Approximating Bayesian inference
with a sparse distributed memory system

Joshua T. Abbott   Jessica B. Hamrick   Thomas L. Griffiths
{ joshua.abbott , jhamrick , tom_griffiths }@berkeley.edu

Overview

Probabilistic models of cognition have had success in explaining complex inductive inferences that people make in everyday life.

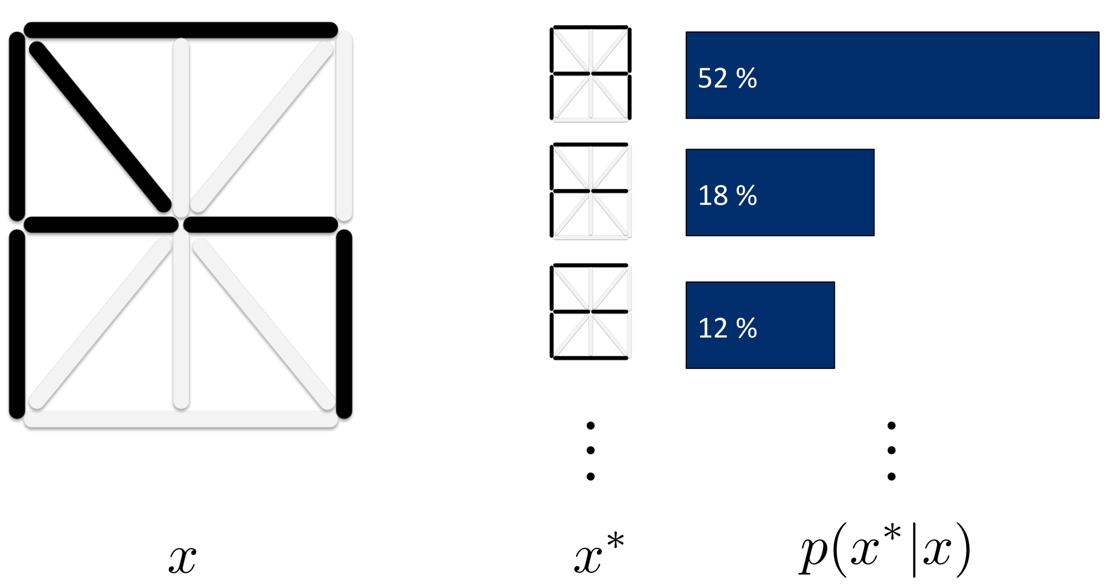
However, these models involve difficult computations over structured representations - raising criticism as to whether human brains do anything like this.

We show that an associative memory system using sparse distributed representations can be interpreted as a Monte Carlo method of approximating Bayesian inference.

This provides an account of how structured representations assumed by probabilistic models could be expressed in the distributed, continuous representations used by connectionist models.

Background

Observe: noisy stimulus   **Infer:** true uncorrupted stimulus



Can recover true stimulus using Bayes' rule to compute the *posterior distribution* over x^*

$$p(x^*|x) = \frac{p(x|x^*)p(x^*)}{\int_{x^*} p(x|x^*)p(x^*)dx^*}$$

Many probabilistic computations can be found as the expectation of a function $f(x^*)$ given x

$$E[f(x^*)|x] = \int f(x^*)p(x^*|x)\,dx^*$$

This expectation can be approximated using a Monte Carlo method known as *importance sampling.*

$$E[f(x^*)|x] \approx \frac{1}{K}\sum_{k=1}^{K} f(x_k^*)\frac{p(x_k^*|x)}{p(x_k^*)}$$

