Lecture 15: Generalization, geometry, and causality

Samuel Gershman

Harvard University

Roadmap

- Learning a solution to one problem is useless in the long run unless some aspect of the solution is generalizable to other problems.

Roadmap

- Learning a solution to one problem is useless in the long run unless some aspect of the solution is generalizable to other problems.
- Generalization from the perspective of causality: generalization is most effective when learning invariant predictors that plausibly capture causal relationships between variables, while disregarding spurious relationships.

Roadmap

- Learning a solution to one problem is useless in the long run unless some aspect of the solution is generalizable to other problems.
- Generalization from the perspective of causality: generalization is most effective when learning invariant predictors that plausibly capture causal relationships between variables, while disregarding spurious 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.

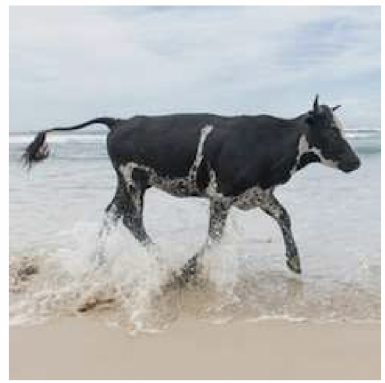
Roadmap

- Learning a solution to one problem is useless in the long run unless some aspect of the solution is generalizable to other problems.
- Generalization from the perspective of causality: generalization is most effective when learning invariant predictors that plausibly capture causal relationships between variables, while disregarding spurious 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, with connections to dreaming and oscillatory plasticity rules.

Recognizing cows with different backgrounds

Convolutional neural networks trained for object recognition 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

[Beery et al 2018]

Causality and generalization

- 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!

Causality and generalization

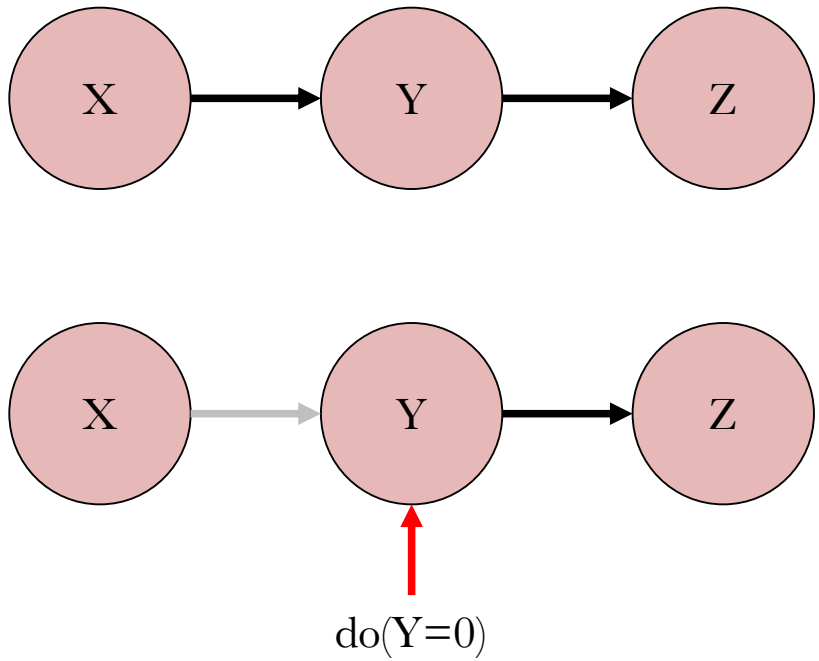
- The essence of the problem is causality: we understand intuitively that pastures don't cause cows to appear in images. Only cows cause the appearance of cows!
- Because cows often graze in pastures, the presence of a pasture makes it more likely that a cow will be there. In other words, we could think of the pasture as a cause of the cow's presence, which in turn causes its appearance in a photo of the pasture.

Causality and generalization

- 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.
- 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.

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 do($Y = 0$), sets Y to 0 and removes the arrow from X to Y. In terms of the example, X is the pasture, Y is the cow presence, and Z is the cow's appearance in the photo. The farmer's intervention is represented by the "do" operator.



Causality and generalization

- The intervention breaks the spurious correlation between X and Y.

Causality and generalization

- The intervention 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).

Causality and generalization

- The intervention 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).

Causality as invariance under intervention

- A system that learns invariant causal mechanisms will generalize correctly to new contexts, whereas a system that learns spurious correlations will not.

Causality as invariance under intervention

- A system that learns invariant causal mechanisms will generalize correctly to new contexts, whereas a system that learns spurious correlations will not.
- We will see how causal invariance is achieved by neural representations with a particular geometry in the high-dimensional space of population activity.

Causality as invariance under intervention

- A system that learns invariant causal mechanisms will generalize correctly to new contexts, whereas a system that learns spurious correlations will not.
- 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.

Causality as invariance under intervention

- Many different strands of thinking about causality have pivoted around some notion of invariance.

Causality as invariance under intervention

- 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.

Causality as invariance under intervention

- 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.

