An Information-Theoretical Approach To Optimizing Task Design For Differentiating Probabilistic Neural Codes

Po-Chen Kuo
University of Washington
pckuo@uw.edu

Edgar Y. Walker
University of Washington
eywalker@uw.edu

Abstract

Bayesian brain hypothesis has been among the leading theories in modeling perceptual decision-making under uncertainty. While many psychophysical studies have provided evidence in support of the brain performing Bayesian computation, how uncertainty information is encoded in sensory neural populations has remained elusive. Specifically, two competing hypotheses propose that early sensory populations encode either the likelihood function (exemplified by probabilistic population codes) or the posterior distribution (exemplified by neural sampling codes) over the stimulus, with the critical distinction being whether stimulus priors would modulate early sensory neural responses. However, differentiating the two probabilistic neural codes experimentally remains challenging, as it is unclear what task design would effectively distinguish the two hypotheses. In this work, we develop an information-theoretical approach to optimizing task stimulus distribution that would best differentiate competing probabilistic neural representations. Our method derives an *information gap*—the expected performance difference between likelihood and posterior decoders applied to sensory population responses following a specific probabilistic neural code—by measuring the KL divergence between true posterior distributions and surrogate posterior distributions utilizing Bayes-optimal estimators for a given task design. On simulated neural populations, we demonstrate that our information-gap measure accurately predicts decoder performance differences across a wide array of settings. Crucially, maximizing the information gap yields stimulus distributions that optimally differentiate likelihood and posterior coding hypotheses. Our framework enables principled, theory-driven experimental design for differentiating probabilistic neural codes, advancing our understanding of how neural populations represent and process sensory uncertainty.

1 Introduction and related work

Perceptual decision-making requires organisms to represent and process sensory information in the face of noisy and ambiguous sensory inputs, thus correctly accounting for the uncertainty associated with observations. The Bayesian brain hypothesis—with theoretical roots tracing to Laplace [1] and von Helmholtz [2]—proposes that the brain maintains internal generative models of the world and performs inference by computing probability distributions over task-relevant latent world states [3, 4]. This framework has proven successful in explaining various aspects of human and animal perception, from multisensory integration to object recognition and motion perception [5, 6, 7]. Extensive behavioral evidence demonstrates that humans and animals perform near optimally in perceptual tasks that require uncertainty estimation [6, 8, 9], strongly suggesting that sensory neural populations encode both task-relevant stimulus features and their associated uncertainty. However,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

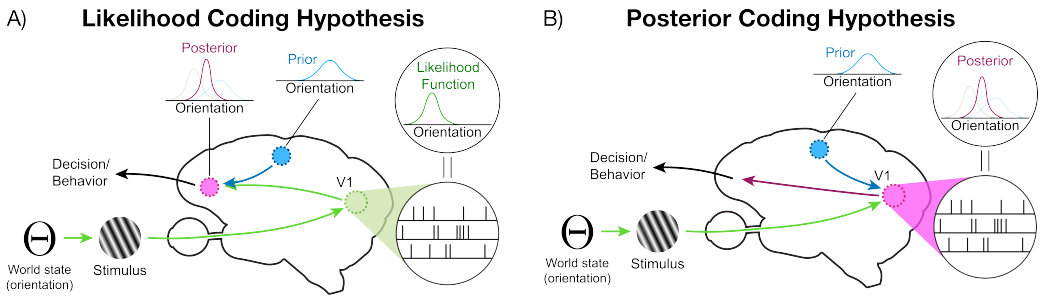

Figure 1: **Two competing hypotheses on how sensory uncertainty information is encoded in early sensory neural populations.** A) Likelihood coding hypothesis (exemplified by the probabilistic population code [15]) proposes that early sensory populations encode the likelihood function over the stimulus features, with posterior computation deferred to downstream areas. B) Posterior coding hypothesis (exemplified by the neural sampling code [16]) posits that early sensory populations readily encode the posterior distribution over hidden world state through incorporating prior knowledge potentially conveyed by feedback connections from higher cortical areas.

the neural implementation of probabilistic computation remains actively debated, and how probability distributions are encoded in the brain is an area of active research [10, 11, 12, 13].

An unresolved question concerns the format of probabilistic representations: Do early sensory populations encode likelihood functions over stimuli, or do they readily represent posterior distributions that incorporate prior knowledge [14]? The **likelihood coding hypothesis** (Fig. 1A) proposes that early sensory populations responding to stimuli (e.g., a drifting grating x) with underlying latent world states (e.g., orientation θ) represent likelihood functions $L(\theta) \equiv p(x|\theta)$. The classic form of probabilistic population code [15, 12] exemplifies this hypothesis, proposing that areas like the primary visual cortex (V1) represent likelihood functions, accounting for the inherent variability in neural responses. Previous work shows that likelihood functions decoded from V1 populations are predictive of animals' trial-by-trial choices and reflects uncertainty associated with the stimuli [12].

