Quantifying Humans' Priors Over Graphical Representations of Tasks

Gecia Bravo Hermsdorff(✉), Talmo Pereira, and Yael Niv

Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
geciah@princeton.edu

Abstract. Some new tasks are trivial to learn while others are almost impossible; what determines how easy it is to learn an arbitrary task? Similar to how our prior beliefs about new visual scenes colors our perception of new stimuli, our priors about the structure of new tasks shapes our learning and generalization abilities [2]. While quantifying visual priors has led to major insights on how our visual system works [5,10,11], quantifying priors over tasks remains a formidable goal, as it is not even clear how to define a task [4]. Here, we focus on tasks that have a natural mapping to graphs. We develop a method to quantify humans' priors over these "task graphs", combining new modeling approaches with Markov chain Monte Carlo with people, MCMCP (a process whereby an agent learns from data generated by another agent, recursively [9]). We show that our method recovers priors more accurately than a standard MCMC sampling approach. Additionally, we propose a novel low-dimensional "smooth" (In the sense that graphs that differ by fewer edges are given similar probabilities.) parametrization of probability distributions over graphs that allows for more accurate recovery of the prior and better generalization. We have also created an online experiment platform that gamifies our MCMCP algorithm and allows subjects to interactively draw the task graphs. We use this platform to collect human data on several navigation and social interactions tasks. We show that priors over these tasks have non-trivial structure, deviating significantly from null models that are insensitive to the graphical information. The priors also notably differ between the navigation and social domains, showing fewer differences between cover stories within the same domain. Finally, we extend our framework to the more general case of quantifying priors over exchangeable random structures.

Keywords: Markov chain Monte Carlo with People (MCMCP)
Representational learning · Structural priors · Task graphs
Human cognition

1 Introduction

1.1 Our Brain Must Utilize Efficient Priors

Our lives are punctuated by a multitude of seemingly disparate new tasks (e.g., navigating in a new place, interacting with new people, writing a new abstract)

A. J. Morales et al. (Eds.): ICCS 2018, SPCOM, pp. 281–290, 2018.
https://doi.org/10.1007/978-3-319-96661-8_30

that we are able to perform with relative ease. Still, if we consider all tasks we could possibly be faced with, we would not be good (at least initially) at most of them (e.g., playing Go). This simple observation leads to a fundamental, yet unanswered, question in cognitive neuroscience: are there essential structural properties that unite the tasks that our brains are "naturally" good at solving, and if so, what are they?

Understanding our brain's representation of a new task (i.e., our prior about the task's structure) is key to answering this question. Indeed, the prior used in a given task sharply constrains how fast and efficiently (if at all) this task can be solved [2]. In particular, the curse of dimensionality [3] suggests that task representations should be compact, filtering out redundancies. However, there is no free lunch; reduced representations also constrain the set of tasks an agent can efficiently solve. Thus, these reduced representations should manifest as priors that leverage on the relevant structure of naturalistic tasks, i.e., tasks that the organism encounters in everyday life and have been relevant over evolutionary time-scales [4].

1.2 Main Contributions

Quantifying priors over tasks is a formidable goal [2,4], if not only for the reason that "what is a task?" is a relatively open-ended question. Here we restrict our attention to tasks that have a natural mapping to graphs as this allows us to quantify their structure using graph theoretical tools. Specifically, our experiments focus on two prominent domains of naturalistic tasks: navigation and social interaction, with nodes and edges representing, for example, regions and borders, or people and relationships.

On the theoretical side, we develop a method to quantify humans' priors over these "task graphs", which combines new modeling approaches with Markov chain Monte Carlo with people (MCMCP) [6,9] – a process whereby an agent learns from data generated by another agent, recursively. Our simulations demonstrate that our method recovers these priors more accurately than a standard MCMC sampling approach. This result is particularly relevant for the "resource constrained" regime, where data are limited and costly to acquire, such as in experiments with human subjects. Moreover, we propose a novel low-dimensional "smooth" parametrization of probability distributions over (non-isomorphic) graphs on the same vertex set. We show that, in the limited data regime, it allows for more accurate recovery of the prior (*in silico* data), and better generalization (in human data). Finally, we extend our framework to the more general case of quantifying priors over exchangeable random structures [14].