Causality as invariance under intervention

- 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.

Study question

Why is generalization fundamentally a causal problem rather than a statistical one?

The trouble with empirical risk minimization

- Recall the empirical risk minimization setup. 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).

The trouble with empirical risk minimization

- Recall the empirical risk minimization setup. 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).
- Goal is to find a conditional distribution $q(s|x)$, a *predictor*, that minimizes the empirical risk:

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

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

The trouble with empirical risk minimization

- Regression example using continuous labels ($s \in \mathbb{R}$) and two continuous inputs ($x = [x_a, x_b]$). Data-generating process (s is an effect of x_a and a cause of x_b):

$$x_a \sim \mathcal{N}(0, \sigma_a^2), \quad x_b \sim \mathcal{N}(s, \sigma_b^2), \quad s \sim \mathcal{N}(x_a, \sigma_s^2)$$

The trouble with empirical risk minimization

- Regression example using continuous labels ($s \in \mathbb{R}$) and two continuous inputs ($x = [x_a, x_b]$). Data-generating process (s is an effect of x_a and a cause of x_b):

$$x_a \sim \mathcal{N}(0, \sigma_a^2), \quad x_b \sim \mathcal{N}(s, \sigma_b^2), \quad s \sim \mathcal{N}(x_a, \sigma_s^2)$$

- Ideally, we would like our classifier to correctly identify the causal structure, relying only on x_a to predict s.

The trouble with empirical risk minimization

- Regression example using continuous labels ($s \in \mathbb{R}$) and two continuous inputs ($x = [x_a, x_b]$). Data-generating process (s is an effect of x_a and a cause of x_b):

$$x_a \sim \mathcal{N}(0, \sigma_a^2), \quad x_b \sim \mathcal{N}(s, \sigma_b^2), \quad s \sim \mathcal{N}(x_a, \sigma_s^2)$$

- 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 $L(q, s, x) = -\log q(s|x)$, is the posterior, $q(s|x) = p(s|x)$, which in this case is a Gaussian with mean:

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

The trouble with empirical risk minimization

- 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.

The trouble with empirical risk minimization

- 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.
- Empirical risk minimization learns a spurious correlation rather than an invariant cause.

The trouble with empirical risk minimization

- 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.
- 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.

Invariant risk minimization

- We generalize the setup by considering a set of contexts (indexed by e), each associated with a distribution $p_e(s, x)$.

Invariant risk minimization

- We generalize the setup by considering a set of contexts (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.

Invariant risk minimization

- We generalize the setup by considering a set of contexts (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.

Invariant risk minimization

- We generalize the setup by considering a set of contexts (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.
- Need a predictor that performs well simultaneously in all the contexts, which eliminates non-causal predictors by stress-testing them in contexts where they fail.

Invariant risk minimization

- 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.

Invariant risk minimization

- 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.
- Feature representation computed by a non-linear encoder $\phi(x)$. The label distribution is assumed to be a log-linear decoder:

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

where w_s is a weight vector and we have included a bias term β_s.

Invariant risk minimization

- Binary classification, with $s \in \{1, 2\}$, with log odds:

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

where $w = w_1 - w_2$ and $\beta = \beta_1 - \beta_2$.

Invariant risk minimization

- Binary classification, with $s \in \{1, 2\}$, with log odds:

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

  where $w = w_1 - w_2$ and $\beta = \beta_1 - \beta_2$.

- When the class-conditional distribution over features is isotropic Gaussian, $\phi(x)|s \sim \mathcal{N}(\bar{\phi}_s, \sigma^2 \mathbf{I})$, the Bayes-optimal weight vector is:

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

Invariant risk minimization

- Binary classification, with $s \in \{1, 2\}$, with log odds:

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

  where $w = w_1 - w_2$ and $\beta = \beta_1 - \beta_2$.

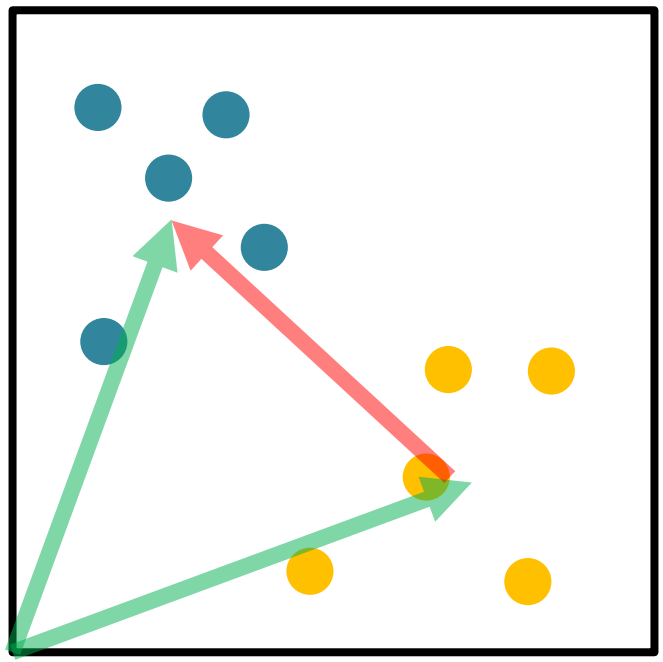
- When the class-conditional distribution over features is isotropic Gaussian, $\phi(x)|s \sim \mathcal{N}(\bar{\phi}_s, \sigma^2 \mathbf{I})$, the Bayes-optimal weight vector is:

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

- 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.