In contrast, motivated by extensive feedback from higher cortical areas that could convey prior information, the **posterior coding hypothesis** (Fig. 1B) posits that sensory populations directly represent posterior distributions over latent world states $p(\theta|x)$, suggesting that even early sensory areas would incorporate knowledge of priors to compute posteriors. The neural sampling code [16] is the most concrete example where a neural population is posited to represent a posterior distribution by drawing a "sample" from the distribution and encoding it in its stochastic responses, thereby suggesting that neural variability naturally reflects the sampling process [11, 17, 18, 19, 20].

The critical distinction between the two probabilistic coding hypotheses lies in whether stimulus priors $p(\theta)$ would modulate early sensory population responses. While existing approaches have demonstrated that specific instantiations of each code can capture observed sensory neural response patterns [12, 18], there is yet to be experimental work aimed to directly distinguish the predictions from each coding hypothesis. A fundamental challenge lies in identifying experimental designs—specifically, stimulus prior distributions—that would maximally differentiate the two coding hypotheses [7, 21]. Since both probabilistic coding hypotheses can often account for similar neural response patterns under traditional experimental conditions, targeted task designs where their predictions diverge maximally are crucial for distinguishing between likelihood and posterior coding schemes.

In this work, we present an information-theoretic framework for designing experiments that optimally differentiate likelihood and posterior coding hypotheses. Our approach quantifies the expected difference in decodable information—which we term the **information gap**—when applying neural network-based decoders to extract likelihood or posterior information from sensory neural populations following either coding scheme. Specifically, we (1) derive analytical expressions for the information gap under both coding hypotheses, evaluted as the Kullback–Leibler (KL) divergence between the true posterior and a surrogate posterior utilizing Bayes-optimal estimators; (2) validate theoretical predictions through simulations with deep neural network decoders applied to synthetic populations; and (3) demonstrate how maximizing the information gap yields stimulus distributions that optimally differentiate the two probabilistic coding hypotheses.

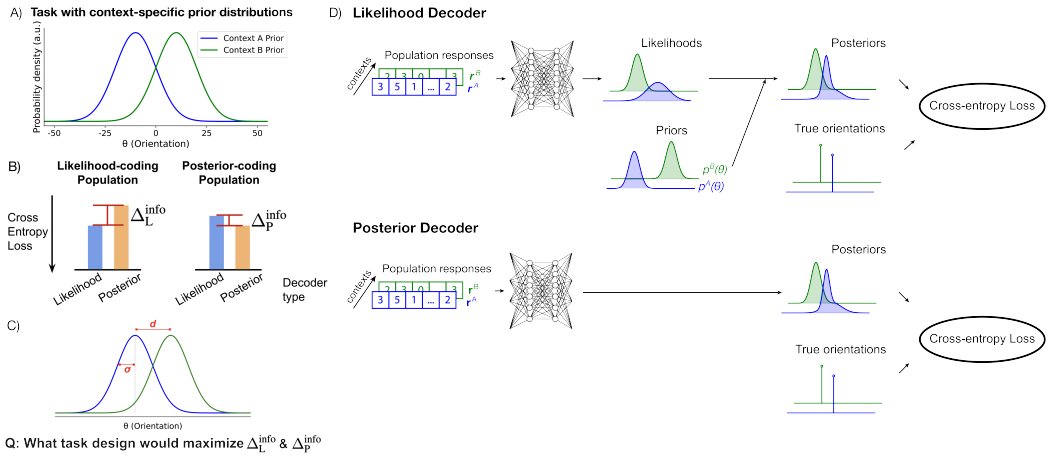

Figure 2: **A decoding approach to differentiating probabilistic neural codes.** A) An experimental paradigm consists of two contexts c with context-specific prior distributions $p^c(\theta)$. B) Schematic for how information gap, the difference in likelihood (blue) and posterior (orange) decoder performances, can indicate whether the underlying neural population encodes the likelihood function (left) or the posterior distributions (right). C) Example task parameters that would be varied to maximize the information gap. D) Schematic for deep neural network-based decoders for decoding the likelihood function (top) or the posterior distribution (bottom).

Our framework provides a principled metric for optimizing experimental designs, establishing the theoretical upper bound on distinguishability between the two coding hypotheses for a given experimental design. By maximizing this metric, we identify stimulus distributions that yield maximally differential decoder performance—enabling rigorous, empirically testable predictions that directly adjudicate between competing theories of probabilistic neural computation.

2 Information gap

We propose to determine whether early sensory populations encode likelihood functions or posterior distributions by examining how varying stimulus priors affects population responses. Classic orientation discrimination tasks under different contexts naturally involve altered stimulus prior distributions, making them ideal for testing this distinction [6, 12]. Our experimental paradigm manipulates priors across two different contexts and examines whether population responses vary according to these changes in stimulus statistics (Fig. 2A)—a design that would leave likelihood encoding population responses invariant across contexts while systematically affecting posterior encoding populations.