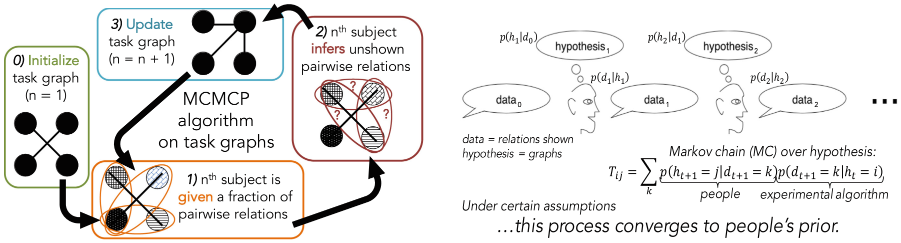
On the experimental side, we have created an online experimental platform[1] with a game-like interface that instantiates our MCMCP algorithm, and allows

[1] Links for some of our experiments:
http://psiturk-geciah.princeton.edu:9001/ (navigation in nature parks),
http://psiturk-geciah.princeton.edu:9003/ (navigation in cities),
http://psiturk-geciah.princeton.edu:9000/ (friendships in workplaces),
http://psiturk-geciah.princeton.edu:9002/ (friendships in school classes).

subjects to interactively draw the task graphs. We use this platform to collect human data on several navigation and social interactions tasks. We show that priors over these tasks have non-trivial structure, deviating significantly from null models that are insensitive to the graphical information. Moreover, the priors are notably different between the navigation and social domains, while exhibiting fewer differences between different tasks in the same domain (city and nature park (navigation); coworkers and students (social)).

2 Markov Chain Monte Carlo with People (MCMCP) over (task) Graphs

Figure 1a illustrates our MCMCP algorithm for generating experiments. For example, in one of our experiments, the subject is told that they are visiting a new city, and informed whether certain pairs of neighborhoods share a border or not (*step 1* in Fig. 1a). They then are asked to guess if the other pairs of neighborhoods share a border or not (*step 2*) by drawing a map using our graph drawing interface. Additionally, to incentivize subjects to give their true prior, they are told that there is an underlying truth (e.g., an actual city), and that they win extra money by correctly guessing the relations obscured.


0) Initialize task graph (n = 1)
3) Update task graph (n = n + 1)
2) nth subject infers unshown pairwise relations
?
?
?
MCMCP algorithm on task graphs
1) nth subject is given a fraction of pairwise relations
$p(h_1|d_0)$
hypothesis$_1$
$p(h_2|d_1)$
hypothesis$_2$
$p(d_1|h_1)$
$p(d_2|h_2)$
data$_0$
data$_1$
data$_2$
...
data = relations shown
hypothesis = graphs
Markov chain (MC) over hypothesis:
$T_{ij} = \sum_k \underbrace{p(h_{t+1}=j|d_{t+1}=k)}_{\text{people}} \underbrace{p(d_{t+1}=k|h_t=i)}_{\text{experimental algorithm}}$
Under certain assumptions
...this process converges to people's prior.

(a) (b)

Fig. 1. *Performing Markov chain Monte Carlo with people (MCMCP) in a given task allows for sampling from their prior over it.* (a) *Schema of our algorithm for generating experiments: (0)* create a task graph for the first subject; *(1)* the subject is given partial information about this graph (a fraction of the pairwise relations chosen at random, here, 3 out of 6); *(2)* they are asked to infer the unshown portions (the remaining pairwise relations); *(3)* construct a new task graph from these responses for the next subject; *(4)* repeat steps *1–3.* (b) *The algorithm interpreted as MCMC.*

