Chapter 15: Generalization, geometry, and causality

Natural environments present agents with a combinatorial space of problems. Learning a solution to one problem is useless in the long run unless some aspect of the solution is generalizable to other problems. The brain's ability to do this effectively is an important computational mystery. One approach to unraveling the mystery, pursued in this chapter, is to look at generalization from the perspective of causality: generalization is most effective when learning invariant predictors that plausibly capture causal relationships between variables, while disregarding spurious (non-causal) relationships. Representations that support invariant prediction have a distinctive parallel geometry that is attested in neural recordings, supporting the idea that the brain organizes its representational architecture to support causal generalization. Several mechanisms for learning invariant predictors are reviewed, with connections to dreaming and oscillatory plasticity rules.

Generalization is something so natural for our brains that it's easy to miss how remarkable it is. Consider the seemingly simple problem of recognizing a cow in an image. People have no trouble with the images shown in Figure 1, yet convolutional neural networks trained for object recognition (see Chapter 8) struggle when cows appear in unusual backgrounds like beaches. Because cows appear mainly in pastures, the networks learn to rely on information contained in the background—a spurious correlation that should be ignored. How does the brain know what to ignore?

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure 1: **Recognizing cows with different backgrounds.** Labels, generated by a convolutional neural network, are shown below each image. Reproduced from Beery et al. (2018).

The essence of the problem is causality: we understand intuitively that pastures don't cause cows to appear in images. Only cows cause the appearance of cows! Because cows often graze in pastures, the presence of a pasture makes it more likely that a cow will be there.

In other words, we could think of the pasture as a cause of the cow's presence, which in turn causes its appearance in a photo of the pasture. Critically, if the farmer arrives and leads the cow back to its stable, the pasture will still appear in the photo but the cow will not. The farmer has "intervened" on the causal structure, severing the correlation between the pasture and the cow appearance.

It's important to understand that interventions are a special form of variation, because they alter causal structure; they aren't simply random samples from the joint distribution of variables.

Figure 2 depicts the situation graphically. Here X is the pasture, Y is the cow presence, and Z is the cow's appearance in the photo (all binary variables in this case). The farmer's intervention is represented by the "do" operator (Pearl, 2009), which removes the causal influence of X on Y. This in turn breaks the spurious correlation between X and Y. In other words, X is not a cause of Z because its effect on Z is not invariant across interventions on Y (a cow will not appear in a photo if it's absent, even if the photo is taken in a pasture). In contrast, Y is a cause of Z because its effect on Z is invariant across interventions on X (a cow will appear in a photo if it's present, even if the photo is taken on a beach).

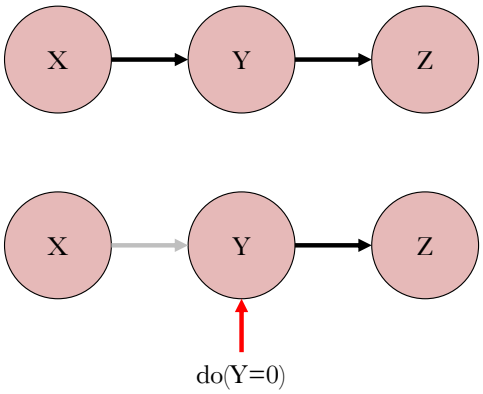

Figure 2: **A causal model.** (Top) The model represented as a directed graph, where nodes represent variables and arrows represent causal dependencies. (Bottom) Intervening on variable Y, represented by $\text{do}(Y = 0)$, sets Y to 0 and removes the arrow from X to Y.

The view of causality as *invariance under intervention* is central to understanding its role in generalization. A system that learns invariant causal mechanisms will generalize correctly to new contexts, whereas a system that learns spurious correlations will not. The goal of this chapter is to unpack this idea more systematically, and then to explore its implications for the brain. We will see how causal invariance is achieved by neural representations with a particular geometry in the high-dimensional space of population activity. This geometry is abstract in the sense that it maintains its structure across contexts, reflecting the underlying causal invariants that are unaffected by interventions on the other variables that collectively comprise the context.

