Lecture 3: Principles of perceptual representation

Samuel Gershman

Harvard University

Roadmap

- If neural computation is the manipulation of representations, what are those representations?

Roadmap

- If neural computation is the manipulation of representations, what are those representations?
- Focusing on perception, we can organize principles of representation into a small set of general principles (efficiency, sparsity, prediction), formulated as different optimization problems.

Efficient coding

- Key idea: neural representations are optimized to communicate information, subject to a set of resource constraints.

Efficient coding

- Key idea: neural representations are optimized to communicate information, subject to a set of resource constraints.
- Constraints: max firing rate, discrete levels (due to spikes), noise.

Efficient coding

- Key idea: neural representations are optimized to communicate information, subject to a set of resource constraints.
- Constraints: max firing rate, discrete levels (due to spikes), noise.
- These constraints mean that a neuron has an upper bound on how much information it can communicate about its inputs. The efficient coding principle states that the neuron should be configured to operate at this upper bound.

Single neuron as a communication channel

- Receives inputs ("messages") about the state (s) that it communicates to downstream neurons via its firing rate x (the channel output).

Single neuron as a communication channel

- Receives inputs ("messages") about the state (s) that it communicates to downstream neurons via its firing rate x (the channel output).
- Firing rates can distinguish M different input levels.

Single neuron as a communication channel

- Receives inputs ("messages") about the state (s) that it communicates to downstream neurons via its firing rate x (the channel output).
- Firing rates can distinguish M different input levels.
- With M levels, a neuron can communicate up to $\log M$ bits per sample. This upper bound is achieved when each firing rate is used with equal frequency across the distribution of inputs.

Quantifying uncertainty

- A neuron's activity x is informative to the extent that it allows downstream neurons to reduce their uncertainty about the state s.

Quantifying uncertainty

- A neuron's activity x is informative to the extent that it allows downstream neurons to reduce their uncertainty about the state s.
- The "surprisal" of observing s (measured in bits) is defined as $-\log p(s)$.

Quantifying uncertainty

- A neuron's activity x is informative to the extent that it allows downstream neurons to reduce their uncertainty about the state s.
- The "surprisal" of observing s (measured in bits) is defined as $-\log p(s)$.
- Uncertainty can be quantified as the average surprisal, or *entropy*:

$$\mathcal{H}[s] = \mathbb{E}[-\log p(s)] = -\sum_s p(s) \log p(s)$$

Quantifying uncertainty

Uncertainty about s after observing neural activity x is the *conditional entropy*:

$$\mathcal{H}[s|x] = \mathbb{E}[-\log p(s|x)] = -\sum_x p(x) \sum_s p(s) \log p(s|x)$$

Information

The uncertainty reduction afforded by observing x is the difference between these two entropies, the *mutual information*:

$$\mathcal{I}[s;x] = \mathcal{H}[s] - \mathcal{H}[s|x] = \mathcal{H}[x] - \mathcal{H}[x|s]$$

(note the symmetry)

Information

- $\mathcal{H}[x]$ measures the variability of firing rates across the distribution of inputs.

Information

- $\mathcal{H}[x]$ measures the variability of firing rates across the distribution of inputs.
- $\mathcal{H}[x|s]$ measures transmission noise. If we assume that transmission noise is negligible, then $\mathcal{H}[x|s] \approx 0$.

Maximizing information

- When $\mathcal{H}[x|s] \approx 0$, maximizing information corresponds to maximizing output entropy $\mathcal{H}[x]$.

Maximizing information

- When $\mathcal{H}[x|s] \approx 0$, maximizing information corresponds to maximizing output entropy $\mathcal{H}[x]$.
- This is achieved when the output response distribution $p(x)$ is uniform, which can be implemented by setting the tuning function to be the cumulative distribution function of the stimulus distribution $p(s)$:

