Explorations in Predictive Coding Models of Hallucinations

Milica Kiš, Martin Takáč

FMFI, UNIBA, Bratislava
Fakulta matematiky, fyziky a informatiky
Univerzity Komenského
Mlynská dolina F1
842 48 Bratislava
Email: milica.kis@fmph.uniba.sk

Abstract

In this paper, we explore the emergence of visual hallucinations in a hierarchical predictive coding model trained on the MNIST digit dataset. Adding precision to the model enables exploration of the role of priors in hallucinations. Forcing sensory precision to low values corresponds to regarding the sensory data as unreliable. Conversely, precision in the upper layers driven to disproportionately high values yields strong priors, corresponding to rigid beliefs characteristic of psychiatric disorders. The aberrant precision weighting is taken as the mechanism underlying the hallucinations. Hallucinations can be induced either by clamping the top layer to a specific digit or by letting the system evolve its internal landscape from an unconstrained internal state. The latter would correspond to spontaneous hallucinations in the absence of sensory input, while the former would result in the generative model maintaining a strong belief despite the sensory evidence.

1 Introduction

Recently, predictive coding has shown promise as an explanatory and exploratory framework for phenomena in computational psychiatry. This influential theory in neuroscience (Friston and Kiebel, 2009) has been used to explain various cognitive phenomena, including those in the domain of human vision.

In predictive coding, processing of neural signals approximates Bayesian inference on the latent causes of sensory data. Observations (likelihoods) are weighed against expectations (priors) to arrive at better beliefs (posterior probabilities or posteriors). Prediction errors signal the discrepancy between likelihoods and priors, which guides adaptive belief updating. The goal of this process is to minimize the free energy. The theory assumes a hierarchical brain structure and bidirectional transmission of neural messages (bottom-up and top-down). From the standpoint of computational psychiatry, positive psychotic symptoms, such as hallucinations and delusions, may be seen as profound disruptions in forming, maintaining, and updating our beliefs about the world.

In this paper, we focus on visual hallucinations of indeterminate etiology.[1] In a general sense, hallucinations signify percepts arising without the corresponding external stimulus. The other most common type is auditory (verbal) hallucinations (AVH), i.e., hearing voices (Corlett et al., 2019).

In the current literature, there are two opposing hypotheses about the nature of priors associated with hallucinations (Corlett et al., 2019; Sterzer et al., 2018). The strong prior hypothesis states that hallucinations result from enhanced top-down predictive signaling (i.e., increased prior precision) relative to sensory-level activity. Therefore, perception relies more on beliefs than on the sensory input. Conversely, a weak prior account (i.e., lower precision) posits relying more on the ambiguous or noisy input from the sensory level, exerting a disproportionate influence on perception. However, these two hypotheses are not mutually exclusive and may operate simultaneously in different conditions.

The additional complexity is that different priors may be associated with different sensory modalities and hierarchical levels of brain organization. Higher-level associative cortical regions exhibit a higher density of recurrent connections than lower sensory regions, with implications for excitation-inhibition dynamics (Corlett et al., 2019). Consequently, priors that are active at different levels have implications for precision dynamics.

2 Hierarchical predictive coding (HPC) network

To test specific hypotheses about the emergence of hallucinatory-like behavior in the visual system arising from aberrant precision weighting, we introduce a hierarchical predictive coding (HPC) architecture based on the model by Tscshantz et al. (2023), with modifications and simplifications. The current model disregards

[1] Apart from psychiatric disorders, such as schizophrenia and post-traumatic stress disorder, hallucinations may be associated with neurological conditions, such as Alzheimer's disease or epilepsy, eye conditions, or may even appear or be induced in the healthy population (Corlett et al. (2019).

the amortized feedforward route, but adds dynamic precision updating at different hierarchical levels.

2.1 Network architecture

We implemented a 4-layer hierarchical PC model with interpretable inference dynamics to replicate the normal activity in the visual cortex (with 4 layers loosely interpreted as corresponding to areas V1, V2, V4, and IT).[2] The lowest, sensory input layer consists of 784 neurons (corresponding to 28 x 28 pixels input MNIST digit images). The second and third layers contain 256 and 64 neurons, respectively, while the top (output) layer has 10 neurons, corresponding to the MNIST digit identity (classes). The reduction from 256 to 64 units implements a dimensional bottleneck consistent with increasing abstraction along the ventral visual stream.