This back-and-forth between data seen by the subjects (relations shown, or "partial graphs") and the resulting hypothesis they infer (completed "task graphs") can be marginalized over the partial graphs to create a Markov chain (MC) over the space of task graphs (Fig. 1b). Assuming that subjects are Markovian, and share the same fixed decision rule, this MC is time-homogeneous. If, in addition, we assume that subjects are "Bayesian", computing Bayes rule using the correct likelihood function[2] and a shared prior, and respond by sampling from their posterior, this MC has as its stationary distribution the subjects' prior over the relevant task graphs.[3] Precisely, this "MCMCP Bayesian model" gives a transition matrix $\underline{\underline{\mathbf{T}}}$ over the relevant non-isomorphic task graphs, with entries:

$$t_{ij} = p(g_j|g_i) = \sum\nolimits_k p(g_j|d_k)p(d_k|g_i) \tag{1}$$

where $p(d_k|g_i)$ is the probability of seeing partial graph d_k by randomly obscuring r relations of the graph g_i (with $r :=$ total number of relations minus number of relations shown), and $p(g_j|d_k)$ is given by Bayes rule using a fixed prior.

3 Results

3.1 Resource Constrained MCMCP

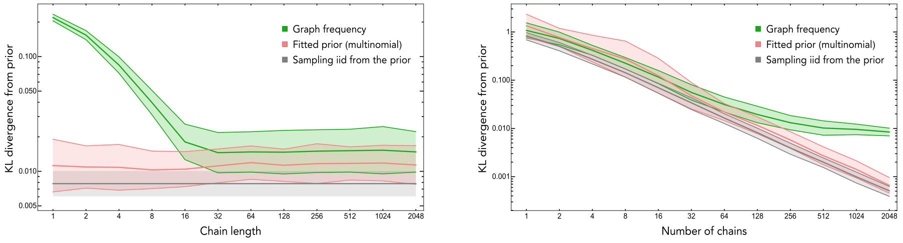
In standard MCMC, one uses samples generated by the algorithm to reconstruct the target (stationary) distribution. This is inefficient in the following sense: to obtain i.i.d. samples, only a small fraction of the iterations are used as samples, as one must discard the initial samples until (hopefully[4]) the chain has converged to its stationary distribution (the so called "burn-in" period). In addition, one might only collect samples every $\mathcal{O}(\tau_c)$ iterations (where $\mathcal{O}(\tau_c)$ is the autocorrelation time) to mitigate correlations. While these issues are generally not a concern when samples are efficiently generated via a computer, in MCMCP the primary bottleneck is due to the use of human subjects.

Fortunately, in our case, we can use data more efficiently by leveraging the additional structure provided by the Bayesian assumption. Specifically, we propose to recover subjects' prior by fitting their choices to our MCMCP Bayesian model (as opposed to using the observed graph frequency as a proxy of the prior as is done in classic MCMC). The unknowns are the probabilities that the prior gives to each of the non-isomorphic graphs on the relevant vertex set. As illustrated in Fig. 2, our fitting method recovers the prior more precisely than a standard MCMC sampling method (especially in the case of constrained chain length). Moreover, aside from using data more efficiently, our approach has additional advantages: it does not have the problem of "guessing" the mixing time (which can vary substantially depending on the prior, number of nodes, and number of relations obscured); and it also allows for experiments to be run in parallel.

[2] I.e., that the partial graphs are generated by randomly erasing a fraction of the relations, which they are told in our experiments.

[3] Of course, we also need to assume that the chain is ergodic; a fair assumption given humans' non-zero probability of doing something strange/unexpected.

[4] As determining convergence can be non-trivial in certain cases, especially when the state space is large.



(a)

(b)

Fig. 2. *Priors can be more precisely recovered by leveraging the MCMCP assumptions.* **(a)** We simulated data from our MCMCP Bayesian model on 5 nodes (34 non-isomorphic graphs), with a prior chosen to give an asymptotic mixing time of $\tau_m \sim 13$ iterations. Each simulation has 2048 data points, split into different chain lengths. We then fit a multinomial prior to these data, and measured the KL divergence from the true prior (that generated the data) for the fitted prior and for the sampled frequency of graphs. Error bars denote ±1 standard deviation. Sampling i.i.d. from the true prior is shown in gray as reference. Notice that using the graph frequency is doomed to fail when the chain length is short as there is not enough time for the chain to approach the prior. Moreover, fitting the prior does better than using the graph frequency even when the chain is much longer than τ_m. **(b)** Here we fix the chain length to 16, and vary the number of chains (same specifications as before). As the number of data points increases, fitting the prior continues to improve, while using the graph frequency asymptotes to a finite error.