$$f(s) \propto P(s) = \int_{s' \leq s} p(s')ds'.$$

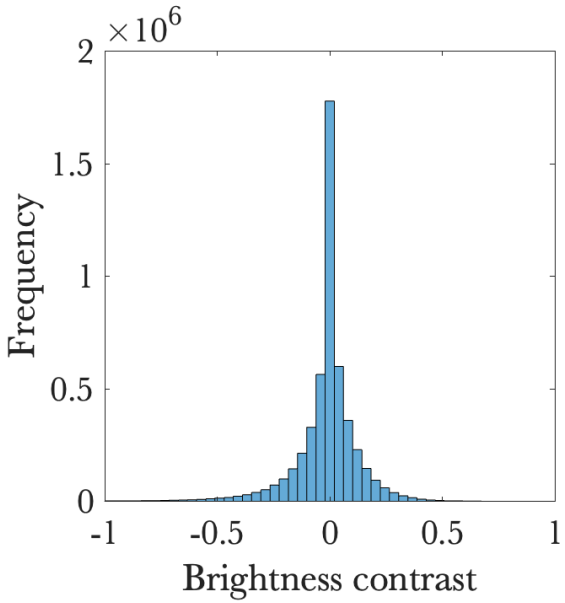
Maximizing information

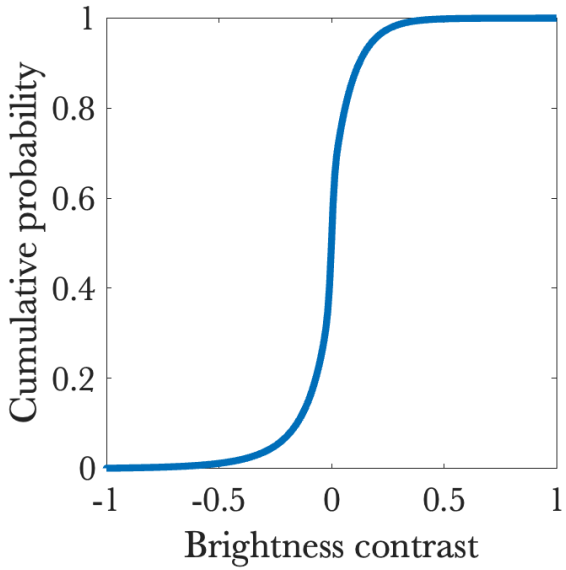
- When $\mathcal{H}[x|s] \approx 0$, maximizing information corresponds to maximizing output entropy $\mathcal{H}[x]$.
- This is achieved when the output response distribution $p(x)$ is uniform, which can be implemented by setting the tuning function to be the cumulative distribution function of the stimulus distribution $p(s)$:

$$f(s) \propto P(s) = \int_{s' \leq s} p(s')ds'.$$

- This tuning curve is a rank transformation, where the firing rate for a stimulus corresponds to its normalized rank in the stimulus distribution $\Rightarrow$ high sensitivity to changes in regions where rank changes quickly.

Brightness contrast statistics

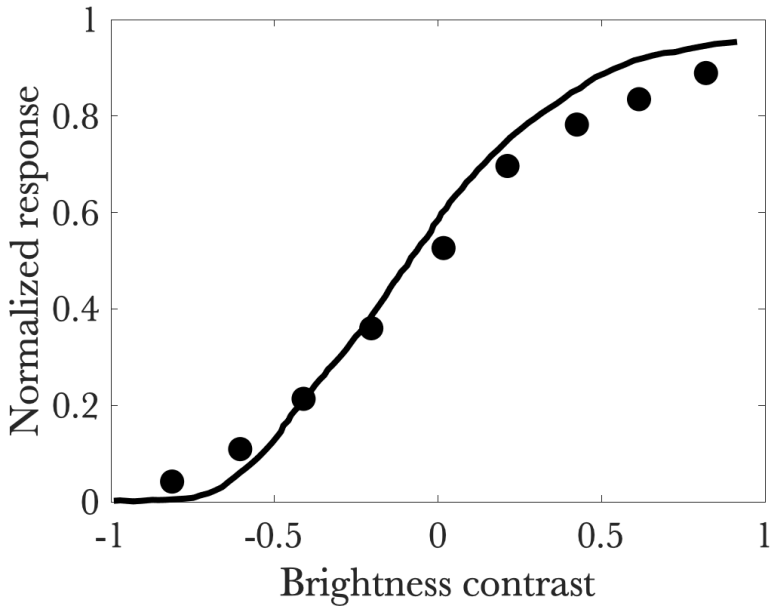




The distribution of brightness contrast in a natural image and the cumulative distribution function.