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.



Invariant risk minimization

- Consider how context affects this picture. Suppose that each context appends a set of extra features that are uncorrelated with the labels.

Invariant risk minimization

- 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.

Invariant risk minimization

- 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.

Context-dependent flexibility

- Assumption that context essentially adds noise to the classification problem is violated in settings where the labels are correlated with the context.

Context-dependent flexibility

- 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.

Context-dependent flexibility

- 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.
- But we still want to seek causal invariants that support abstraction across contexts.

Context-dependent flexibility

- We relax the invariant risk minimization problem:

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

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.

Context-dependent flexibility

- We relax the invariant risk minimization problem:

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

  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.

Context-dependent flexibility

- 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:

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

where θ is the angle between the weight vectors.

Context-dependent flexibility

- 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:

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

  where θ is the angle between the weight vectors.
- The regularizer penalizes both large weights (the first two terms) *and* large angles between the weight vectors.

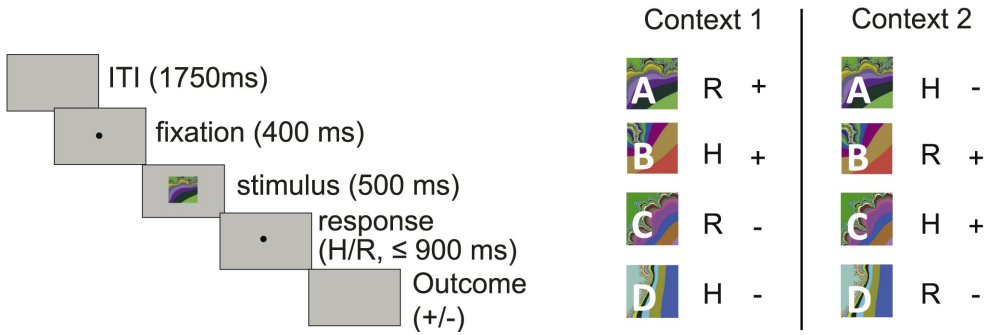
Context-dependent flexibility

- 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:

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

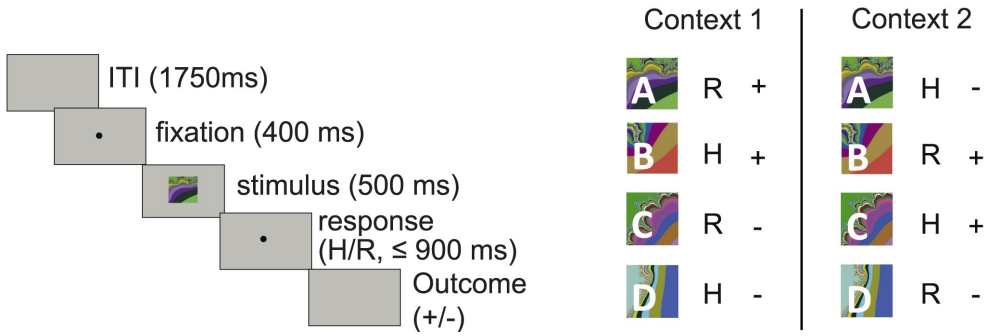
  where θ is the angle between the weight vectors.
- The regularizer penalizes both large weights (the first two terms) *and* large angles between the weight vectors.
- Recall that the optimal unregularized weight vector is proportional to the coding direction (the difference between the class-conditional 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., $\theta \approx 0$).

Representational geometry in the brain



- Bernardi et al [2020] trained monkeys to perform a context-dependent decision making task with two possible responses ($a \in \{R, H\}$) to an image (x). Monkeys received reward based on a context-dependent reward function.

Representational geometry in the brain



- Bernardi et al [2020] trained monkeys to perform a context-dependent decision making task with two possible responses ($a \in \{R, H\}$) to an image (x). Monkeys received reward based on a context-dependent reward function.
- Context switched every 50-70 trials.

Representational geometry in the brain

- To perform well on this task, monkeys should represent the task structure in such a way that the correct action can be decoded from $\phi(x, e)$.