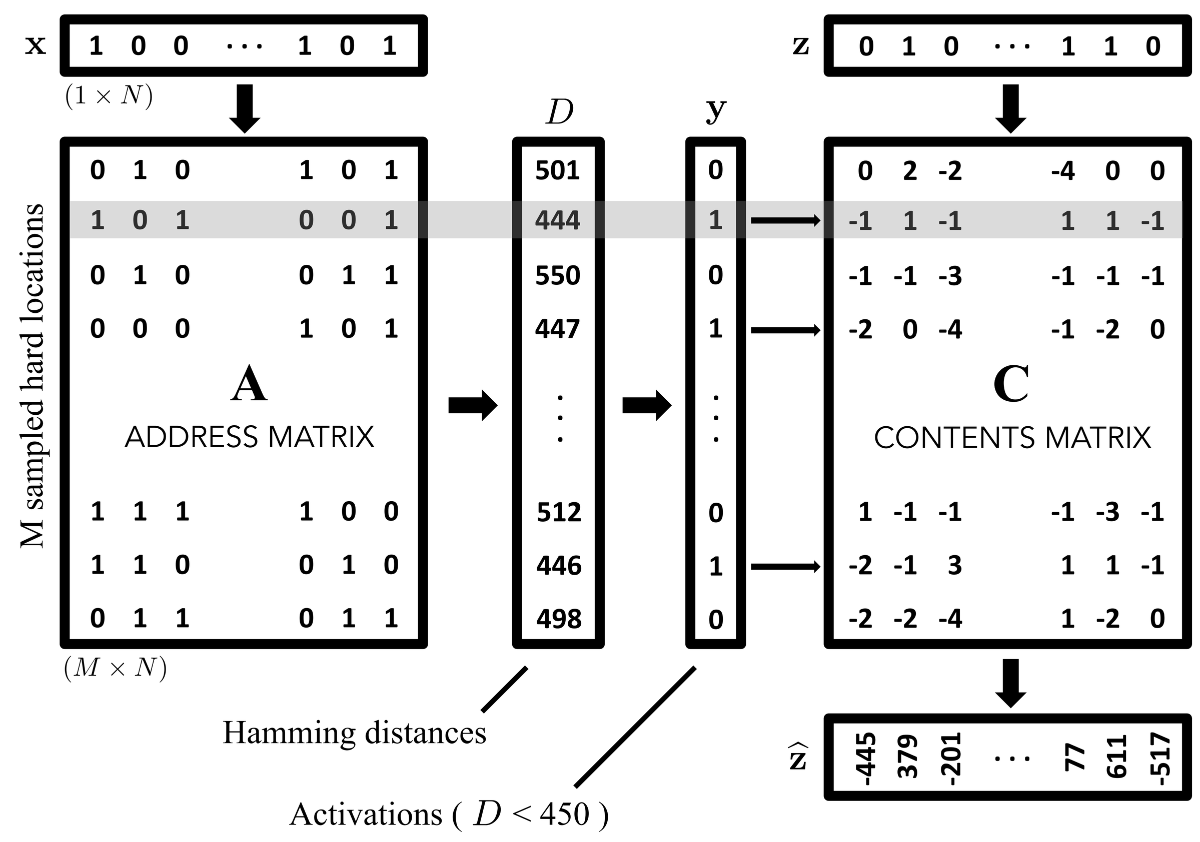
$$= \frac{1}{K}\sum_{k=1}^{K} f(x_k^*)\frac{p(x|x_k^*)p(x_k^*)}{p(x_k^*)p(x)}$$

$$= \alpha(x)\sum_{k=1}^{K} f(x_k^*)p(x|x_k^*)$$

Essentially, the *posterior* distribution can be approximated using a set of K samples $\{x_k^*\}$ drawn from the *prior*, $p(x^*)$, weighted by the *likelihood*.

Previous work has shown that *importance sampling* can be implemented in a common psychological process model – an exemplar model [1].

Sparse Distributed Memory



Writing: $\mathbf{z}$ is written to memory address $\mathbf{x}^*$ by manipulating the contents of all sampled addresses $\mathbf{a}_j$ within a Hamming distance D of $\mathbf{x}^*$

$$\mathbf{y} = \Theta_D(\mathbf{A}\,\mathbf{x}^*)$$
$$\mathbf{C} = \mathbf{C} + \mathbf{z}\,\mathbf{y}$$

Reading: $\hat{\mathbf{z}}$ is read from memory address $\mathbf{x}$ by summing the contents of all sampled addresses $\mathbf{a}_j$ within a Hamming distance D of $\mathbf{x}$

$$\mathbf{y} = \Theta_D(\mathbf{A}\,\mathbf{x})$$
$$\hat{\mathbf{z}} = \mathbf{C}^T\,\mathbf{y}$$

Main idea: use an SDM to store and retrieve exemplars, allowing us to build on the equivalence between exemplar models and importance samplers.