Efficient coding in the blowfly eye

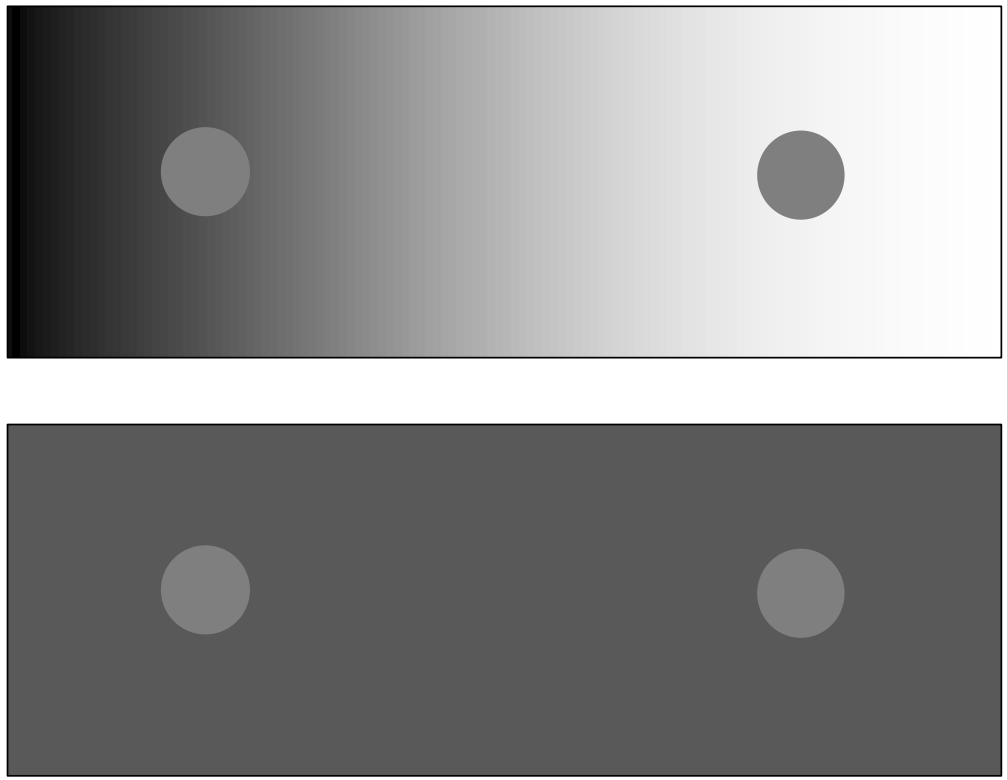
Normalized responses of large monopolar cells at different contrast levels. The line shows the cumulative distribution function of contrast estimated from images of the fly's natural environment.



[Laughlin 1981]

Brightness illusions

Same stimulus maps onto different firing rates depending on the stimulus distribution, such that a brightness difference is perceived where there is no objective difference.



Predictive coding

- The rank transformation removes information about the mean (stimulus expectation), encoding only deviations (prediction errors).

Predictive coding

- The rank transformation removes information about the mean (stimulus expectation), encoding only deviations (prediction errors).
- This strategy supports metabolic efficiency: only generate spikes in response to incompressible surprises.

Study question

Representations are conceptualized here in terms of tuning functions. What are the limitations of this conceptualization when compared with a more dynamical view of neural computation?

Efficient coding with a convolutional population

- Idealized population of neurons with identical, uniformly spaced tuning functions; each idealized tuning function is a shifted copy of a "prototype" tuning function $\tilde{f}$.

Efficient coding with a convolutional population

- Idealized population of neurons with identical, uniformly spaced tuning functions; each idealized tuning function is a shifted copy of a "prototype" tuning function $\tilde{f}$.
- Tiling property:

$$\sum_d f_d(s - s_d) \approx 1$$

where $\{s_d\}$ is a set of evenly spaced points in stimulus space (the *stimulus lattice*).

Efficient coding with a convolutional population

- Idealized population of neurons with identical, uniformly spaced tuning functions; each idealized tuning function is a shifted copy of a "prototype" tuning function $\tilde{f}$.
- Tiling property:

$$\sum_d f_d(s - s_d) \approx 1$$

  where $\{s_d\}$ is a set of evenly spaced points in stimulus space (the *stimulus lattice*).
- The optimal tuning function warps the preferred stimuli to maximize an approximation of the mutual information, the Fisher information $J(s)$, subject to an upper bound G on the population firing rate:

$$f^* = \underset{f}{\operatorname{argmax}}\, \mathbb{E}[\log J(s)|f]$$

Fisher information

- Fisher information is defined as:

$$J(s) = \sum_x p(x|s) \left[\frac{\partial}{\partial s} \log p(x|s) \right]^2$$

where x is the spike count vector for the population.

Fisher information

- Fisher information is defined as:

$$J(s) = \sum_x p(x|s) \left[\frac{\partial}{\partial s} \log p(x|s) \right]^2$$

  where x is the spike count vector for the population.
- Can be interpreted as an approximation of the mutual information, but more tractable to analyze.