1 Causality and invariance

Many different strands of thinking about causality have pivoted around some notion of invariance. They all have in common the assertion that causal relationships are "law-like" in the sense that they generalize across many contexts. Conversely, contexts are interventions that leave the causal relationships intact. Because each context is associated with a different distribution over observations, knowledge of causal relationships enables a form of "out of distribution" generalization.

For a philosophical exposition, see Woodward (2005). A plethora of terms in other fields have been used to denote closely related ideas: *robustness* (Bühlmann, 2020), *autonomy* (Haavelmo, 1944), *ignorability* (Rubin, 1978).

The critical question is how we can obtain causal knowledge from observational data. The next section describes a formal framework that (under some assumptions) guarantees invariance across the space of all contexts, provided invariance is satisfied in a sufficiently diverse set of training contexts.

1.1 From empirical to invariant risk minimization

Recall the empirical risk minimization setup from Chapter 8. We are given a dataset of M input-label pairs, $\{x_m, s_m\}_{m=1}^M$, where x_m is an input (e.g., an image) and s_m is its label (e.g., an object category or a continuous feature like size). The goal is to find a conditional distribution $q(s|x)$, a *predictor*, that minimizes the empirical risk:

$$\hat{L}(q) = \frac{1}{M} \sum_m L(q, s_m, x_m), \tag{1}$$

where $L(q, s, x)$ is a loss function.

Now consider the following regression example using continuous labels ($s \in \mathbb{R}$) and two continuous inputs ($x = [x_a, x_b]$). Suppose that the data-generating process has the following structure:

$$x_a \sim \mathcal{N}(0, \sigma_a^2) \tag{2}$$

$$x_b \sim \mathcal{N}(s, \sigma_b^2) \tag{3}$$

$$s \sim \mathcal{N}(x_a, \sigma_s^2). \tag{4}$$

Here s is an effect of x_a and a cause of x_b. Ideally, we would like our classifier to correctly identify the causal structure, relying only on x_a to predict s. Unfortunately, this will not generally be the case for empirical risk minimization. The Bayes-optimal predictor, obtained by minimized by the cross-entropy loss (see Chapter 8), is the posterior, $q(s|x) = p(s|x)$, which in this case is a Gaussian with mean $\hat{s}$:

Cross-entropy loss: $L(q, s, x) = -\log q(s|x)$.

$$\hat{s} = \frac{\sigma_b^2}{\sigma_s^2 + \sigma_b^2} x_a + \frac{\sigma_s^2}{\sigma_s^2 + \sigma_b^2} x_b. \tag{5}$$

In the limit $\sigma_b^2 \to 0$, x_b becomes a deterministic effect of s, and its coefficient goes to 1, whereas the coefficient for x_a (the correct causal

predictor) goes to 0. This is a case where empirical risk minimization learns a spurious correlation rather than an invariant cause. In essence, the problem is that there is no way to learn invariant causes without some source of variance.

To address this problem, we need to generalize the setup by considering a set of contexts or environments (indexed by e), each associated with a distribution $p_e(s, x)$. This provides the source of variance that allows us to an identify invariant predictor that captures the causal structure. Naively, you might think that you could just pool all the contexts together and apply the standard empirical risk minimization approach. However, this can still lead to fitting spurious correlations if the variance of the non-causal variables is small. What's needed instead is a predictor that performs well simultaneously in all the contexts, which eliminates non-causal predictors by stress-testing them in contexts where they fail. This is known as *invariant risk minimization* (Arjovsky et al., 2019).

In order to guarantee that causal variables can be identified, the predictor needs access to a sufficiently rich feature representation, such that at least some of the features correspond to causal variables. In Chapter 8, we decomposed $q(s|x)$ into two parts: a non-linear encoder $\phi(x)$, followed by a log-linear decoder:

$$q(s|\phi(x)) \propto \exp\left[\beta_s + w_s \cdot \phi(x)\right], \tag{6}$$

where w_s is a weight vector and we have included a bias term β_s. With this parametrization, the feature representation corresponds to the encoder. A rich feature representation typically entails that the encoding matrix $\Phi = [\phi(x_1), \ldots, \phi(x_M)]$ is high-dimensional and has a large rank (to ensure diversity of features).