Our approach leverages a decoding framework to distinguish the probabilistic information content encoded in neural populations. As schematized in Fig. 2B, the core insight is that decoder performance degrades when attempting to extract mismatched probabilistic content: if a neural population encodes likelihood functions, a decoder trained to extract likelihood information should outperform one extracting posterior information, and vice versa for posterior-coding populations. This differential performance between likelihood and posterior decoders thus serves as a diagnostic tool for identifying the underlying probabilistic code. Building on recent advances in neural decoding [12], we employ deep learning-based decoders that can effectively extract the encoded information while incorporating the structural assumptions of each probabilistic coding scheme (Fig. 2D).

However, it is unclear what stimulus prior distributions would lead to maximal differentiabiltiy between the two probabilistic coding hypotheses (Fig. 2C). While intuition suggests using maximally different context priors, this would limit stimulus overlap across contexts, hence preventing observation of how different priors modulate responses to identical stimuli. This tradeoff—requiring sufficient prior differences to generate distinguishable responses under posterior coding while maintaining adequate overlap for meaningful comparisons—cannot be resolved through intuition alone. We therefore develop an information-theoretic framework that quantifies the expected decoder performance difference to systematically optimize experimental designs.

Experimental paradigm Consider a generative model of sensory observations $\theta \to x$, where x represents noisy sensory observations (e.g. drifting gratings) generated according to the conditional distribution $p(x|\theta)$ given the hidden world state θ (e.g. orientation). We consider an experimental task as introduced in Fig. 2A with two contexts $c = \{A, B\}$, each with associated context frequencies $p(c)$ and context-specific priors $p(\theta|c) \equiv p^c(\theta)$.

Given neural population response vectors $\boldsymbol{r}$, our goal is to quantify the difference in decoder accuracies between a likelihood decoder $g_{\text{L}}(\boldsymbol{r})$ and a posterior decoder $g_{\text{P}}(\boldsymbol{r})$, trained to extract likelihood functions and posterior distributions, respectively. We adopt an information theoretical approach to derive the *expected* difference in decoder performance—an information-theoretical quantity we termed *information gap*—for both probabilistic coding hypotheses under the theoretical assumption of perfect decoding of the sensory information. Although any empirical decoder would underestimate the true sensory information content, we posit that the theoretical limits would serve as important reference points in evaluating the task design. Below we derive the information gap under each of the two probabilistic coding hypotheses.

Information gap for likelihood-coding hypothesis For discretized sensory observations $x \in \{x_i\}$, given a task design specified by $(p(c), p^c(\theta))\ \forall c \in \{A, B\}$ and a generative model $p(x_i|\theta)$, the information gap $\Delta_{\text{L}}^{\text{info}}$ (expected difference between likelihood and posterior decoder accuracies) for a likelihood-coding population $\boldsymbol{r}_{\text{L}} \sim p(x|\theta)$ can be expressed as (see Appendix A.1 for full derivation):

$$\begin{aligned}\Delta_{\text{L}}^{\text{info}} &\coloneqq \mathbb{E}_{p(x_i,c)}\big[D_{\text{KL}}(p^c(\theta|x_i) \,||\, q^*_{P,i}(\theta))\big] \\ &= \sum_{x_i} \Big\{ D_{\text{KL}}(p^A(\theta|x_i) \,||\, q^*_{P,i}(\theta)) \cdot p(c=A) \Big[\sum_{\theta} p(x_i|\theta) p^A(\theta)\Big] + \\ &\qquad D_{\text{KL}}(p^B(\theta|x_i) \,||\, q^*_{P,i}(\theta)) \cdot p(c=B) \Big[\sum_{\theta} p(x_i|\theta) p^B(\theta)\Big] \Big\} \end{aligned} \tag{1}$$

where $p^c(\theta|x_i)$ is the true posterior given observation x_i, which is the output of the likelihood decoder, and $q^*_{P,i}(\theta)$ denotes the surrogate posterior produced by the posterior decoder using Bayes-optimal estimators. The surrogate posterior $q^*_{P,i}(\theta)$ is given by:

$$q^*_{P,i}(\theta) = \frac{[p(c=A)p^A(\theta) + p(c=B)p^B(\theta)] \cdot p(x_i|\theta)}{\sum_{\theta'} \{[p(c=A)p^A(\theta') + p(c=B)p^B(\theta')] \cdot p(x_i|\theta')\}} \tag{2}$$

Since likelihood-coding populations $\boldsymbol{r}_L$ contain no prior information, a posterior decoder trained on such population responses cannot perfectly reconstruct the posterior. Instead, the posterior decoder output converges to a Bayes-optimal estimator of context-dependent posteriors determined by the context distributions $p(c)$ and $p^c(\theta)$.