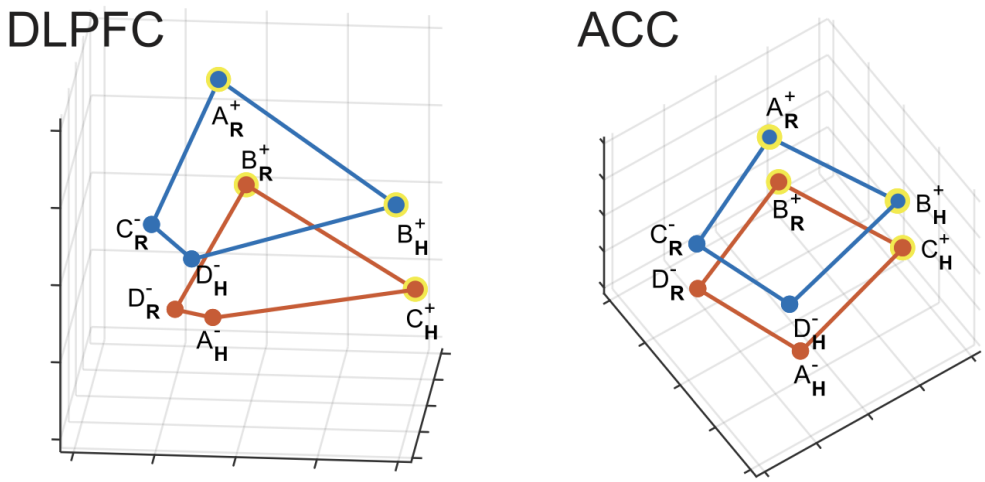
Representational geometry in the brain

- 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.

Representational geometry in the brain

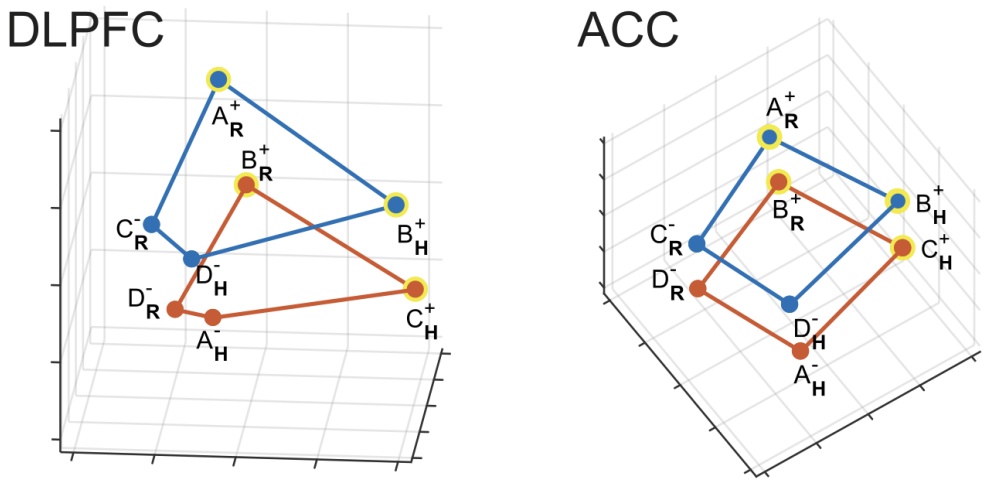
- To perform well on this task, monkeys should represent the task structure in such a way that the correct action can be decoded from $\phi(x, e)$.
- For a linear decoder, this only requires that there is some weight vector w that separates the correct and incorrect actions for each context.
- However, this admits spurious correlations that can lead to poor generalization.

Representational geometry in the brain



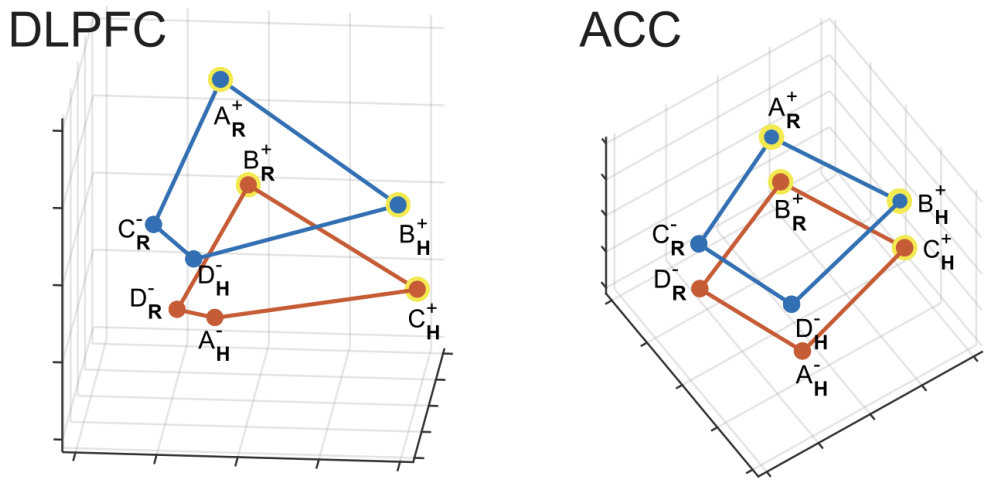
- We should expect invariant predictors to exhibit parallelism in the representational geometry: the angle between coding directions for different contexts should be close to 0.

Representational geometry in the brain



- 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 in the prefrontal cortex and hippocampus: neural representations for the two contexts look like approximately translated copies of one another.