For simplicity, we will focus on binary classification, with $s \in \{1, 2\}$, where the log odds takes the following form:

$$\log \frac{q(s=1|\phi(x))}{q(s=2|\phi(x))} = \beta + w \cdot \phi(x), \tag{7}$$

where $w = w_1 - w_2$ and $\beta = \beta_1 - \beta_2$. When the class-conditional distribution over features is Gaussian, $\phi(x)|s \sim \mathcal{N}(\bar{\phi}_s, \Sigma)$, the Bayes-optimal weight vector and bias are given by:

$$w^* = \Sigma^{-1}(\bar{\phi}_1 - \bar{\phi}_2) \tag{8}$$

$$\beta^* = \log \frac{p(s=1)}{p(s=2)} - \frac{1}{2}(\bar{\phi}_1 + \bar{\phi}_2)w^*, \tag{9}$$

where $p(s)$ is the prior distribution over class labels. For isotropic covariances, $\Sigma = \sigma^2 \mathbf{I}$, the optimal weight vector simplifies further:

$$w^* \propto \bar{\phi}_1 - \bar{\phi}_2. \tag{10}$$

The optimal weight vector is proportional to the *coding direction*—the direction in feature space that maximally separates the two classes, which in this case corresponds to the difference between the class-conditional means (Figure 3).

Let's now consider how context affects this picture. Suppose that each context appends a set of extra features that are uncorrelated with the labels. This will change the optimal bias, but will *not* change the optimal weight vector. Thus, w^* defines an invariant predictor: it can be applied across all contexts.

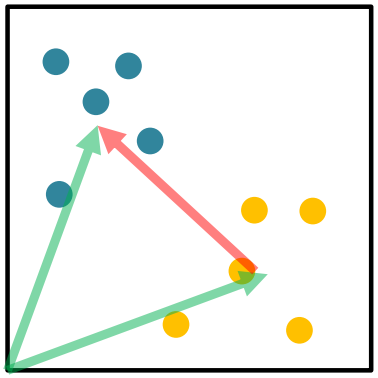

Figure 3: **Binary classification.** The two green arrows represent the vectors pointing at the class-conditional means. The red arrow shows the coding direction obtained by vector subtraction.

1.2 *Context-dependent flexibility*

Our assumption that context essentially adds noise to the classification problem is violated in settings where the labels are correlated with the context. These settings require context-dependent flexibility, where the weight vector is allowed to vary across contexts. At the same time, we still want to seek causal invariants that support abstraction across contexts. To this end, we relax the invariant risk minimization problem as follows:

$$w_e^* = \operatorname*{argmin}_w \hat{L}_e(w) + \lambda \|\bar{w} - w\|^2, \qquad \bar{w} = \frac{1}{N}\sum_e w_e, \tag{11}$$

Note that the subscript on the weight vector now indicates context, not class.

where $\hat{L}_e(w)$ is the empirical risk given weight vector w in context e, and $\bar{w}$ is the average weight vector across all N contexts. The second term regularizes each context-dependent weight towards the average weight; the parameter λ controls the strength of this regularization. When λ is large, w_e^* will tend to be invariant across contexts.

The regularizer implies an intriguing geometric property. Suppose we have two contexts, $e \in \{1, 2\}$. If we sum the regularization terms across contexts and apply the Law of Cosines, we get:

$$\|\bar{w} - w_1\|^2 + \|\bar{w} - w_2\|^2 = \|w_1\|^2 + \|w_2\|^2 - 2\|w_1\|\|w_2\|\cos(\theta), \tag{12}$$

where θ is the angle between the weight vectors. Thus, the regularizer penalizes both large weights (the first two terms) and large angles between the weight vectors. Recall that the optimal unregularized weight vector is proportional to the coding direction (the difference between the class-conditional means). This suggests that if we also optimize an encoder $\phi(x, e)$ defined jointly over sensory inputs and context, we should find representations where the coding directions are approximately parallel across contexts (i.e., $\theta \approx 0$). This geometric property can be found in parts of the brain, as we discuss next.