2.2 Initialization

Generative (top-down) weights are initialized by sampling from a normal distribution (due to the theoretical Gaussian assumptions). Precisions are initiated as vectors of ones.

Clamping. The highest-level latent representation is clamped to the target label (a one-hot vector) during training, thereby implementing supervised learning.

2.3 Network performance metrics

Under similar (Gaussian) assumptions, Mean Square Error (MSE, summed across all layers) serves as a proxy for minimizing the variational free energy across epochs. Reconstruction quality is estimated from the lowest, sensory layer. It is presented as the percentage of drop in MSE relative to the first epoch (assuming that the energy is minimized). Reconstructions allow visual comparison with the original MNIST digits (Fig. 1).

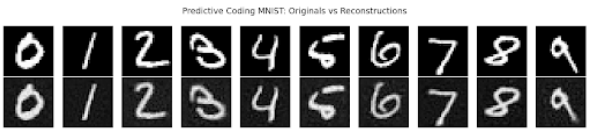

Fig. 1: Original MNIST digits vs. reconstructions (generative model).

2.3.1 Hyperparameters

The following hyperparameters were used: (1) **inference rate** (κ), controlling the step size of latent state updates, a fast component, (2) **learning rate for weights** (η), typically set to a small value, e.g. $\eta = 0.001$, a slow learning component, (3) **number of inference steps** (typically 30), (4) **number of training epochs** (typically 100), (5) **learning rate for precisions** (K) (typically 0.005).

[2] This mapping is conceptual and does not reflect strict anatomical correspondence.

2.3.2 Precisions in PC

In predictive coding, precisions (π) reflect the brain's confidence in the prediction errors at each hierarchical level. A high precision means the system strongly trusts the error signal from that layer, and thus assigns more weight to that error when updating its beliefs (latent states). Conversely, low precision would decrease the weighting of the prediction error. Precision is formalized as the inverse of variance or negative entropy (Corlett et al., 2019). Changes in precision are implemented biologically via neuromodulators like dopamine, noradrenaline, or acetylcholine (Feldman and Friston, 2010). Precision is thought to be encoded by the postsynaptic gain of neurons reporting the prediction error. The gain control adjusts the weighting of precision errors. Synaptic gain is a fundamental biophysical property of neuronal function, and it describes the sensitivity of a postsynaptic neuron's membrane potential to variations of strength of the presynaptic input (Nour and Dolan, 2022). Relationship between precision, priors and posteriors is presented in Fig. 2. Posterior belief is shifted toward the prior or sensory evidence (likelihood) in proportion to their relative precision (as the inverse of variance).

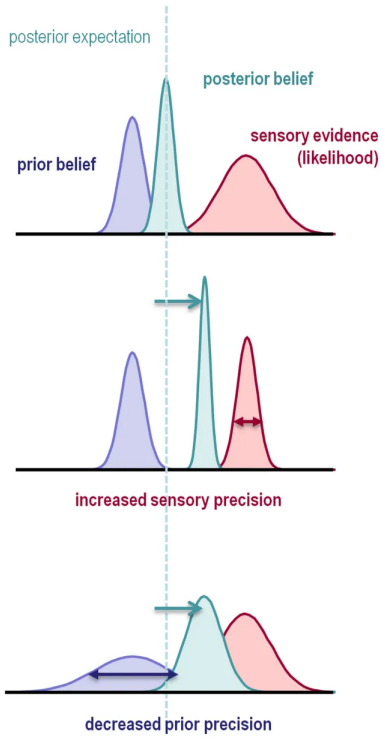

Fig. 2: Relationship between precision, priors and posteriors. Adapted from Adams et al. (2013).

Dynamic precision updating. The precision, being the inverse of the variance $\pi = \frac{1}{\sigma^2}$, can be iteratively updated at each level l using the formula:

$$\Delta\pi_l = -K * \left(\epsilon_l^2 - \frac{1}{\pi_l}\right) , \quad (1)$$

where the squared error signal ϵ_l^2 (the difference between expectation and the bottom-up signal at level l) is conveniently taken to represent the variance. All precisions are initiated as vectors of ones. This formula computes a sliding estimate of the inverse variance of each component of the error vector. Precision increases when errors are small and decreases when errors are large. The order of updates is also important. First, the inference updates settle latent states, then the precisions adapt to current uncertainty, and, finally, the weights learn from precision-weighted errors.