Fisher information

- Fisher information is defined as:

$$J(s) = \sum_x p(x|s) \left[\frac{\partial}{\partial s} \log p(x|s) \right]^2$$

  where x is the spike count vector for the population.
- Can be interpreted as an approximation of the mutual information, but more tractable to analyze.
- For independent Poisson neurons:

$$J(s) = \sum_d \frac{f_d'(s)^2}{f_d(s)}$$

Optimal tuning

- Parametrized tuning function:

$$f_d(s) = g(s_d^*)\tilde{f}(\Gamma(s) - s_d)$$

with gain $g(s)$, warping function $\Gamma(s)$, and preferred stimulus (after warping) $s_d^* = \Gamma^{-1}(s_d)$, where $\Gamma(s)$ is the CDF of density $\gamma(s)$ controlling resource allocation across neurons.

Optimal tuning

- Parametrized tuning function:

$$f_d(s) = g(s_d^*)\tilde{f}(\Gamma(s) - s_d)$$

with gain $g(s)$, warping function $\Gamma(s)$, and preferred stimulus (after warping) $s_d^* = \Gamma^{-1}(s_d)$, where $\Gamma(s)$ is the CDF of density $\gamma(s)$ controlling resource allocation across neurons.

- Optimal solution:

$$\gamma(s) \propto p(s), \qquad g(s) \propto G$$

Optimal tuning

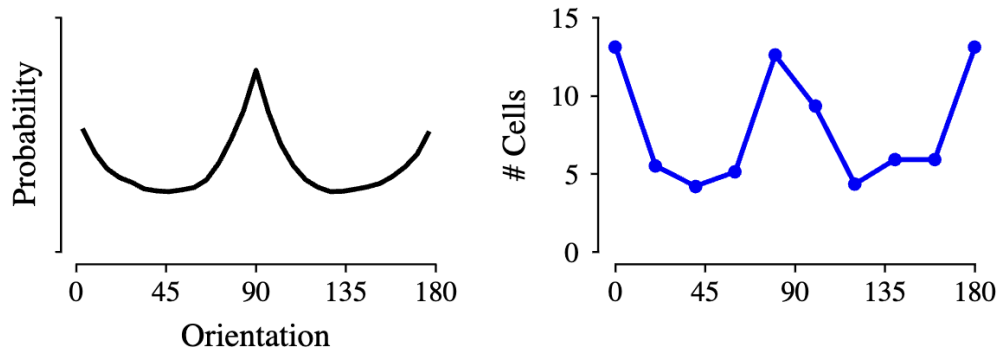
- Under the Ganguli & Simoncelli model, the distribution of preferred stimuli should match the prior distribution: optimal preferred stimuli correspond to samples from $p(s)$.

Optimal tuning

- Under the Ganguli & Simoncelli model, the distribution of preferred stimuli should match the prior distribution: optimal preferred stimuli correspond to samples from $p(s)$.
- This is because the optimal warping function is the CDF of the stimulus distribution, and the preferred stimuli are obtained by taking the inverse CDF evaluated at each stimulus on the stimulus lattice.

Optimal tuning

Distribution of orientations in natural scenes, and the distribution of preferred orientations in V1.



[Ganguli & Simoncelli 2010]

Fisher information: the fundamental unit of analysis

- Under efficient coding [Wei & Stocker 2015]:

$$J(s) \propto p(s)^2$$

Fisher information: the fundamental unit of analysis

- Under efficient coding [Wei & Stocker 2015]:

$$J(s) \propto p(s)^2$$

- Thus, we can formalize the efficient coding principle without making specific claims about tuning functions. It provides a powerful abstraction that we will discuss more in the next lecture.

Sparsity

- Only a small proportion of neurons are active at any given time.

Sparsity

- Only a small proportion of neurons are active at any given time.
- For example, on average 2.5% of V1 neurons are active for each natural image [Yoshida et al, 2020]. Of these responsive neurons, only 5.4% of them exhibited overlap between pairs of images.

Sparsity

- Sensory data arise from many different causes, only a few of which are present at any given moment.