3.2 A Natural Low-Dimensional Parametrization of Distributions Over Graphs

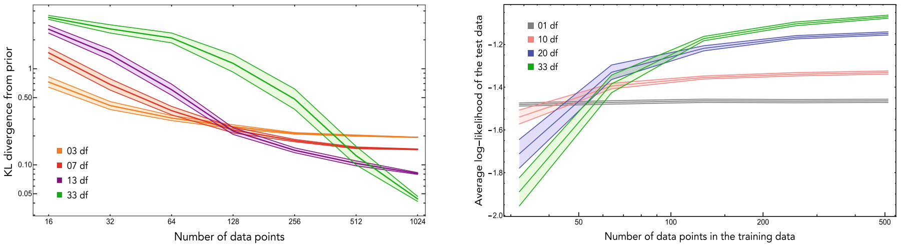
The number of non-isomorphic graphs $G(n)$ on n nodes grows superexponentially [15]; given limited data, even for moderate n we cannot sufficiently sample them all. For these cases, to obtain informative priors, we need to extend the probabilities to graphs that were not sampled. Our approach is to find a natural low-dimensional parameterization of the prior. Specifically, we propose to use the following form for the prior $\underline{\mathbf{p}}$:

$$\underline{\mathbf{p}} \propto \mathrm{ER}(1/2) * \exp \sum_{b=2}^{G(n)} c_b \underline{\mathbf{v}}_b \tag{2}$$

where c_b are the coefficients to be fit, $\underline{\mathbf{v}}_b$ is the b^{th} right eigenvector (ordered by decreasing eigenvalues) of the transition matrix $\underline{\underline{\mathbf{T}}}$ from Eq. 1, with the data generated by obscuring one relation of the underlying graph, and using an ER(1/2) (Erdős-Rényi model with $p = 1/2$) distribution as the prior.

This choice has several interesting properties,[5] for example, when $c_{i>2} = 0$, there is a unique correspondence between $c_2 \in (-\infty, \infty)$ and an $ER(p)$ prior with

[5] Formal details about this parametrization and its extension to hypergraphs to appear in a paper under preparation by Bravo Hermsdorff and Gunderson.



(a)

(b)

Fig. 3. *Benefits of our low-dimensional smooth parametrization of the prior in the limited data regime.* Error bars denote ±1 standard error (SE). **(a)** *Improved accuracy.* We simulated data using the same specifications as in Fig. 2b, and fit the prior for several values of *df*. Notice that when data are limited, using a low-dimensional parametrization recovers the prior more accurately, but as the number of data points increases, using the full multinomial ($df = G(n) - 1$) does best. **(b)** *Better generalization.* We used 1210 data points from a single social cover story on 5 nodes. We randomly split them into test (698 data points) and training data, and fit the prior for several values of *df*. Notice that, in accord with the bias-variance tradeoff, using a low-dimensional parameterization for the prior results in better generalization (higher log-likelihood of the unseen data) when data are scarce, but as the number of data points increases, using the full multinomial does best.