Representational geometry in the brain



- 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 in the prefrontal cortex and hippocampus: neural representations for the two contexts look like approximately translated copies of one another.
- Suggests that representations in these areas are optimized for extracting invariant predictors.

Representational geometry in the brain

- Parallelism limits flexibility, because the representation needs to be factorized:

$$\phi(x, e) = f(x) + g(e).$$

Representational geometry in the brain

- Parallelism limits flexibility, because the representation needs to be factorized:

$$\phi(x, e) = f(x) + g(e).$$

- This structure guarantees that context-dependent factors are uncorrelated with input-dependent factors that arise from the class-conditional distribution $p(x|s)$.

Representational geometry in the brain

- Parallelism limits flexibility, because the representation needs to be factorized:

$$\phi(x, e) = f(x) + g(e).$$

- 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*.

Representational geometry in the brain

- Parallelism limits flexibility, because the representation needs to be factorized:

$$\phi(x, e) = f(x) + g(e).$$

- 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*.
- Fully factorized case: shattering dimensionality is the rank of the representation matrix Φ; if all the columns (features) are linearly independent, the rank is the number of features. Usually much smaller than the number of possible dichotomies (2^M).

Representational geometry in the brain

- To achieve greater flexibility, add an "interaction" term $\psi(x, e)$:

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e)$$

where $\epsilon \geq 0$ controls the strength of the interaction term.

Representational geometry in the brain

- To achieve greater flexibility, add an "interaction" term $\psi(x, e)$:

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e)$$

  where $\epsilon \geq 0$ controls the strength of the interaction term.
- As long as ϵ is close to 0, parallelism will be approximately satisfied.

Representational geometry in the brain

- To achieve greater flexibility, add an "interaction" term $\psi(x, e)$:

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e)$$

  where $\epsilon \geq 0$ controls the strength of the interaction term.
- As long as ϵ is close to 0, parallelism will be approximately satisfied.
- 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.

Representational geometry in the brain

- To achieve greater flexibility, add an "interaction" term $\psi(x, e)$:

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e)$$

  where $\epsilon \geq 0$ controls the strength of the interaction term.
- As long as ϵ is close to 0, parallelism will be approximately satisfied.
- 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.

Representational geometry in the brain

- To achieve greater flexibility, add an "interaction" term $\psi(x, e)$:

$$\phi(x, e) = f(x) + g(e) + \epsilon\psi(x, e)$$

  where $\epsilon \geq 0$ controls the strength of the interaction term.
- As long as ϵ is close to 0, parallelism will be approximately satisfied.
- 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.
- Consistent with observations from Bernardi et al [2020]: PFC and hippocampus exhibited high shattering dimensionality, despite also having high parallelism.

Offline mechanisms for causal learning

- Our treatment of invariant prediction has relied on an extrinsic source of variance: an agent has to actually experience different contexts in order to discover invariant predictors.

Offline mechanisms for causal learning

- Our treatment of invariant prediction has relied on an extrinsic source of variance: an agent has to actually experience different contexts in order to discover invariant predictors.
- Fortunately, the brain is not truly a prisoner of experience—it can synthesize counterfactual data, providing itself with an alternative source of variance.

Learning from randomized data: a function of dream sleep?

- **Domain randomization** is a powerful and simple method for improving the generalization capabilities of machine learning systems.

Learning from randomized data: a function of dream sleep?

- **Domain randomization** is a powerful and simple method for improving the generalization capabilities of machine learning systems.
- Originally developed in robotics, where systems are first trained on simulated data before being deployed in the real world.

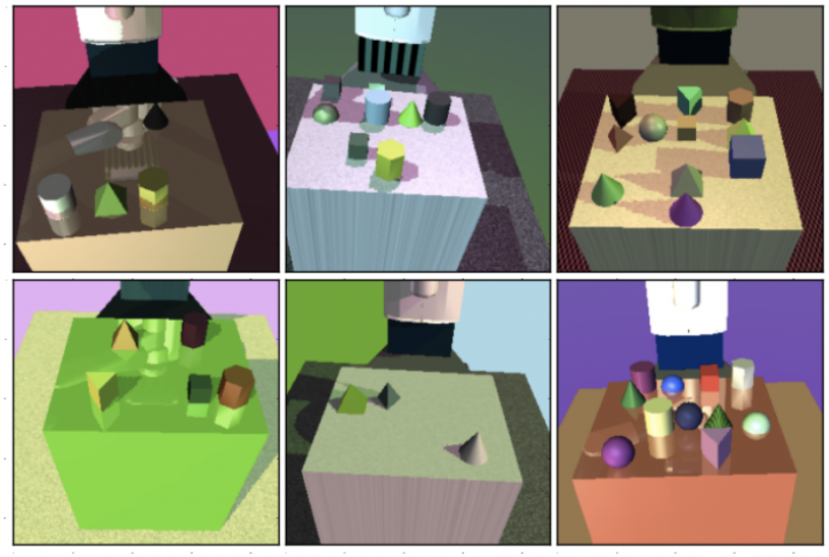
Learning from randomized data: a function of dream sleep?