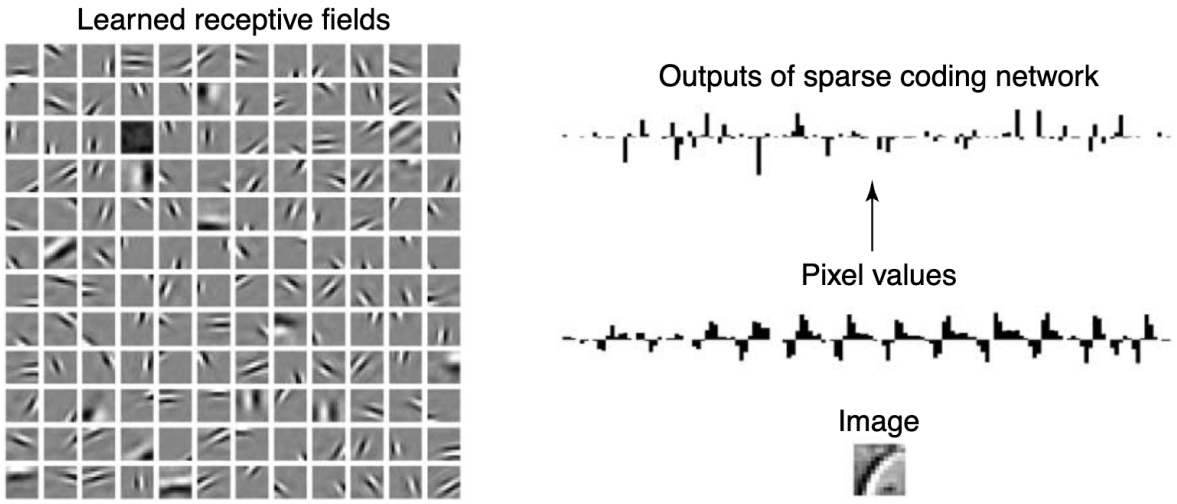
Sparsity

- Sensory data arise from many different causes, only a few of which are present at any given moment.
- If your eyes scan a scene, the high-dimensional time series of retinal images arises from different glimpses of a slowly changing object set. The retinal images live on a low-dimensional subspace defined by the set of currently active causes (objects).

Sparsity

- Sensory data arise from many different causes, only a few of which are present at any given moment.
- If your eyes scan a scene, the high-dimensional time series of retinal images arises from different glimpses of a slowly changing object set. The retinal images live on a low-dimensional subspace defined by the set of currently active causes (objects).
- Perceptual inference: which causes are active at any given moment? Perceptual learning: what is the stable mapping between causes and images?

Visual receptive fields from sparse coding



[Olshausen & Field, 2004]

The metabolic argument for sparsity

- Two major contributors to energy consumption: spiking and synaptic transmission.

The metabolic argument for sparsity

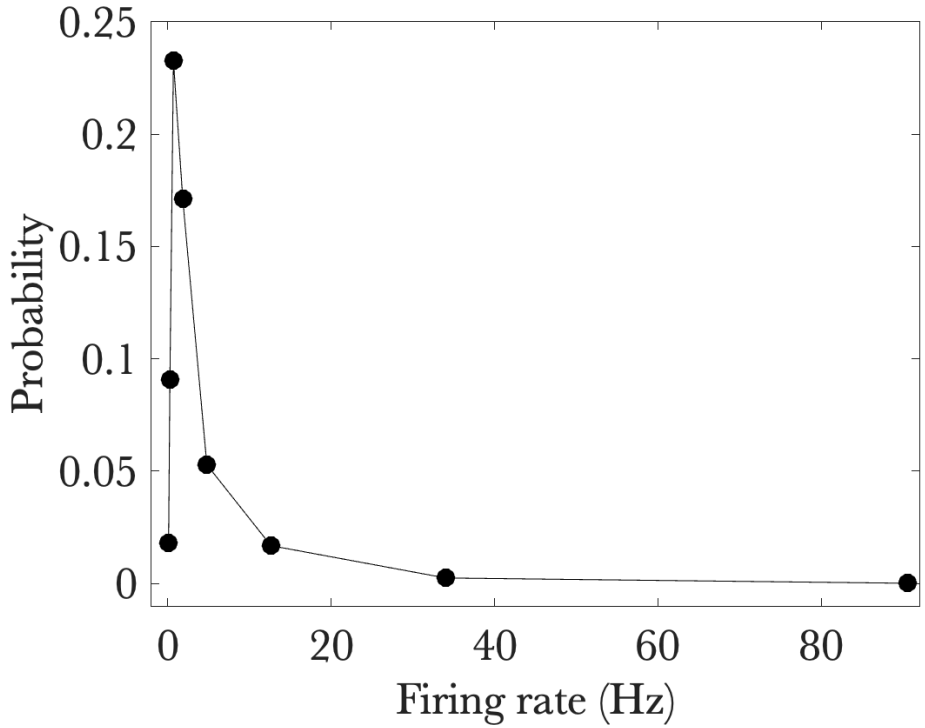
- Two major contributors to energy consumption: spiking and synaptic transmission.
- A single spike in human cortex costs 2.4×10^9 molecules of ATP [Lennie 2003].