3 Precision experiments

Replicating hallucinatory-like effects by manipulating precisions.

3.1 Manipulating precisions across layers

Adding precision to the model enables exploration of the role of priors in hallucinations. Forcing sensory precision to low values corresponds to regarding the sensory data as unreliable. Conversely, precision in the upper layers driven to disproportionately high values yields strong priors, corresponding to rigid beliefs characteristic of psychiatric disorders. We created an imbalanced hierarchy as a way to implement aberrant precision weighting, which is taken as the mechanism underlying hallucinations. Hallucinations (or emerging hallucinatory-like structures) can be induced either by clamping the top layer to a specific digit or by letting the system explore its unconstrained internal landscape. The former would result in the generative model maintaining a strong belief despite the sensory evidence (blank input in this case, see Fig. 3), while the latter would correspond to spontaneous hallucinations in the absence of sensory input (see Fig. 4). The rationale behind this choice is that hallucinations (including visual) are usually defined as the false sensory perception in the absence of external sensory stimuli.

This model enables testing different hypotheses within the same architecture (using the same inference dynamics). Precision dynamics can be shifted by assigning different values to various layers, sensory input could be either blank, regular or corrupted or noisy, while clamping enables constraining or freeing the inference dynamics.

3.2 Two hallucination regimes

3.2.1 Top-down hallucinations (strong priors)

In this regime, we are clamping the top layer to a specific digit. The network is generating data from internal strong priors. We are observing increase in free energy (negative values of the recorded drop in free energy). The values are presented in Tab. 1. For comparison, drop in free energy and MSE during regular inference is 95.79%, and 98.31%, respectively.

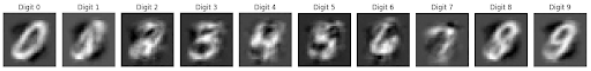

Fig. 3: Hallucinated digits from strong priors

$Hallucination$	$FEdrop$
1	-5168.15%
2	-5549.41%
3	-5795.73%
4	-5873.69%
5	-5299.63%

Tab. 1: Negative free energy drop (i.e., increase) in strong prior condition.

3.2.2 Spontaneous hallucinations

Network samples its own generative model in five iterations in unconstrained generative regime, in the absence of sensory input. We also observe the negative drop of free energy, i.e., the increase. This is expected because the network is generating an image (i.e., a digit) from a blank input, leading to a large prediction error and thus an increase in free energy. The values are presented in Tab. 2.

Fig. 4: Spontaneous hallucinations

$Hallucination$	$FEdrop$
1	-9609.93%
2	-9259.43%
3	-9333.59%
4	-9797.32%
5	-9978.10%

Tab. 2: Negative free energy drop (i.e., increase) in spontaneous condition.

Free energy increases more strongly in the spontaneous condition, reflecting the absence of constraints