Information gap for posterior-coding hypothesis For discretized sensory observations $x \in \{x_i\}$, given a task design specified by $(p(c), p^c(\theta))\ \forall c \in \{A, B\}$ and a generative model $p(x_i|\theta)$, the information gap $\Delta_{\text{P}}^{\text{info}}$ (expected difference between likelihood and posterior decoder accuracies) for a posterior-coding population $\boldsymbol{r}_{\text{P}} \sim p(\theta|x)$ is evaluated as (see Appendix A.1 for full derivation):

$$\begin{aligned}\Delta_{\text{P}}^{\text{info}} &\coloneqq \mathbb{E}_{p(x_i,c)}\big[D_{\text{KL}}(p^c(\theta|x_i) \,||\, q^{c*}_{L,i}(\theta))\big] \\ &= \sum_{(x_j,x_k)} \Big\{ D_{\text{KL}}(p^A(\theta|x_j) \,||\, q^{A*}_{L,j}(\theta)) \cdot p(c=A) \Big[\sum_{\theta} p(x_j|\theta) p^A(\theta)\Big] \\ &\qquad + D_{\text{KL}}(p^B(\theta|x_k) \,||\, q^{B*}_{L,k}(\theta)) \cdot p(c=B) \Big[\sum_{\theta} p(x_k|\theta) p^B(\theta)\Big] \Big\} \end{aligned} \tag{3}$$

where $p^c(\theta|x_i)$ is the true posterior given observation x_i, which is the output of the posterior decoder, and $q^*_{L,i}(\theta)$ denotes a surrogate posterior which is the posterior distribution associated with the output of likelihood decoders utilizing Bayes-optimal estimators. The sum includes only pairs (x_j, x_k) that satisfy the condition expressed below in Eq. 4 as these pairs represent scenarios where identical population responses $\boldsymbol{r}_P$ (encoding the same posterior distribution across the two contexts $c \in \{A, B\}$, i.e., $\boldsymbol{r}^A_{P,j} \approx \boldsymbol{r}^B_{P,k}$) must map to different likelihood functions ($p(x_j|\theta)$ and $p(x_k|\theta)$ respectively), preventing the likelihood decoder from achieving perfect decoding.

$$\forall_\theta,\ p^A(\theta|x_j) = p^B(\theta|x_k) \ \Leftrightarrow\ \forall_\theta,\ p^A(\theta) \cdot p(x_j|\theta) \propto p^B(\theta) \cdot p(x_k|\theta) \tag{4}$$

With this, the surrogate posteriors for the pair (x_j, x_k) $q^*_{L,i}(\theta)$ is given by:

$$q^{A*}_{L,j}(\theta) = \frac{\ell^*_{jk}(\theta)p^A(\theta)}{Z^A_j[\ell^*_{jk}(\theta)]}, \quad q^{B*}_{L,k}(\theta) = \frac{\ell^*_{jk}(\theta)p^B(\theta)}{Z^B_k[\ell^*_{jk}(\theta)]}$$

where $\ell^*_{jk}(\theta)$ denotes the output of the likelihood decoder on the posterior-coding population, approaching the Bayes-optimal estimator of the likelihood functions. $Z^A_j[\ell^*_{jk}(\theta)]$ and $Z^B_k[\ell^*_{jk}(\theta)]$ are normalization constants dependent on $\ell^*_{jk}(\theta)$, defined as:

$$Z^A_j[\ell^*_{jk}(\theta)] \coloneqq \sum_\theta p^A(\theta)\ell^*_{jk}(\theta), \quad Z^B_k[\ell^*_{jk}(\theta)] \coloneqq \sum_\theta p^B(\theta)\ell^*_{jk}(\theta)$$

The Bayes-optimal likelihood function estimator $\ell^*_{jk}(\theta)$ can be found by solving the following implicit equation using fixed-point iteration (see Appendix A.1 for detail):

$$\ell^*_{jk}(\theta) = \frac{\rho^A_j p^A(\theta|x_j) + \rho^B_k p^B(\theta|x_k)}{\frac{\rho^A_j}{Z^A_j[\ell^*_{jk}(\theta)]}p^A(\theta) + \frac{\rho^B_k}{Z^B_k[\ell^*_{jk}(\theta)]}p^B(\theta)} \tag{5}$$

where ρ^A_j and ρ^B_k denote the frequencies of each context given observing neural population responses coming from $\boldsymbol{r}^A_{\mathrm{P},j}$ or $\boldsymbol{r}^B_{\mathrm{P},k}$. Let us first define:

$$S^A_j \coloneqq p(c=A)\sum_\theta p^A(\theta)p(x_j|\theta), \quad S^B_k \coloneqq p(c=B)\sum_\theta p^B(\theta)p(x_k|\theta)$$