The metabolic argument for sparsity

- Two major contributors to energy consumption: spiking and synaptic transmission.
- A single spike in human cortex costs 2.4×10^9 molecules of ATP [Lennie 2003].
- Cortical neurons need to spike on average less than once per second in order to satisfy the energy budget. This is remarkably low given the fact that studies have reported spike rates of up to 100 Hz.

Stimulus responses in auditory cortex

Neurons typically spike close to 1 Hz, but can infrequently achieve much higher spike rates.



[Hromadka et al 2008]

The metabolic argument for sparsity

- For a "strong" response to a stimulus (spike rate of 10 Hz over 200 ms), the energy budget could support concurrent spiking in 0.3% of neurons.

The metabolic argument for sparsity

- For a "strong" response to a stimulus (spike rate of 10 Hz over 200 ms), the energy budget could support concurrent spiking in 0.3% of neurons.
- Even allowing for large transient increases in glucose consumption during intense sensory stimulation, the average spike rate can only increase by a few spikes per second $\Rightarrow$ 4% of concurrently active neurons spiking at 50 Hz.

The metabolic argument for sparsity

- For a "strong" response to a stimulus (spike rate of 10 Hz over 200 ms), the energy budget could support concurrent spiking in 0.3% of neurons.
- Even allowing for large transient increases in glucose consumption during intense sensory stimulation, the average spike rate can only increase by a few spikes per second $\Rightarrow$ 4% of concurrently active neurons spiking at 50 Hz.
- Takeaway: **metabolic constraints necessitate sparsity of neural activity**.

Study question

The energy efficiency of the brain is remarkable (its power usage is comparable to a dim light bulb). What might we learn about the design of energy-efficient artificial systems from studying the brain?

Representations optimized for prediction

- The future can, at least partially, be predicted from the past.

Representations optimized for prediction

- The future can, at least partially, be predicted from the past.
- This predictability can be exploited by perceptual systems.

Anticipatory saccades

- Our high-acuity (foveal) vision is limited to a small portion of the visual field (approximately twice the width of your thumbnail held at arm's length).

Anticipatory saccades

- Our high-acuity (foveal) vision is limited to a small portion of the visual field (approximately twice the width of your thumbnail held at arm's length).
- We perceive much more than the central 2 degrees because our eyes are making frequent saccades—ballistic, high-velocity movements to salient regions of the visual field.

Anticipatory saccades

- Our high-acuity (foveal) vision is limited to a small portion of the visual field (approximately twice the width of your thumbnail held at arm's length).
- We perceive much more than the central 2 degrees because our eyes are making frequent saccades—ballistic, high-velocity movements to salient regions of the visual field.
- Saccades to unpredictable stimuli usually take around 200 ms. In contrast, saccades to predictable stimuli can be initiated *even before the stimulus appears*.

Anticipatory saccades

- Our high-acuity (foveal) vision is limited to a small portion of the visual field (approximately twice the width of your thumbnail held at arm's length).
- We perceive much more than the central 2 degrees because our eyes are making frequent saccades—ballistic, high-velocity movements to salient regions of the visual field.
- Saccades to unpredictable stimuli usually take around 200 ms. In contrast, saccades to predictable stimuli can be initiated *even before the stimulus appears*.
- This is useful in a fast-changing but predictable world, where predictive saccades can increase the rate of information flow.

Covert attention

- Even without eye movements, prediction can improve perception.

Covert attention

- Even without eye movements, prediction can improve perception.
- A centrally presented cue, indicating the likely future location of a target, speeds detection of the target when it appears in the cued location, and slows down detection when it appears unexpectedly in an uncued location [Posner 1980].