2 *Representational geometry in the brain*

Bernardi et al. (2020) trained monkeys to perform a context-dependent decision making task in which monkeys could make one of two responses ($a \in \{R, H\}$) to an image (x). They received reward based on a context-dependent reward function. The context switched every 50-70 trials. While monkeys performed this task, the researchers recorded neurons in the hippocampus and two prefrontal areas (the dorsolateral prefrontal cortex, and the anterior cingulate cortex).

To perform well on this task, monkeys should represent the task structure in such a way that the correct action can be decoded from $\phi(x, e)$. For a linear decoder, this only requires that there is some weight vector w that separates the correct and incorrect actions for each context. However, as we've already discussed, this admits spurious correlations that can lead to poor generalization. The framework developed in the last section suggests that we should expect invariant predictors to exhibit parallelism in the representational geometry: the angle between coding directions for different contexts should be close to 0. This was indeed the case for all three brain areas; two examples are shown in Figure 4, where neural representations for the two contexts look like approximately translated copies of one another. This suggests that representations in these areas are optimized for extracting invariant predictors.

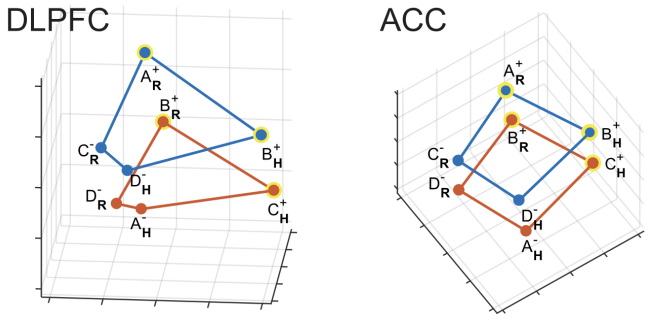

Figure 4: **Representational geometry in two prefrontal areas.** Each point corresponds to neural activity (projected into 3D using multidimensional scaling) for a single context-value-action combination, where the letters denote stimuli, the superscripts denote value (+ reward, - unrewarded), and the subscripts denote actions. The colors correspond to the two contexts. DLPFC: dorsolateral prefrontal cortex; ACC: anterior cingulate cortex. Reproduced from Bernardi et al. (2020).

On its own, parallelism doesn't guarantee flexibility. In fact, parallelism strongly limits flexibility. This is because to achieve perfect parallelism, the representation needs to be factorized:

$$\phi(x, e) = f(x) + g(e). \tag{13}$$

This structure guarantees that context-dependent factors are uncorrelated with input-dependent factors that arise from the class-conditional distribution $p(x|s)$. To quantify flexibility, we can ask how many different dichotomies of M points can be linearly separated (i.e., correctly discriminated by a linear decoder) by a given

representation—the *shattering dimensionality*. In the fully factorized case, the shattering dimensionality is the rank of the representation matrix Φ; if all the columns (corresponding to features) are linearly independent, then the rank is simply the number of features. This is usually much smaller than the total number of possible dichotomies (2^M).

One way to achieve greater flexibility is to add an "interaction" term $\psi(x, e)$:

The interaction term corresponds to what Rigotti et al. (2013) call *nonlinear mixed selectivity*.

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e), \tag{14}$$

where $\epsilon \geq 0$ controls the strength of the interaction term. As long as ϵ is close to 0, parallelism will be approximately satisfied. Importantly, even a small non-zero value of ϵ is sufficient to guarantee that *all* dichotomies are linearly separable, as long as the rank of the representation matrix is at least M. Thus, sacrificing a small amount of parallelism can enable a huge gain in flexibility. This is consistent with observations from Bernardi et al. (2020), who found that all three brain exhibited high shattering dimensionality, despite also having high parallelism.

This result is known as Cover's Theorem (Cover, 1965).

3 *Offline mechanisms for causal learning*

So far, our treatment of invariant prediction has relied on an extrinsic source of variance. In other words, an agent has to actually experience different contexts in order to discover invariant predictors. As such, the prison of experience severely constrains the scope of causal learning. Fortunately, the brain is not truly a prisoner of experience—it can synthesize counterfactual data, providing itself with an alternative source of variance.