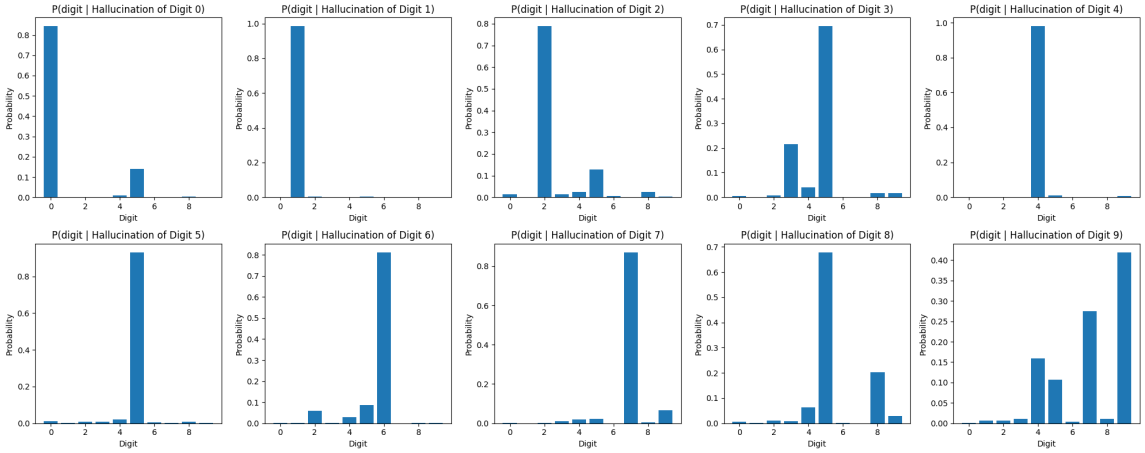

Fig. 5: Probability distributions in strong prior top-down condition

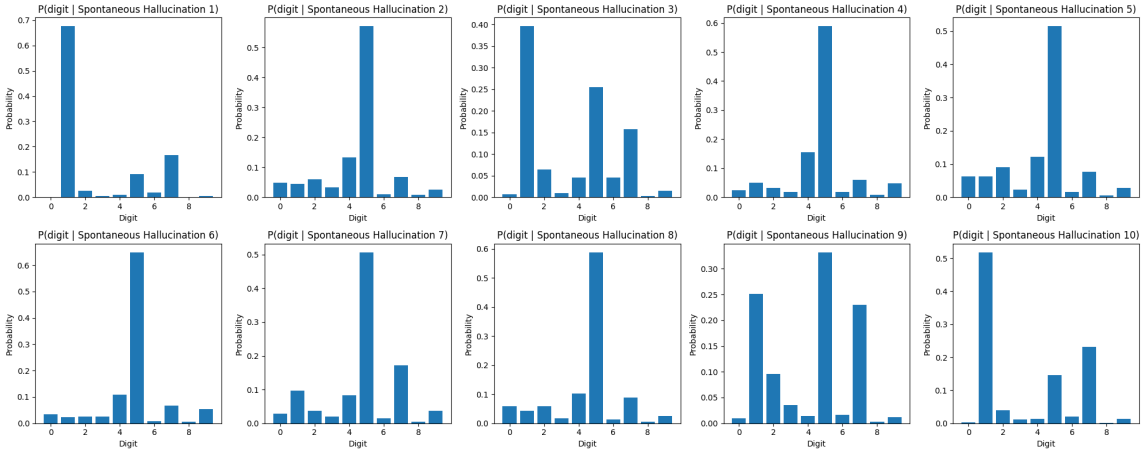

Fig. 6: Probability distributions in spontaneous condition

and the need to resolve large internally generated prediction errors during convergence.

4 Logistic regression for estimating the classification confidence

In order not to be relying solely on the observed visual similarity with the original MNIST digits, we used Logistic Regression for estimating the classification confidence (i.e., probability distribution of digits given the hallucinated image). Classification confidence was higher in the case of top-down strong prior hallucinations (with the exception of digits 3, 8, and 9) (Fig. 5), while for spontaneous hallucinations probabilities were more distributed, which might suggest that spontaneous generation is more ambiguous, and that the system is trying to explore various internal landscapes (Fig. 6)). The resulting ambiguity for digits 3, 8, and 9 might lead to speculation about the perceived similarities of rounded and semi-rounded shapes in these figures, which also might have been encoded in their latent representations.

5 Conclusion

Exploring hallucinatory-like behavior in a hierarchical predictive coding model approximating human vision offers a glimpse into internal dynamics of aberrant precision-weighting, which is thought to underlie many psychiatric conditions. Although these deficits may be tested in experimental conditions (mismatch negativity, sensory attenuation), computational modeling offers an alternative method of testing these hypotheses.

Future directions. Another possible direction in our research is testing the hypothesis of weak priors. Under these assumptions, hallucinations arise from insufficient top-down constraints on perception, such that noisy or ambiguous sensory input is not adequately regularized by prior expectations. In the model, we could achieve this by reducing the precisions of higher hierarchical layers (which would correspond to a weak prior) and providing a noisy or corrupted input, allowing sensory ambiguities to drive inference.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K. and Powers, A. R. (2019). Hallucinations and strong priors. *Trends in cognitive sciences*, 23(2):114–127.

Feldman, H. and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221.

Mancini, V. and Nour, M. M. (2025). If mismatch negativity is the answer, what is the question? on the nature of predictive coding abnormalities in psychosis. *Biological Psychiatry Global Open Science*, 5(1):100412.

Nour, M. M. and Dolan, R. J. (2022). Synaptic gain abnormalities in schizophrenia and the potential relevance for cognition. *Biological psychiatry*, 91(2):167–169.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M. and Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological psychiatry*, 84(9):634–643.

Tscshantz, A., Millidge, B., Seth, A. K. and Buckley, C. L. (2023). Hybrid predictive coding: Inferring, fast and slow. *PLoS computational biology*, 19(8):e1011280.