- **Domain randomization** is a powerful and simple method for improving the generalization capabilities of machine learning systems.
- Originally developed in robotics, where systems are first trained on simulated data before being deployed in the real world.
- **Reality gap**: poor performance in the real world despite good performance in simulation, often because learning algorithms fit spurious correlations in the simulated data.

Learning from randomized data: a function of dream sleep?

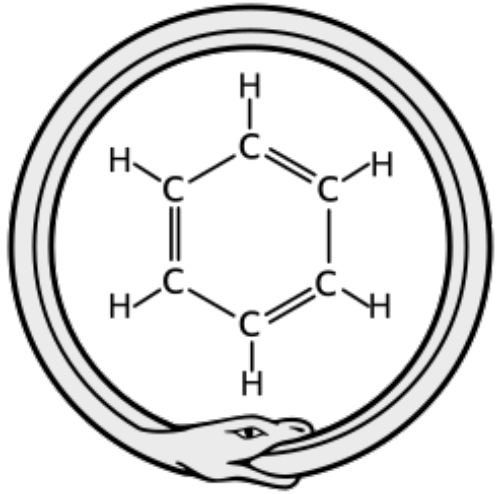
- **Domain randomization** is a powerful and simple method for improving the generalization capabilities of machine learning systems.
- Originally developed in robotics, where systems are first trained on simulated data before being deployed in the real world.
- **Reality gap**: poor performance in the real world despite good performance in simulation, often because 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. This provides the source of variance needed to learn invariant predictors.

Images generated from a randomized simulator



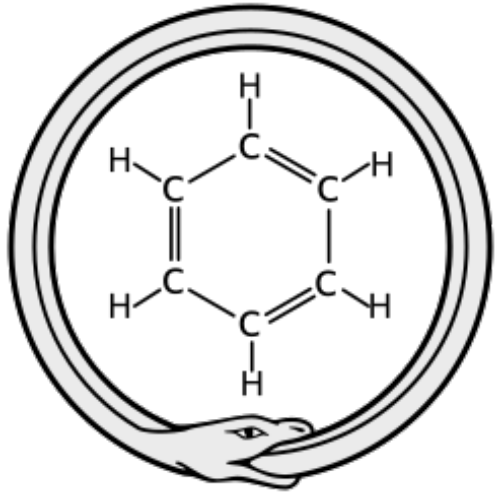
[Tobin et al 2017]

Learning from randomized data: a function of dream sleep?



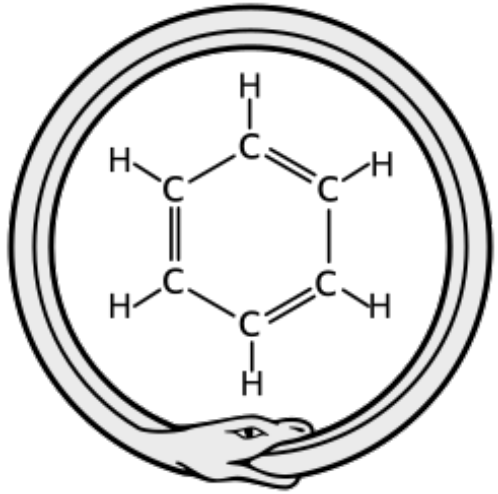
- Dream sleep might serve a function similar to domain randomization [Hoel 2021].

Learning from randomized data: a function of dream sleep?



- Dream sleep might serve a function similar to domain randomization [Hoel 2021].
- Dreams often occur in response to repetitive task training. People trained on a virtual maze navigation task reported task-related mental imagery during sleep, and the occurrence of this imagery predicted later performance on a test with random starting positions [Wamsley et al 2010].

Learning from randomized data: a function of dream sleep?



- Dream sleep might serve a function similar to domain randomization [Hoel 2021].
- Dreams often occur in response to repetitive task training. People trained on a virtual maze navigation task reported task-related mental imagery during sleep, and the occurrence of this imagery predicted later performance on a test with random starting positions [Wamsley et al 2010].
- Suggests that dreaming does not merely improve memory—it also improves generalization

Learning from randomized data: a function of dream sleep?

- Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep.

Learning from randomized data: a function of dream sleep?

- Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep.
- Human performance on a visual texture discrimination task declined over several training sessions, except when subjects took a nap between sessions [Mednick et al 2002].

Learning from randomized data: a function of dream sleep?

- Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep.
- Human performance on a visual texture discrimination task declined over several training sessions, except when subjects took a nap between sessions [Mednick et al 2002].
- Performance in this task is retinotopically specific: changing the location of the stimulus to an untrained region of visual space rescued performance.

Learning from randomized data: a function of dream sleep?

- Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep.
- Human performance on a visual texture discrimination task declined over several training sessions, except when subjects took a nap between sessions [Mednick et al 2002].
- 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 retinotopic visual areas 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 degrade performance.

Learning from randomized data: a function of dream sleep?