SDMs as importance samplers

Writing: Let $w(\mathbf{a}_j, \mathbf{x}^*)$ be the probability of writing to address $\mathbf{a}_j$ given an input address $\mathbf{x}^*$, and satisfying the constraint:

$$\lim_{M\to 2^N} w(\mathbf{a}_j, \mathbf{x}^*) = \prod_{i=1}^{N}\delta(x_i^* - a_{ji})$$

After writing K (address, data) pairs $(\mathbf{x}_k^*, \mathbf{z}_k)$, the value of the counter associated with bit i of address $\mathbf{a}_j$ is:

$$\mathbf{c}_j = \sum_{k=1}^{K} w(\mathbf{a}_j, \mathbf{x}_k^*)\,\mathbf{z}_k$$

Reading: Let $r(\mathbf{x}, \mathbf{a}_j)$ be the probability that we read from address $\mathbf{a}_j$ given input $\mathbf{x}$. The output of the SDM for a particular set of addresses $\mathbf{A}$ is:

$$\hat{\mathbf{z}} = \sum_{j=1}^{M}\mathbf{c}_j\; r(\mathbf{x}, \mathbf{a}_j)$$

To see how this output behaves for *any* SDM, we consider the expected value of $\hat{\mathbf{z}}$ over sampled sets of addresses $\mathbf{A}$:

$$E_{\mathbf{A}}[\hat{\mathbf{z}}|\mathbf{x}] = E_{\mathbf{A}}\left[\sum_{j=1}^{M}\left(\sum_{k=1}^{K} w(\mathbf{a}_j, \mathbf{x}_k^*)\,\mathbf{z}_k\right) r(\mathbf{x}, \mathbf{a}_j)\right]$$

$$= \sum_{k=1}^{K}\mathbf{z}_k \cdot E_{\mathbf{A}}\left[\sum_{j=1}^{M} w(\mathbf{a}_j, \mathbf{x}_k^*)\,r(\mathbf{x}, \mathbf{a}_j)\right]$$

In the limit, as our address space grows larger, the expected value of $\hat{\mathbf{z}}$ read from the SDM will be:

$$\lim_{M\to 2^N} E_{\mathbf{A}}[\hat{\mathbf{z}}|\mathbf{x}] = \sum_{k=1}^{K}\mathbf{z}_k\int_{\mathbf{a}_j}\delta(\mathbf{x}_k^* - \mathbf{a}_j)\,r(\mathbf{x}, \mathbf{a}_j)\;\mathrm{d}\mathbf{a}_j$$

$$E_{\mathbf{A}}[\hat{\mathbf{z}}|\mathbf{x}] = \sum_{k=1}^{K}\mathbf{z}_k\,r(\mathbf{x}, \mathbf{x}_k^*)$$

Simulations

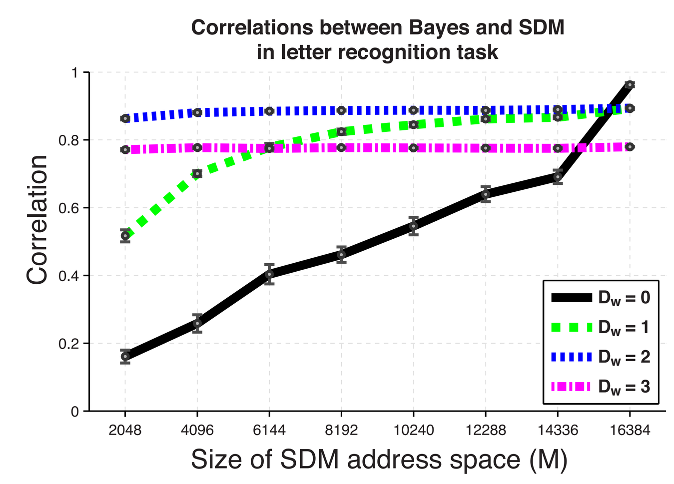
Letter Recognition Task
(Bayesian likelihood matches SDM read rule)

Bayesian model: we set the prior, $p(\mathbf{x}^*)$, to be proportional the relative letter frequency of English text. The likelihood is defined to be uniformly distributed over the number of possible bit strings in a hypersphere of radius D_r

$$p(\mathbf{x}|\mathbf{x}^*) \approx r(\mathbf{x}, \mathbf{x}^*) = \begin{cases}\left[\sum_{d=1}^{D_r}\binom{N-d+1}{d}\right]^{-1} & |\mathbf{x} - \mathbf{x}^*| \le D_r \\ 0 & \text{otherwise}\end{cases}$$