Then the context frequencies ρ^A_j and ρ^B_k are evaluated as:

$$\rho^A_{jk} \coloneqq p(c=A|\boldsymbol{r}=\boldsymbol{r}^A_{\mathrm{P},j}\vee\boldsymbol{r}^B_{\mathrm{P},k}) = S^A_j/(S^A_j+S^B_k),\ \rho^B_{jk} \coloneqq p(c=B|\boldsymbol{r}=\boldsymbol{r}^A_{\mathrm{P},j}\vee\boldsymbol{r}^B_{\mathrm{P},k}) = S^B_k/(S^A_j+S^B_k)$$

In summary, our information-theoretic framework provides analytical expressions for the information gap—the expected difference in decoder performances measured as cross-entropy—for both likelihood-coding hypothesis (Eq. 1) and posterior-coding hypothesis (Eq. 3). The key insight enabling the above analytic derivation stems from identifying the Bayes-optimal estimators for decoding mismatched probabilistic information content, for instance, when decoding the posterior distribution from a likelihood-coding population (Eq. 2) or when decoding the likelihood function from a posterior-coding population (Eq. 5). We next validate the decoders and the expected information gap by training and applying the decoders on two synthetic neural populations following different information coding hypotheses (likelihood vs posterior-coding) and compare the empirical difference in the the decoder performances to that predicted by our information gap measure as derived above. We will then demonstrate how maximizing information gap enables targeted experimental design that optimally differentiate the two probabilistic coding hypotheses.

3 Simulation experiments

To validate that the information gap accurately predicts decoder performance differences for both probabilistic coding hypotheses, we conducted comprehensive simulation experiments. We constructed synthetic likelihood-coding and posterior-coding neural populations and trained both likelihood and posterior decoders on these populations (Fig. 2). These simulations serve two complementary purposes: validating our theoretical framework and providing practical insights into the scaling properties and convergence behavior of the information gap measure.

Task design: Gaussian context priors We consider Gaussian context priors motivated by known neurophysiological experiments [11, 12]. In this task, subjects perform an orientation discrimination task with two contexts $c \in \{A, B\}$, with the context for each session sampled randomly from the two with equal probability. Within each session, the trial-to-trial hidden world state θ (i.e. the orientation) is drawn from context-specific Gaussian prior distributions $p^c(\theta) = \mathcal{N}(\mu^c, (\sigma^c)^2)$, where μ^c and $(\sigma^c)^2)$ are the task-specific parameters. In the simulation, we adopt an identical variance for the two Gaussian priors $\sigma^A = \sigma^B = \sigma$, but in principle this could be varied when considering task designs.

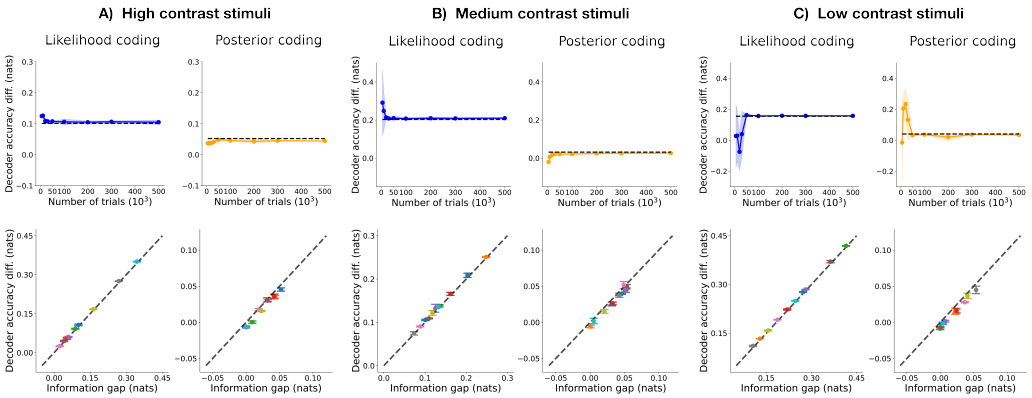

Figure 3: **Information gap accurately predicts the decoder performance difference on simulated likelihood-coding and posterior-coding populations.** A) (top) On simulated neural populations encoding the likelihood function (left, blue) or the posterior distributions (right, orange) presented with high contrast stimuli, the difference between the likelihood and posterior decoder performances converges to the theoretical value of information gap as the total number of trials used for decoder training increases (shaded areas denote the s.t.d. across 5 random seeds); (bottom) Across multiple task design parameters, theoretical values of information gap (x-axis) accurately predicts the decoder performance difference on the simulated neural populations (y-axis), for both the likelihood-coding populations (left) and the posterior-coding populations (right). (Each color marks one set of task parameters that are used for both types of simulated populations; Error bars denote the s.t.d. across 5 random seeds.) B) Same for medium contrast stimuli and C) for low contrast stimuli.