3.1 *Learning from randomized data: a function of dream sleep?*

Domain randomization (Tobin et al., 2017) is a powerful and simple method for improving the generalization capabilities of machine learning systems. It was originally developed in robotics, where systems are first trained on simulated data before being deployed in the real world. A common pitfall for such systems is the *reality gap*: poor performance in the real world despite good performance in simulation. One reason this happens is that the learning algorithms fit spurious correlations in the simulated data. Domain randomization addresses this problem by randomizing aspects of the simulator (e.g., viewpoint, color, texture) that are independent of the underlying physical laws (Figure 5). This provides the source of variance needed to learn invariant predictors.

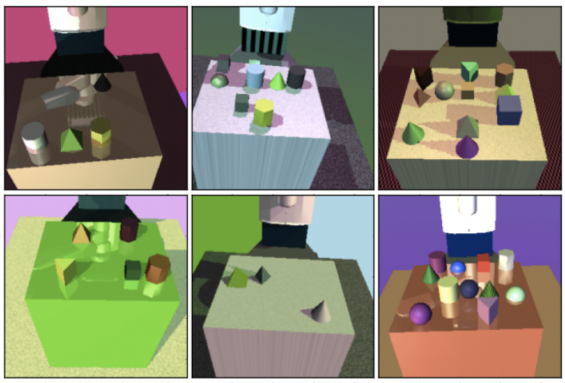

Figure 5: **Images generated from a randomized simulator.** Reproduced from Tobin et al. (2017).

Hoel (2021) has argued that dream sleep might serve a similar function. Dreams often occur in response to repetitive task training. For example, many people trained on a virtual maze navigation task reported task-related mental imagery during sleep, and the occurrence of this imagery was predictive of subsequent performance on a later test with random starting positions (Wamsley et al., 2010). This suggests that dreaming does not merely improve memory—it also improves generalization (see Lewis et al., 2018, for further examples).

The chemist Friedrich Kekulé famously discovered the circular structure of benzene based on a dream in which a snake bit its own tail.

Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep. Mednick et al. (2002) showed that human performance on a visual texture discrimination task declined over several training sessions, except when subjects took a nap between sessions. One explanation centers on the fact that performance in this task is retinotopically specific: changing the location of the stimulus to an untrained region of visual space rescued performance. If some neurons in early visual areas (which are retinotopically organized) are relatively insensitive to visual texture, then these would constitute spurious correlations when stimuli are consistently presented in a particular location. Overfitting these spurious correlations would then lead to performance degradation. If sleeping generates variation not available during waking, it could ameliorate overfitting by breaking the spurious correlations. While there is no direct evidence for this hypothesis, it is consistent with the finding that activation of visual areas during REM sleep (when dreams typically take place) tend to be broader than activation during waking (Igawa et al., 2001).

If dreaming prevents overfitting by generating variation for learning, then we should expect patterns of activation that look different between sleeping and waking states. Analyses of dream diaries indicate that dream content is typically related to recent waking experi-

ence, but is rarely a simple replay of experience (Fosse et al., 2003). Another avenue into this question is the measurement of hippocampal place cells, which are known to activate during sleep. Spatial locations can be decoded from these activations, enabling comparison of decoded trajectories between sleeping and waking states. Notably, these trajectories are not the same (Stella et al., 2019). The same conclusion is reached when examining activations during quiet rest periods, when animals are awake but not moving much (Gupta et al., 2010). These findings are in broad agreement with the hypothesis that offline activation generates a form of domain randomization.