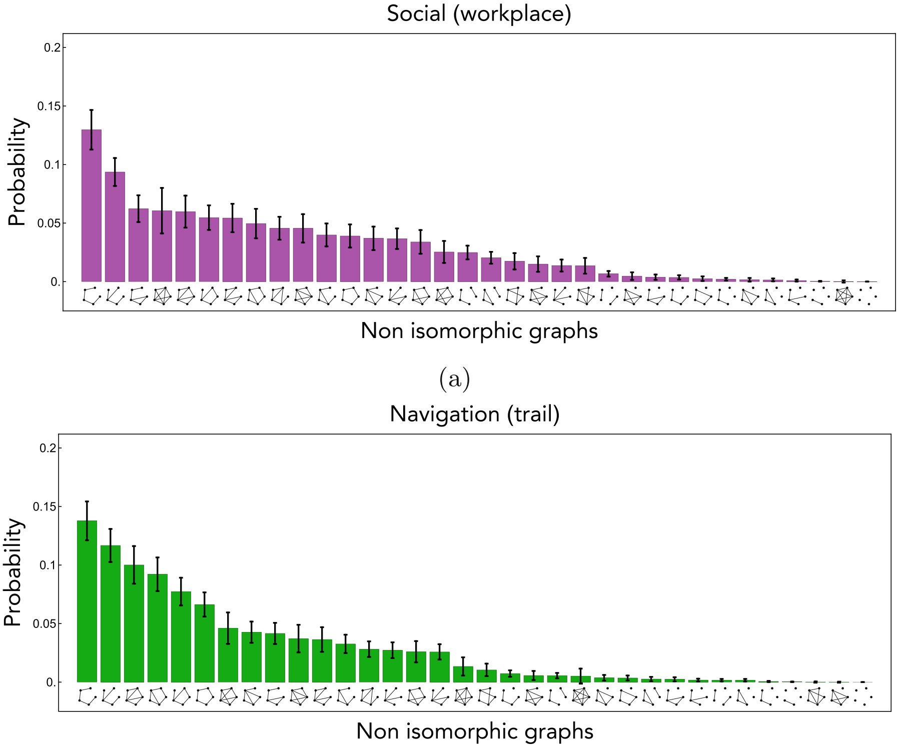
$p \in (0, 1)$. "Smoother" priors, parameterized by the number of degrees of freedom $1 \leq df \leq G(n) - 1$, are obtained by including only the longer-decaying modes (i.e., $c_{i>df+1} = 0$). In Fig. 3a, we show the effect of df when applying our model to simulated data. When all the graphs are sufficiently sampled, $df = G(n) - 1$ (equivalent to a full multinomial model) recovers the prior more accurately. In the limited data regime, however, using fewer coefficients does better, as it avoids overfitting. Figure 3b illustrates the associated improved generalization using human data.

3.3 Subjects' Priors Have Non-trivial Graphical Structure

We compared several models for the priors using leave-p-out cross-validation (CV) [1] on data from our online experiments.[6] In particular, we considered several choices of df for our smooth parametrization of the prior, a full multinomial prior, and two "null models" (in the sense that they are not sensitive to graphical structure): *(1)* Erdős-Rényi (ER) prior, where the edges probabilities are independent and identical random variables; and *(2)* Prior over average degree, where the edge probabilities are no longer independent but are completely exchangeable (equivalent to only counting the *number* of edges).

For all priors we considered (different cover stories and numbers of nodes), a smooth parametrization of the prior did better (higher average log-likelihood in

[6] Our experiments use graphs on 4 to 10 nodes.

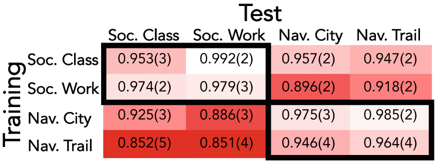


(a)

(b)

Fig. 4. *Deviations from basic null models suggests that the priors have non-trivial graphical structure.* To generate error bars, we simulated our Bayesian MCMCP model using the same set of partial graphs and the prior fitted to human data, and then fit our model to these simulated data. The error bars denote ±1 standard deviation from the mean fitted prior in the simulated data. **(a)** *Social:* Using 481 data points of a cover story of inferring the friendship network in a workplace with 5 people, we performed leave-10-out cross validation (CV). The figure displays the fit with the highest CV score (defined as the average log-likelihood per trial in the test sets). This fitted prior (31 *df*, CV score: -2.52, SE: 0.062) deviates significantly (t-test p-value: $< .0001$) from both null models (ER model, CV score: -2.79, SE: 0.043; and prior over average degree, CV score: -2.71, SE: 0.056). **(b)** *Navigation:* We performed the same analysis on a dataset (537 data points) of a cover story of inferring the trail map of a nature park with 5 sights. As before, the best fitted prior has 31 *df* (CV score: -2.41, SE: 0.066), and deviates significantly from both null models (ER model, CV score: -2.79, SE: 0.049; and prior over average degree, CV score: -2.71, SE: 0.061).