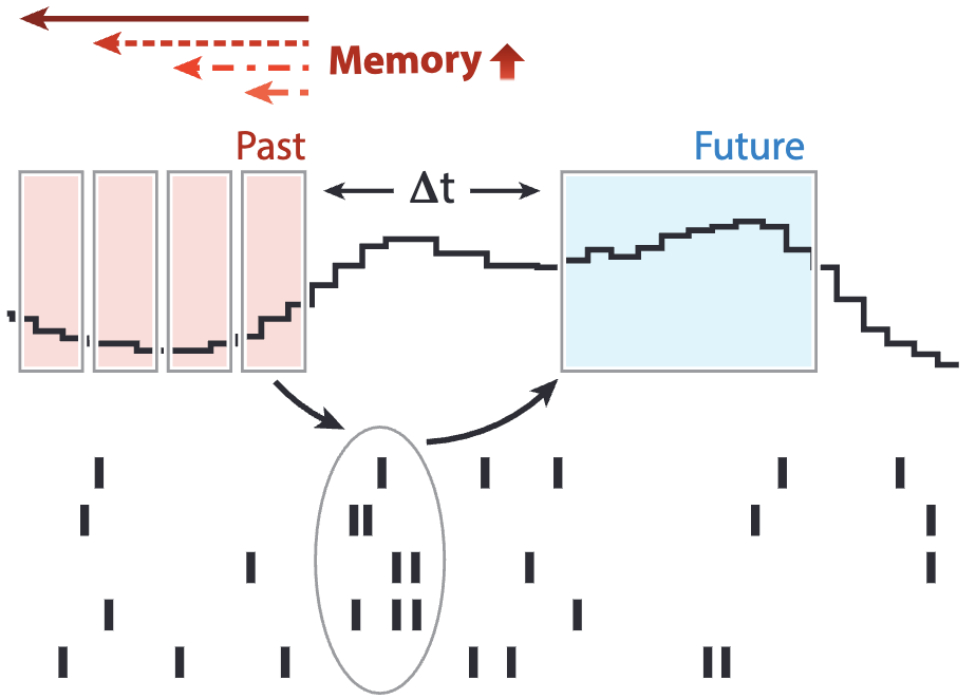
The predictive information bottleneck principle

- Let s_{past} denote the history of stimuli, and s_{future} denote future stimuli that haven't been observed yet.

The predictive information bottleneck principle

- Let s_{past} denote the history of stimuli, and s_{future} denote future stimuli that haven't been observed yet.
- A population of neurons encodes the stimulus history into its spiking activity x. If this population carries predictive information, it should be able to predict the future trajectory of the stimulus over some timescale.

The prediction problem



[Rust & Palmer 2021]

The predictive information bottleneck solution

- An optimal predictive representation $x = f^*(s_{\text{past}})$ should maximize predictability of the future subject to a constraint on memory of the past [Bialek et al, 2001]:

$$f^* = \underset{f}{\text{argmax}}\, \mathcal{I}[x; s_{\text{future}}], \qquad \text{subject to } \mathcal{I}[s_{\text{past}}; x] \leq C.$$

The predictive information bottleneck solution

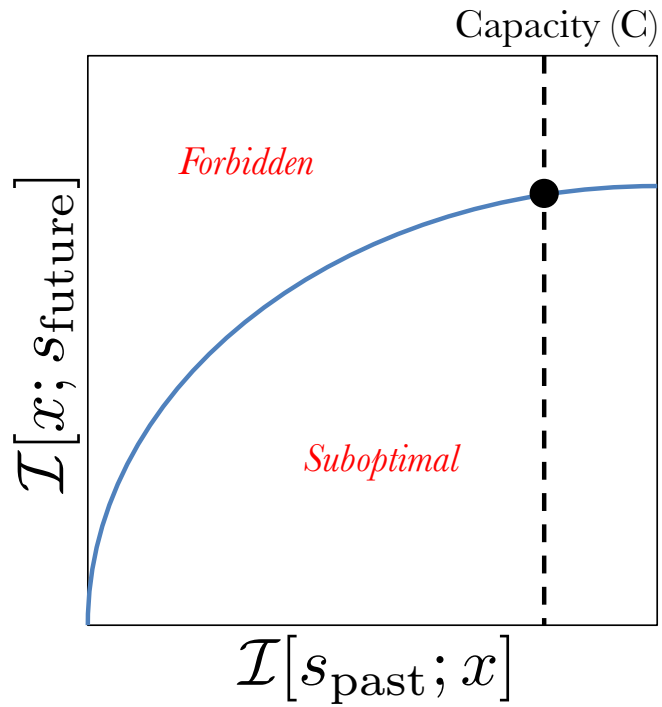
- An optimal predictive representation $x = f^*(s_{\text{past}})$ should maximize predictability of the future subject to a constraint on memory of the past [Bialek et al, 2001]:

$$f^* = \underset{f}{\text{argmax}}\, \mathcal{I}[x; s_{\text{future}}], \qquad \text{subject to } \mathcal{I}[s_{\text{past}}; x] \leq C.$$

- For different choices of the capacity parameter C, we can chart an optimality frontier. This tells us the highest achievable predictive information for a given constraint on memory capacity.