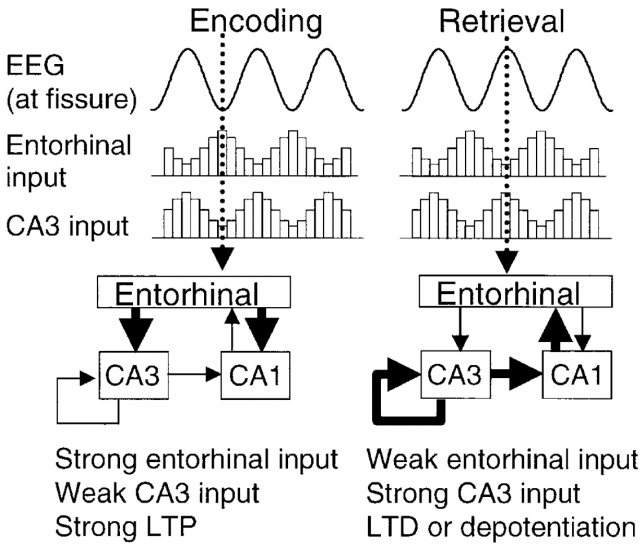

Figure 6: **Separation of encoding and retrieval phases by theta oscillations in the hippocampus.** EEG: electroencephalography; LTP: long-term potentiation; LTD: long-term depression. CA3 and CA1 are subfields of the hippocampus. The fissure is an anatomical landmark at the input to CA1. Reproduced from Hasselmo et al. (2002).

3.2 Oscillating inhibition and contrastive learning

Another way to generate variation in the service of learning is through transient alteration of brain activity. Brain oscillations are a promising candidate for this function. In particular, the hippocampal theta rhythm (4-8 Hz), which arises from inhibitory interneurons (Allen and Monyer, 2015), controls both the level of activity, the relative strength of feedforward vs. feedback/recurrent pathways, and the direction of plasticity (summarized in Figure 6). Hippocampal excitatory (pyramidal) neurons tend to have the highest firing rates near the peak of the theta oscillation (Fox et al., 1986). This phase also coincides with the strongest recurrent activity in subfield CA3 and greater long-term depression at hippocampal synapses (Huerta and Lisman, 1995). An opposite profile is observed at the trough of the theta oscillation: lower firing rates, stronger feedforward activity from entorhinal cortex, and greater long-term potentiation. On the basis of these data, Hasselmo et al. (2002) proposed that a function

of theta oscillations in the hippocampus is to separate encoding and retrieval phases.

The fact that plasticity is still happening (albeit in the opposite direction) during the "retrieval" phase suggests that this isn't pure retrieval, but rather a different kind of encoding. Instead of encoding vs. retrieval, we could think of phase-dependent plasticity as implementing a form of *contrastive learning*, with Hebbian plasticity during the theta trough and anti-Hebbian plasticity during the theta peak (Norman et al., 2006; Ketz et al., 2013). Oscillating inhibition plays an important role here, by generating "positive" examples near the trough (when inhibition is high) and "negative" examples near the peak (when inhibition is low). Positive examples correspond to plausibly invariant causal mechanisms: these reflect patterns of covariation that survive when inhibition intervenes on spurious correlations. Hebbian plasticity strengthens these patterns. When inhibition is reduced near the peak, spurious correlations are revealed and then weakened by anti-Hebbian plasticity. In this way, contrastive learning with oscillating inhibition can learn causal representations.

The idea of using contrastive learning to discover causal representations has also been explored in the machine learning literature (Mitrovic et al., 2020; Wang and Jordan, 2024).

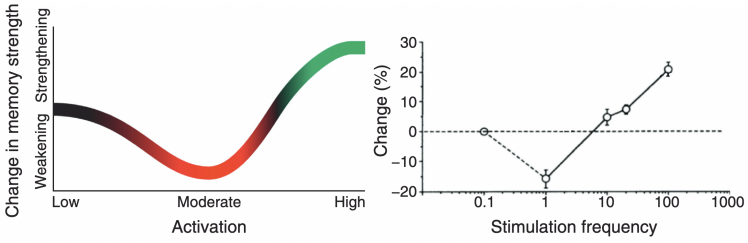

Figure 7: **The nonmonotonic plasticity hypothesis.** (Left) Hypothetical relationship between activation level and plasticity. Strongly activated memories are strengthened; moderately activated memories are weakened. (Right) Data from Kirkwood et al. (1996) showing long-term depression for intermediate-frequency stimulation and long-term potentiation for high-frequency stimulation. Reproduced from Ritvo et al. (2019).

Oscillating inhibition combined with contrastive learning produces a form of plasticity that varies nonmonotonically with neural activation level (Figure 7). When inhibition is high, only the most invariant patterns survive—these get strengthened by Hebbian learning. When inhibition is low, both spurious and invariant patterns are active—these get weakened by anti-Hebbian learning.