- Overtraining can sometimes lead to performance degradation, possibly due to overfitting, which can be reversed with sleep.
- Human performance on a visual texture discrimination task declined over several training sessions, except when subjects took a nap between sessions [Mednick et al 2002].
- 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 retinotopic visual areas 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 degrade performance.
- If sleeping generates variation unavailable during waking, it could ameliorate overfitting by breaking spurious correlations.

Learning from randomized data: a function of dream sleep?

- If dreaming prevents overfitting by generating variation for learning, then we should expect patterns of activation that look different between sleeping and waking states.

Learning from randomized data: a function of dream sleep?

- 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 experience, but is rarely a simple replay of experience [Fosse et al 2003].

Learning from randomized data: a function of dream sleep?

- 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 experience, but is rarely a simple replay of experience [Fosse et al 2003].
- Spatial trajectories decoded from hippocampus differ between sleeping and waking states [Stella et al 2019].

Learning from randomized data: a function of dream sleep?

- 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 experience, but is rarely a simple replay of experience [Fosse et al 2003].
- Spatial trajectories decoded from hippocampus differ between sleeping and waking states [Stella et al 2019].
- These findings are in broad agreement with the hypothesis that offline activation generates a form of domain randomization.

Oscillating inhibition

- Another way to generate variation in the service of learning: transient alteration of brain activity.

Oscillating inhibition

- Another way to generate variation in the service of learning: transient alteration of brain activity.
- The hippocampal theta rhythm (4-8 Hz), which arises from inhibitory interneurons, controls both the level of activity, the relative strength of feedforward vs. feedback/recurrent pathways, and the direction of plasticity.

Oscillating inhibition

- Another way to generate variation in the service of learning: transient alteration of brain activity.
- The hippocampal theta rhythm (4-8 Hz), which arises from inhibitory interneurons, controls both the level of activity, the relative strength of feedforward vs. feedback/recurrent pathways, and the direction of plasticity.
- Hippocampal excitatory (pyramidal) neurons tend to have the highest firing rates near the peak of the theta oscillation.

Oscillating inhibition

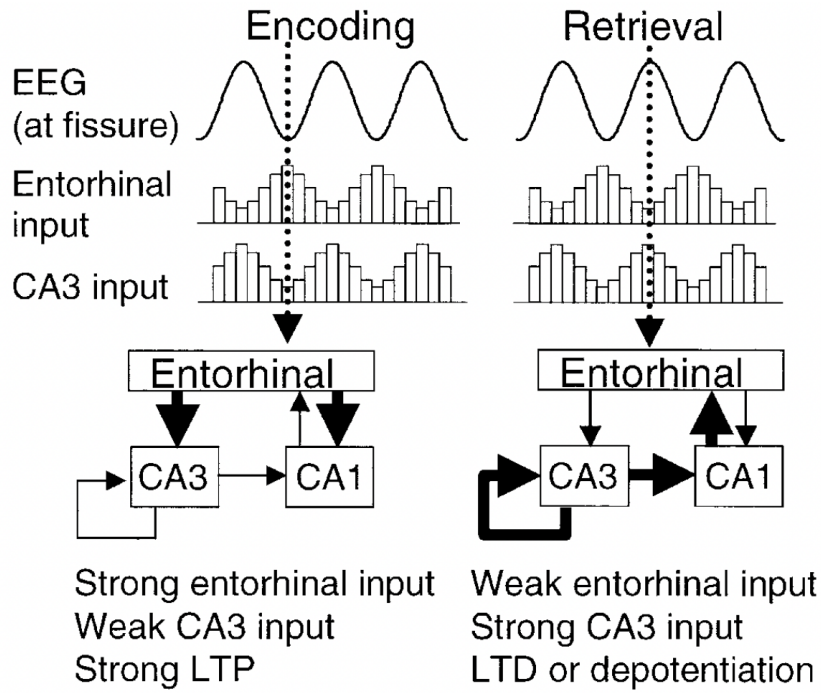
- Another way to generate variation in the service of learning: transient alteration of brain activity.
- The hippocampal theta rhythm (4-8 Hz), which arises from inhibitory interneurons, controls both the level of activity, the relative strength of feedforward vs. feedback/recurrent pathways, and the direction of plasticity.
- Hippocampal excitatory (pyramidal) neurons tend to have the highest firing rates near the peak of the theta oscillation.
- This phase also coincides with the strongest recurrent activity in subfield CA3 and greater long-term depression at hippocampal synapses.

Oscillating inhibition

- Another way to generate variation in the service of learning: transient alteration of brain activity.
- The hippocampal theta rhythm (4-8 Hz), which arises from inhibitory interneurons, controls both the level of activity, the relative strength of feedforward vs. feedback/recurrent pathways, and the direction of plasticity.
- Hippocampal excitatory (pyramidal) neurons tend to have the highest firing rates near the peak of the theta oscillation.
- This phase also coincides with the strongest recurrent activity in subfield CA3 and greater long-term depression at hippocampal synapses.
- Opposite profile at the trough of the theta oscillation: lower firing rates, stronger feedforward activity from entorhinal cortex, and greater long-term potentiation.