We simulate a noisy sensory observation x by drawing from the conditional distribution defined by the given generative model $p(x|\theta)$. Note that this stochastic process can be seen as implicitly capturing both intrinsic neuronal noise and extrinsic stimulus features. This generative model can be experimentally manipulated through stimulus parameters such as contrast—lower contrast increases observation variance, reflecting greater sensory uncertainty. The generative model $p(x|\theta)$ is modeled as Gaussian distributions to reflect Gaussian orientation tuning curves in early V1 and to capture the effect of different contrast levels by systematically varying standard deviations.

For synthetic neural population responses, we implemented neurons to follow Gaussian tuning curves with Poisson variability [12]. After sensory observations x were sampled from the generative model $p(x|\theta)$ for given stimuli θ, they are encoded through Gaussian tuning curves to yield mean firing rate per neuron. Likelihood-coding populations $\boldsymbol{r}_L$ respond solely based on the observations x, while posterior-coding populations $\boldsymbol{r}_P$ have firing rates additionally modulated by the context-specific prior $p^c(\theta)$, thus effectively encoding the posterior $p^c(\theta|x) \propto p(x|\theta) \cdot p^c(\theta)$. In both cases, spike counts were then generated from the mean rates using Poisson distribution. As described in Fig. 2D, deep neural networks are trained with cross-entropy loss to serve as flexible, powerful decoders of probabilistic distributions from simulated neural population responses [12].

Scaling and convergence We first examine the scaling and convergence properties of the theoretical prediction of information gap. The top row of Fig. 3 demonstrates convergence of the empirical difference in decoder performances on simulated neural populations across stimulus contrast levels. As training data (number of trials) increases, decoder performance differences for both simulated populations—likelihood-coding (blue) and posterior-coding (orange)—rapidly converge to the theoretical values of information gap (dashed lines) computed via Eq. 1 and 3. This convergence of the simulated decoder performance difference to our derived information gap prediction suggests that the information gap successfully captures the asymptotic decoder performance difference.

Validation across parameter space Next, we assess the validity of the theoretical prediction of information gap across a wide range of task parameters and generative models. To this end, we systematically vary the generative model by varying the level of sensory "contrast" and for each contrast level we simulated populations under at least ten different sets of task parameters. The bottom row of Fig. 3 systematically validate theoretical predictions across diverse task parameters.

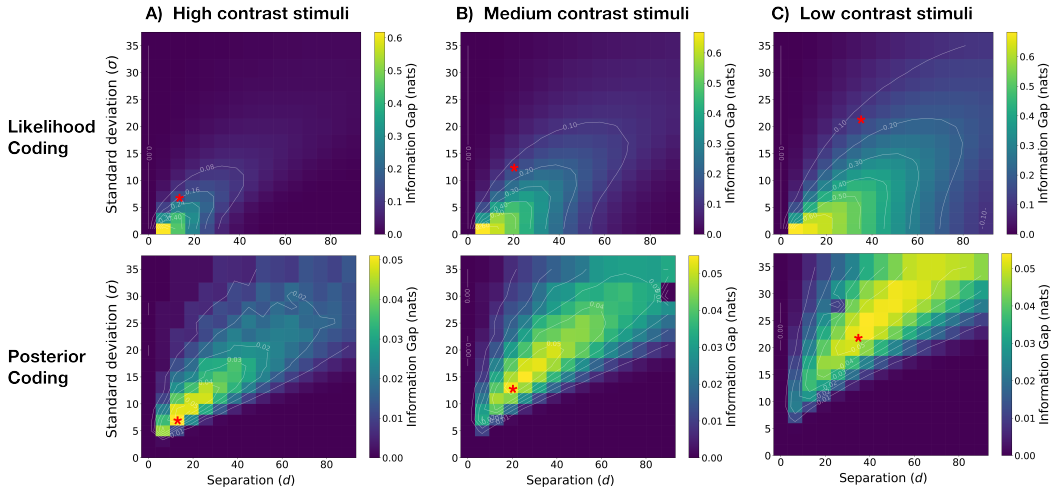

Figure 4: **Information gap landscapes inform practical task designs that optimally differentiate probabilistic representations in neural populations.** A) Information gap as a function of task parameters (d: separation between context priors, and σ: context prior standard deviations) for both the likelihood coding hypothesis (top) and the posterior coding hypothesis (bottom) when presented with high contrast stimuli. The red asterisks identify strategic task designs that achieve the tradeoff where posterior-coding information gap approaches its maximum while likelihood-coding maintains sufficient discriminative signal. B) Same for medium contrast stimuli and C) for low contrast stimuli.

Plotting empirical decoder performance differences against theoretical values of information gap for each task design parameter reveals remarkable agreement for both probabilistic coding schemes.

A notable finding is that information gaps for likelihood-coding populations exceed those for posterior-coding populations by approximately an order of magnitude. Our theoretical framework provides an intuitive explanation: in likelihood coding, every stimulus contributes to the information gap calculation, whereas in posterior coding, only observation pairs satisfying Eq. 4 contribute. This asymmetry suggests that distinguishing probabilistic decoders in posterior-coding populations presents greater experimental challenges, requiring careful task design to achieve sufficient statistical power.