4 *Conclusion*

This chapter aimed to demystify some central aspects of high-level intelligence: generalization, abstraction, and causal knowledge. We started with the principle that causality is invariance under intervention: a causal relationship between variables is precisely the structure that remains intact when other aspects of the world change. This principle ties causality tightly to generalization, since invariance is the abstraction needed to make predictions in new contexts. We showed how invariance manifests geometrically as a parallel struc-

ture in neural representations, reflecting the factorization of causal and contextual variables. The parallelism cannot be perfect, however, because some non-linear interaction between these variables (mixed selectivity) is needed to endow the system with a sufficiently rich feature set for generalization across many different prediction problems.

Finally, the chapter discussed biologically plausible algorithms for escaping from the prison of experience: domain randomization by dreaming, and contrastive learning by oscillatory plasticity. Both algorithms have in common the idea that the source of variance needed for learning invariant predictors can be generated intrinsically by the brain. This broadly agrees with theories from cognitive science about the role of counterfactual simulation in causal learning (Gerstenberg, 2024).

This treatment of "causal representation" picks up a thread that was begun in Chapter 3, exploring different principles of representation. For a more comprehensive discussion of causal representation learning, see Schölkopf et al. (2021).

Study questions

1. Why is generalization fundamentally a causal problem rather than a statistical one?
2. Why does perfect parallelism limit flexibility?
3. What kinds of experiments can you think of to test the domain randomization hypothesis about dreaming?

References

Allen, K. and Monyer, H. (2015). Interneuron control of hippocampal oscillations. *Current Opinion in Neurobiology*, 31:81–87.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183:954–967.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science*, 35:404–426.

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334.

Fosse, M. J., Fosse, R., Hobson, J. A., and Stickgold, R. J. (2003). Dreaming and episodic memory: a functional dissociation? *Journal of Cognitive Neuroscience*, 15:1–9.

Fox, S., Wolfson, S., and Ranck Jr, J. (1986). Hippocampal theta rhythm and the firing of neurons in walking and urethane anesthetized rats. *Experimental Brain Research*, 62:495–508.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, 28:924–936.

Gupta, A. S., Van Der Meer, M. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, 65:695–705.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Hasselmo, M. E., Bodelón, C., and Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, 14:793–817.

Hoel, E. (2021). The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2.

Huerta, P. T. and Lisman, J. E. (1995). Bidirectional synaptic plasticity induced by a single burst during cholinergic theta oscillation in CA1 in vitro. *Neuron*, 15:1053–1063.

Igawa, M., Atsumi, Y., Takahashi, K., Shiotsuka, S., Hirasawa, H., Yamamoto, R., Maki, A., Yamashita, Y., and Koizumi, H. (2001). Activation of visual cortex in REM sleep measured by 24-channel NIRS imaging. *Psychiatry and Clinical Neurosciences*, 55:187–188.

Ketz, N., Morkonda, S. G., and O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, 9:e1003067.

Kirkwood, A., Rioult, M. G., and Bear, M. F. (1996). Experience-dependent modification of synaptic plasticity in visual cortex. *Nature*, 381:526–528.

Lewis, P. A., Knoblich, G., and Poe, G. (2018). How memory replay in sleep boosts creative problem-solving. *Trends in Cognitive Sciences*, 22:491–503.

Mednick, S. C., Nakayama, K., Cantero, J. L., Atienza, M., Levin, A. A., Pathak, N., and Stickgold, R. (2002). The restorative effect of naps on perceptual deterioration. *Nature Neuroscience*, 5:677–681.

Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. (2020). Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*.

Norman, K. A., Newman, E., Detre, G., and Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, 18:1577–1610.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497:585–590.

Ritvo, V. J., Turk-Browne, N. B., and Norman, K. A. (2019). Nonmonotonic plasticity: how memory retrieval drives learning. *Trends in Cognitive Sciences*, 23:726–742.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634.

Stella, F., Baracskay, P., O'Neill, J., and Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling Brownian diffusion. *Neuron*, 102:450–461.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ international conference on intelligent robots and systems*, pages 23–30. IEEE.

Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A., and Stickgold, R. (2010). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology*, 20:850–855.

Wang, Y. and Jordan, M. I. (2024). Desiderata for representation learning: A causal perspective. *Journal of Machine Learning Research*.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.