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.



[Hasselmo et al 2012]

Oscillating inhibition and contrastive learning

- The fact that plasticity is still happening during the "retrieval" phase suggests that this isn't pure retrieval.

Oscillating inhibition and contrastive learning

- The fact that plasticity is still happening during the "retrieval" phase suggests that this isn't pure retrieval.
- Phase-dependent plasticity could implement a form of **contrastive learning**, with Hebbian plasticity during the theta trough and anti-Hebbian plasticity during the theta peak.

Oscillating inhibition and contrastive learning

- The fact that plasticity is still happening during the "retrieval" phase suggests that this isn't pure retrieval.
- Phase-dependent plasticity could implement a form of **contrastive learning**, with Hebbian plasticity during the theta trough and anti-Hebbian plasticity during the theta peak.
- Oscillating inhibition can generate "positive" examples near the trough (when inhibition is high) and "negative" examples near the peak (when inhibition is low).

Oscillating inhibition and contrastive learning

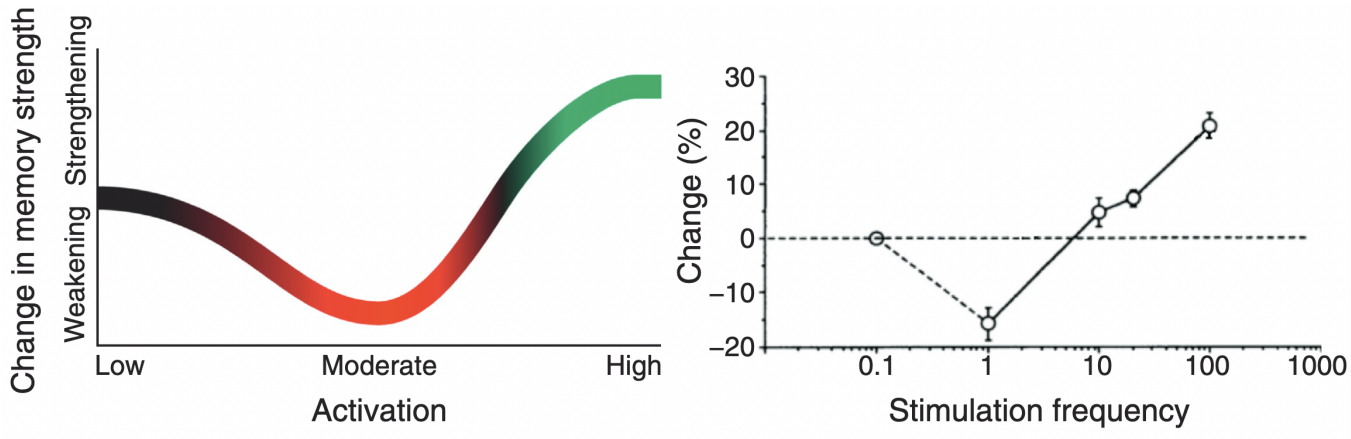
- The fact that plasticity is still happening during the "retrieval" phase suggests that this isn't pure retrieval.
- Phase-dependent plasticity could implement a form of **contrastive learning**, with Hebbian plasticity during the theta trough and anti-Hebbian plasticity during the theta peak.
- Oscillating inhibition can generate "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.

Oscillating inhibition and contrastive learning

- The fact that plasticity is still happening during the "retrieval" phase suggests that this isn't pure retrieval.
- Phase-dependent plasticity could implement a form of **contrastive learning**, with Hebbian plasticity during the theta trough and anti-Hebbian plasticity during the theta peak.
- Oscillating inhibition can generate "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.

The nonmonotonic plasticity hypothesis

(Left) Hypothetical relationship between activation level and plasticity. Strongly activated memories are strengthened; moderately activated memories are weakened. (Right) Long-term depression for intermediate-frequency stimulation and long-term potentiation for high-frequency stimulation.



[Ritvo et al 2019; Kirkwood et al 1996]

Summary

- Central aspects of high-level intelligence: generalization, abstraction, and causal knowledge.

Summary

- Central aspects of high-level intelligence: generalization, abstraction, and causal knowledge.
- Causal invariance under intervention: a causal relationship between variables is the structure that remains intact when other aspects of the world change.

Summary

- Central aspects of high-level intelligence: generalization, abstraction, and causal knowledge.
- Causal invariance under intervention: a causal relationship between variables is 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.

Summary

- Invariance manifests geometrically as a parallel structure in neural representations, reflecting the factorization of causal and contextual variables.

Summary

- Invariance manifests geometrically as a parallel structure in neural representations, reflecting the factorization of causal and contextual variables.
- Parallelism cannot be perfect: 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.

Summary

- Invariance manifests geometrically as a parallel structure in neural representations, reflecting the factorization of causal and contextual variables.
- Parallelism cannot be perfect: 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.
- Domain randomization by dreaming or contrastive learning by oscillatory plasticity can generate sources of variance needed for learning invariant predictors, escaping the prison of experience.