SDM model: we sample exemplars $\mathbf{x}^*$ of the original letters from the prior, $p(\mathbf{x}^*)$, and write them at inputs $\mathbf{z} = \mathbf{x}^*$

Results: Average correlations between Bayesian posterior mean and SDM output (D_r = 2)



Property Induction Task
(SDM read rule matches Bayesian likelihood)

Bayesian model: we observe a set of examples $\mathbf{d}$ of a concept C (known to have property K), and we aim to calculate the probability a novel object y is also a member of C

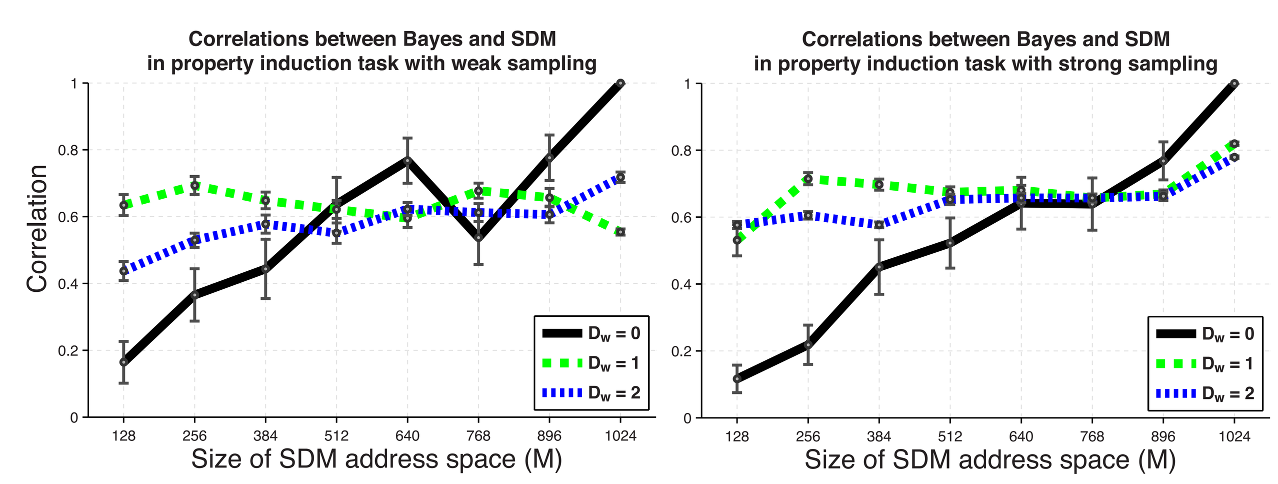
$$p(y \in C|\mathbf{d}) = \sum_{h\in\mathcal{H}} p(y \in C|h)p(h|\mathbf{d})$$

weak sampling:
$$p(\mathbf{d}|h) = \begin{cases}1 & \{d_n\} \subset h \\ 0 & \text{otherwise}\end{cases}$$

strong sampling:
$$p(\mathbf{d}|h) = \begin{cases}1/|h|^n & \{d_n\} \subset h \\ 0 & \text{otherwise}\end{cases}$$

SDM model: location $\mathbf{x}^*$ corresponds to a hypothesis h, and the content $\mathbf{z}$ corresponds to data $\mathbf{d}$. For both weak and strong sampling, we set the SDM read rule $r(\mathbf{d}, h)$ by weighting the selection vector $\mathbf{y}$ by $p(\mathbf{d}|h)$

Results: Average correlations between Bayesian posterior and SDM output (D_r = 0)



Conclusions

We demonstrated that an associative memory system using sparse, distributed representations to store and retrieve exemplars can approximate Bayesian inference.

The critical advance is that this is done using distributed representations: arbitrary hypotheses can be represented, and arbitrary distributions of exemplars encoded by a single architecture.

References:
1. Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, 17(4), 443–464.
2. Kanerva, P. (1988). *Sparse Distributed Memory*. Cambridge, MA: MIT Press.

Acknowledgments: This work was supported by grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research and a Berkeley Fellowship awarded to JBH.