the CV test sets) than the two null models. Additionally, in all cases, the best fit had a relatively high number of *df*, suggesting that there is non-trivial graphical structure in subjects' priors. Figure 4 displays results for priors over 5 nodes. Moreover, as illustrated in Fig. 5 when comparing subjects' priors over different



Training \ Test	Soc. Class	Soc. Work	Nav. City	Nav. Trail
Soc. Class	0.953(3)	0.992(2)	0.957(2)	0.947(2)
Soc. Work	0.974(2)	0.979(3)	0.896(2)	0.918(2)
Nav. City	0.925(3)	0.886(3)	0.975(3)	0.985(2)
Nav. Trail	0.852(5)	0.851(4)	0.946(4)	0.964(4)

Fig. 5. *Domain-dependent priors.* We compared the priors of 4 experiments with different cover stories in 5 nodes by randomly splitting these 4 datasets in training (382 data points) and test (96 data points) data, and fitting a full multinomial model to each training set individually as well as to the aggregated 4 training sets. For each test set, we calculated the likelihood ratio (LR) between the individual fits and the aggregated fit. Notice that the LR is larger when the training and test data share the same domain (as indicate by the block diagonal structure of the table). Moreover, the fact that all LR are smaller than 1 indicates that we still need to collect more data.

cover stories on the same number of nodes, the priors between the navigation and social domains were notably different, while showing fewer differences between different contexts in the same domain (navigation: city and nature park; social: coworkers and students). This raises the interesting possibility that priors over task graphs are sensitive to the more abstract structure of a task (domain) rather than its specific context, which would allow for broader generalization.

3.4 General Framework for Quantifying Priors over Exchangeable Random Structures

Finally, we extend our results to the more general case of quantifying priors over exchangeable random structures [14], where the partial data are generated by randomly obscuring a given fraction of the sequence. The relevant parameters are: $\mathcal{A}$, the alphabet; ℓ, the string length; m, the number of relations obscured; and $\mathcal{G}$, the group under which the sequence is exchangeable. For example, for unordered binary strings, $\mathcal{A} = \{0, 1\}$ and $\mathcal{G}$ is the *full* permutation group $\mathcal{S}_\ell$ acting on the entries of strings of length ℓ; for simple graphs, the binary string is length $\ell = \binom{n}{2}$ and $\mathcal{G}$ is the permutation group $\mathcal{S}_n$, where n is the number of *nodes*. An element in $\mathcal{G}$ induces a permutation of the indices $\{1, \ldots, \ell\}$, and thus a permutation of the elements in $\mathcal{A}^\ell$. This action of $\mathcal{G}$ induces an equivalence relation on $\mathcal{A}^\ell$ ($x \sim y$ if $\exists g \in \mathcal{G}$ s.t. $x = g.y$), partitioning it into equivalence classes. For example, for unordered binary strings, they are partitioned into $\ell + 1$ sets, one for each possible sum $(0, ..., \ell)$; for simple graphs, the partitions correspond to the non-isomorphic graphs on n nodes. The condition of exchangeability under $\mathcal{G}$ means that probabilities are assigned to these partitions, with elements in the same partition having equal probability.

4 Discussion

In this work, we develop a formal framework to quantify humans' priors over exchangeable random structures, and apply it to the case of non-isomorphic graphs representing the structure of several navigation and social tasks (instantiated using our new online experimental platform). We believe that navigation and social interaction tasks are a good starting point as they are an integral part of humans' lives and their different structure can provide interesting comparisons. Although we are currently collecting more data to improve our statistics, and doing further detailed analysis of the data (e.g., modeling priors over larger number of nodes, and building generative models that explains subjects' priors), the results reported here are encouraging, in the sense that there appears to be non-trivial domain-dependent structure in subjects' priors.