The predictive information bottleneck solution

Blue curve: optimality frontier. The dashed line shows a hypothetical capacity parameter.



Predictive information in the retina

- The retina is one of the earliest stages of vision in which signatures of prediction are present.

Predictive information in the retina

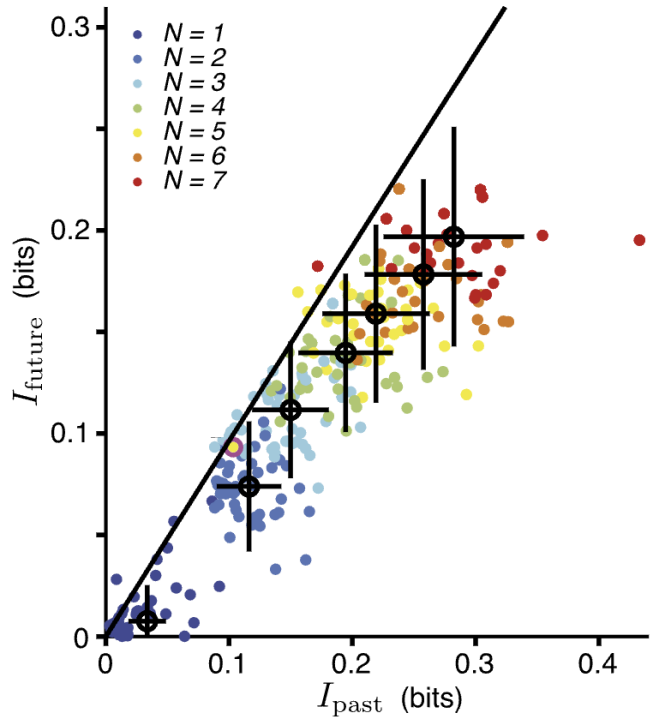
- The retina is one of the earliest stages of vision in which signatures of prediction are present.
- A moving bar evokes a wave of activation in retinal ganglion cells that tracks the leading edge of the bar [Berry et al, 1999].

Predictive information in the retina

- The retina is one of the earliest stages of vision in which signatures of prediction are present.
- A moving bar evokes a wave of activation in retinal ganglion cells that tracks the leading edge of the bar [Berry et al, 1999].
- This is remarkable given that firing latency of retinal ganglion cells to unpredictable flashes is around 50 ms. The population apparently learns to compensate for this delay by anticipating the bar position.

Predictive information in the retina

Color denote groups of cells of different sizes (N). Black line: optimality frontier.



[Palmer et al, 2015]

Predictiveness vs. efficiency

- Although both the predictive information bottleneck and predictive coding solutions involve predictions, what they do with these predictions is quite different.

Predictiveness vs. efficiency

- Although both the predictive information bottleneck and predictive coding solutions involve predictions, what they do with these predictions is quite different.
- In the predictive information bottleneck, only the predictively useful information is kept.

Predictiveness vs. efficiency

- Although both the predictive information bottleneck and predictive coding solutions involve predictions, what they do with these predictions is quite different.
- In the predictive information bottleneck, only the predictively useful information is kept.
- In contrast, predictive coding discards predictive information by only encoding prediction errors. Since sparse coding can arise from efficient coding, predictiveness may also sometimes be at odds with sparsity.

Predictiveness vs. efficiency

- Although both the predictive information bottleneck and predictive coding solutions involve predictions, what they do with these predictions is quite different.
- In the predictive information bottleneck, only the predictively useful information is kept.
- In contrast, predictive coding discards predictive information by only encoding prediction errors. Since sparse coding can arise from efficient coding, predictiveness may also sometimes be at odds with sparsity.
- Note, however, that predictions still need to be retained in some form (not necessarily spiking activity) in order to compute prediction errors.

Study question

Efficiency, sparsity, and prediction principles are partly complementary but sometimes contradictory. How might these principles be reconciled into a single unifying framework of perceptual representation? To what extent are they incompatible?

Summary

- While no single principle can explain all the relevant empirical phenomena, a small set of principles has a remarkably wide scope.

Summary

- While no single principle can explain all the relevant empirical phenomena, a small set of principles has a remarkably wide scope.
- Unifying idea: representations should be optimized to encode information that is useful for certain tasks (reconstruction, inference, prediction).