Overall, these simulation results establish that our information-theoretic framework accurately predicts decoder performance differences for neural populations following either probabilistic coding hypothesis, providing a quantitative foundation for designing targeted, theory-driven experiments to differentiate the fundamental theories of probabilistic neural representations.

4 Task optimization for differentiating probabilistic neural codes

Given the high degree of agreement between the empirical and predicted information gap measures as computed in Eq. 1 and 3, we now leverage this framework to optimize experimental designs that maximally differentiate the two probabilistic coding hypotheses. We systematically explore the task parameter space to identify experimental conditions that would yield maximum discriminative power.

Information gap landscape We evaluated information gap across the two-dimensional task parameter space defined by: (1) the distance between the means of the two Gaussian context priors $d = |\mu^A - \mu^B|$, and (2) the shared standard deviation for both Gaussian context priors $\sigma = \sigma^A = \sigma^B$. The landscapes of information gap across three different levels of stimulus contrasts are presented in Fig. 4, with the top row showing the results for populations following the likelihood-coding hypothesis and the bottom row for those following the posterior-coding hypothesis.

Key patterns emerge from the information landscape analysis. First, we observe that the information gap landscape depends on the stimulus contrast level, suggesting that experimental design should be tailored to specific stimulus features such as the contrast level. In addition, decreasing stimulus contrast expands the parameter region yielding substantial information gaps for both probabilistic

coding hypotheses. This agrees with the intuition that prior information becomes increasingly influential when the sensory observation alone provides insufficient information for reliable inference.

Strategic task design Crucially, for a given contrast level, the information gap landscapes strongly depends on the underlying population coding hypotheses, revealing an inherent trade-off to consider when optimizing experimental design: task parameters that maximize the discriminability for likelihood-coding populations diverge from those optimal for posterior-coding populations. This divergence necessitates strategic selection of experimental parameters that balance discriminative power across both hypotheses.

The marked asymmetry in information gap magnitudes—with posterior-coding values typically an order of magnitude smaller than those of likelihood-coding populations—suggests prioritizing parameters that maximize posterior-coding discriminability while maintaining adequate likelihood-coding sensitivity. The red asterisks in Figure 4 identify such strategic "sweet spots" where posterior-coding information gap approaches its maximum while that under likelihood-coding hypothesis maintains sufficient discriminative signal.

Practical task implementation The information landscapes provide concrete guidance for experimental design. For high-contrast stimuli, optimal discrimination occurs with small prior separations ($d \approx$ 15-20) and smaller standard deviation s($\sigma \approx$ 5-10). As contrast decreases, the optimal region shifts toward larger prior separations and wider standard deviations.

Importantly, our framework enables researchers to systematically optimize experimental designs for differentiating between specific neural coding theories by quantifying expected information gaps across parameter spaces. This optimization directly identifies experimental parameters that maximize statistical power within practical constraints, transforming parameter selection from an exploratory process into a theory-driven procedure that substantially increases the probability of obtaining decisive empirical results. The resulting information gap landscapes serve as navigational maps, revealing parameter combinations that most effectively distinguish whether early sensory populations encode likelihood functions or posterior distributions.

5 Discussion and conclusions

We presented an information-theoretic framework that enables targeted experimental design to resolve a fundamental question: whether early sensory neural populations encode likelihood functions or posterior distributions. We derive analytical expressions for the *information gap*—the expected decoder performance difference when extracting mismatched probabilistic content. This measure quantifies how effective an experimental design can distinguish between competing probabilistic coding hypotheses, providing precise predictions validated through extensive simulations. Most critically, maximizing this information gap yields principled experimental designs that optimally discriminate between probabilistic neural codes, enabling decisive experiments to resolve this theoretical debate.

By providing the first principled methodology for experimentally distinguishing how neural populations represent uncertainty, this work would help addresses fundamental questions about Bayesian computation in the brain. More broadly, this approach bridges computational theory and experimental neuroscience. Our framework provides a general blueprint for testing competing neural coding theories through information-theoretic optimization. By quantifying whether experiments can distinguish between hypotheses before data collection, this work demonstrates how theoretical frameworks can directly accelerate empirical discovery in neuroscience.

Scope and limitations Our framework provides a foundation that can be extended in several directions. First, while we focus on discretized stimuli with Gaussian priors for analytical tractability, the mathematical framework naturally extends to continuous spaces and arbitrary distributions through numerical methods. Second, although our analysis considers pure likelihood or posterior coding, the decoding approach can be used to and characterize hybrid schemes, providing a tool to explore the full spectrum of probabilistic representations. Third, our simulations employ Poisson variability models that capture essential response statistics; incorporating more complex features such as noise correlations and response nonlinearities would further strengthen predictions for biological neural populations with richer details. These extensions represent opportunities to refine the framework as experimental techniques and computational resources continue to advance.