It is important to highlight that our model makes several assumptions about subjects' behavior that could not hold in practice (e.g., that they are "Bayesian"), and that are hard to test empirically (due, e.g., to interactions among the effects of the different assumptions). We argue that for our method to shed light on our understanding of our priors over tasks structure, it is more important to establish whether the priors we obtain from our experiments are in fact "meaningful" (in the sense of being used in practice) than whether the data violate certain assumptions or not.[7] Hence, to test the behavioral relevance of our measured priors, we are beginning to test if subjects learn faster/have better performance in experimental tasks (distinct from our MCMCP experiments) with structures consistent with these priors. Additionally, we are working on developing principled ways to test if analogous real-world datasets have structure similar to the priors.

In sensory and motor neuroscience, quantifying humans' priors over these systems has led to significant insights [7,8,12,13]. For example, several visual illusions can be understood as resulting from priors that encode the structure of naturalistic scenes [10,11]. Analogously, understanding our beliefs about the structure of new tasks could lead to a deeper understanding of our learning and generalization abilities (as well as their failure modes) [2,4]. We hope that the approach we proposed will pave the way towards the more ambitious goals of formalizing what a general "task" is, and of providing unifying principles that explain what makes a given task easy or hard for a human to solve.

References

1. Arlot, S., Alan, C.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010). https://doi.org/10.1214/09-SS054
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
3. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton (1957)

[7] Science is full of examples of reduced models (thermodynamics, effective field theories, magnetohydrodynamics, etc.) that provide powerful predictions even when applied to systems that violate their assumptions.

4. Botvinick, M., Weinstein, A., Solway, A., Barto, A.: Reinforcement learning, efficient coding, and the statistics of natural tasks. Curr. Opin. Behav. Sci. **5**, 71–77 (2015). https://doi.org/10.1016/j.cobeha.2015.08.009
5. Brady, T.F., Konkle, T., Alvarez, G.A.: Compression in visual working memory: using statistical regularities to form more efficient memory representations. J. Exp. Psychol. Gen. **138**, 487–502 (2009). https://doi.org/10.1037/a0016797
6. Canini, K.R., Griffiths, T.L., Vanpaemel, W., Kalish, M.L.: Revealing human inductive biases for category learning by simulating cultural transmission. Psychon. Bull. Rev. **21**, 785–793 (2014). https://doi.org/10.3758/s13423-013-0556-3
7. Field, D.F.: What the statistics of natural images tell us about visual coding. Proc. SPIE Int. Soc. Opt. Eng. **1077**, 269–276 (1989). https://doi.org/10.1117/12.952724
8. Graziano, M.S.A.: Cortical action representations. In: Toga, A.W., Poldrack, R.A. (eds.) Brain Mapping: An Encyclopedic Reference. Elsevier, Amsterdam (2014). http://www.princeton.edu/~graziano/Graziano_encyclopedia_2015.pdf
9. Griffiths, T., Kalish, M.: Language evolution by iterated learning with Bayesian agents. Cogn. Sci. **31**, 441–480 (2007). https://doi.org/10.1080/15326900701326576
10. Howe, C.Q., Purves, D.: The Müller-Lyer illusion explained by the statistics of image source relationships. Proc. Natl. Acad. Sci. **102**(4), 1234–1239 (2005). https://doi.org/10.1073/pnas.0409314102
11. Howe, C.Q., Yang, Z., Purves, D.: The Poggendorff illusion explained by natural scene geometry. Proc. Natl. Acad. Sci. **102**(21), 7707–7712 (2005). https://doi.org/10.1073/pnas.0502893102
12. Lewicki, M.S.: Efficient coding of natural sounds. Nat. Neurosci. **5**(4), 356–363 (2002). https://doi.org/10.1038/nn831
13. Orbán, G., Fiser, J., Aslin, R.N., Lengyel, M.: Bayesian learning of visual chunks by human observers. Proc. Natl. Acad. Sci. **105**(7), 2745–2750 (2008). https://doi.org/10.1073/pnas.0708424105
14. Orbanz, P., Roy, D.M.: Bayesian models of graphs, arrays and other exchangeable random structures. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 437–461 (2015). https://doi.org/10.1109/TPAMI.2014.2334607
15. The on-line encyclopedia of integer sequences (OEIS). https://oeis.org/A000088