stimulus onset.
3. Stimulus flashes that corresponded to a 'change' were those trials in which the image changed (i.e., was different from the image in the preceding flash) and the mouse was engaged in the task (defined by having a rolling reward rate of at least 2 rewards/min).
4. Stimulus flashes that corresponded to a 'non-change' were those flashes that occurred between 4 and 10 flashes after the start of a behavioral trial and before the image changed, which did not have an omission or follow an omission, on which the mouse did not lick, and while the mouse was engaged.
5. The three variability metrics in the change/no-change direction were computed separately for every image, between change and non-change distributions. The metrics were then averaged across all 8 images (for the familiar session) and across all 6 novel images (for the novel session; the two shared familiar images were ignored).
6. The variability metrics in the between-image direction were computed separately for every pair of images in the non-change class and for every pair in the change class. The variability metrics were then averaged across all images pairs across both classes.

7. The active-passive comparison was performed on the 6 novel images.
8. Each line in Fig. 2c-n corresponds to a different mouse. Statistical significance was assessed across mice using one-sided (paired) Wilcoxon signed-rank tests.

The selection criteria in Step 1 above yielded 39 mice with both familiar and novel sessions, with 525.15 ± 98.63 units in familiar sessions, and 423.97 ± 80.67 units in novel sessions (mean $\pm$ standard deviation). We also obtain 82.18 ± 19.57 trials of each image for the non-change condition and 23.78 ± 6.06 trials of each image for the change condition.

E Details of the Analysis on Artificial Neural Networks

1. We train a feedforward convolutional neural network with five layers on the MNIST classification task. The architecture of the CNN is as follows: (0) Starting with 28×28 MNIST digits as input; (1) 6-channel 5×5 convolutional layer with 2-pixel padding, followed by 2×2 max-pooling down to $6 \times 14 \times 14$; (2) 16-channel 5×5 convolutional layer with no padding, followed by 2×2 max-pooling down to $16 \times 5 \times 5$; (3) A fully-connected linear layer to 32 hidden units; (4) A fully-connected linear layer to 16 hidden units; (5) A fully-connected linear layer to 10 hidden units, representing a 1-hot encoded output. We use tanh activations in all but the final layer.
2. The network is trained on vertical MNIST digits with an Adam optimizer (with a learning rate of 0.001), for 10 epochs. We will refer to this network as the base network below. We do not use any form of early stopping, so that all networks we compare are trained on the same amount of data.
3. After the initial training, the lower baseline is assessed by testing the base network at different test rotation angles (see the x-axis in Fig. 3Right).
4. For the upper baseline, we retrain the last three layers of the base network on digits rotated uniformly at random between ± 60 degrees, in a supervised manner.
5. Our main result retrains the base network with noise drawn from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, added to the hidden layer activations after layer 2 (after the tanh non-linearity). A different realization of noise is used for each sample that is passed through the network.
6. The covariance structure of the noise, Σ, is computed from the training data in an unsupervised manner as follows: (i) For every training data point, we run a forward pass through the base network of vertical digits as well as the same digits, rotated uniformly at random between ± 60 degrees. (ii) We extract the hidden layer activations after layer 2 (with dimension $16 \times 5 \times 5$) for the vertical and rotated stimuli—call these X and X^{rot} respectively. (iii) The covariance matrix is then computed as $\Sigma := \mathrm{Cov}(X^{rot} - X)$, as shown in Fig. 3Left. Note that the training labels are never used in this process, and that a common covariance matrix is computed for all MNIST digits.
7. As a control, and to test the importance of the precise structure of the noise covariance, we also retrain the base network with noise injection that is sampled from the diagonal of Σ, i.e., $\Sigma \odot I$, where $\odot$ represents an element-wise product.
8. For each of the four test settings, we train (and retrain) 10 networks with different random initializations. This is shown in the errorbars of Fig. 3Right, representing one standard *deviation* in the distribution of outputs.