References

[1] Pierre Simon de Laplace. *Théorie analytique des probabilités*, volume 7. Courcier, 1820.

[2] H von Helmholtz. Versuch einer erweiterten anwendung des fechnerschen gesetzes im farbensystem. *Z. Psychol. Physiol. Sinnesorg*, 2:1–30, 1891.

[3] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[4] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[5] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55(1):271–304, 2004.

[6] Ahmad T Qamar, R James Cotton, Ryan G George, Jeffrey M Beck, Eugenia Prezhdo, Allison Laudano, Andreas S Tolias, and Wei Ji Ma. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences*, 110(50):20332–20337, 2013.

[7] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual review of neuroscience*, 37(1):205–220, 2014.

[8] Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

[9] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

[10] Tianming Yang and Michael N Shadlen. Probabilistic reasoning by neurons. *Nature*, 447(7148): 1075–1080, 2007.

[11] Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

[12] Edgar Y Walker, R James Cotton, Wei Ji Ma, and Andreas S Tolias. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23(1):122–129, 2020.

[13] Laurence Aitchison, Jannes Jegminat, Jorge Aurelio Menendez, Jean-Pascal Pfister, Alexandre Pouget, and Peter E Latham. Synaptic plasticity as bayesian inference. *Nature neuroscience*, 24 (4):565–571, 2021.

[14] Ralf M Haefner, Jeff Beck, Cristina Savin, Mehrdad Salmasi, and Xaq Pitkow. How does the brain compute with probabilities? *arXiv preprint arXiv:2409.02709*, 2024.

[15] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[16] Patrik Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. *Advances in neural information processing systems*, 15, 2002.

[17] József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3): 119–130, 2010.

[18] Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

[19] Richard D Lange and Ralf M Haefner. Task-induced neural covariability as a signature of approximate bayesian learning and inference. *PLoS computational biology*, 18(3):e1009557, 2022.

[20] Suhas Shrinivasan, Konstantin-Klemens Lurz, Kelli Restivo, George Denfield, Andreas Tolias, Edgar Walker, and Fabian Sinz. Taking the neural sampling code very seriously: A data-driven approach for evaluating generative models of the visual system. *Advances in Neural Information Processing Systems*, 36:21945–21959, 2023.

[21] Sabyasachi Shivkumar, Richard Lange, Ankani Chattoraj, and Ralf Haefner. A probabilistic population code based on neural samples. *Advances in neural information processing systems*, 31, 2018.

A Technical Appendices and Supplementary Material

A.1 Information gap derivation

Consider a generative model of sensory observations $\theta \rightarrow x$, where x is the noisy sensory observation (e.g. a drifting grating stimulus) generated according to $p(x|\theta)$, the likelihood function determined by the generative model, given the hidden state of the environment θ (e.g. true orientation of the drifting grating stimulus). Consider an experimental setup where there are two possible contexts: $c = \{A, B\}$ with their associated context priors $p(\theta|c) \equiv p^c(\theta)$.

Given a sensory observation x and a context c, the context-dependent posterior distribution of θ (denoted as $p^c(\theta|x) \equiv p(\theta|x, c)$) is given by the Baye's rule:

$$
\begin{aligned}
p^c(\theta|x) &= \frac{p^c(\theta, x)}{p^c(x)} \\
&= \frac{p^c(x|\theta) \cdot p^c(\theta)}{\sum_{\theta'} p^c(x|\theta') \cdot p^c(\theta')}, \quad \text{Since the generative process } \theta \rightarrow x \text{ is independent of } c, \\
&= \frac{p(x|\theta) \cdot p^c(\theta)}{\sum_{\theta'} p(x|\theta') \cdot p^c(\theta')} \\
&\propto p(x|\theta) \cdot p^c(\theta)
\end{aligned}
$$

For a given neural population response vector $\boldsymbol{r}$, consider two possible probabilistic neural coding hypotheses:

1. **Likelihood coding**: the neural population responses $\boldsymbol{r}_{\mathrm{L}}$ is hypothesized to encode the likelihood function of the stimulus $p(x|\theta)$.
2. **Posterior coding**: the neural population response $\boldsymbol{r}_{\mathrm{P}}$ is hypothesized to encode the posterior distribution of the hidden state given the stimulus $p(\theta|x)$.

We consider whether it is possible to differentiate the probabilistic information content encoded in given neural population responses through a decoding approach. The intuition behind the decoding approach is that if the neural population is encoding the likelihood function, then a decoder decoding the likelihood function should lead to a better performance then a decoder decoding the posterior distribution; vice versa if the neural population is encoding the posterior distribution. In other words, decoder performance degrades when trying to decode mismatched probabilistic content, such that the difference in decoder performance when decoding the likelihood function versus decoding the posterior distribution can be used to differentiate whether a given neural population is encoding the likelihood function (likelihood-coding) or the posterior distribution (posterior-coding).

Consider applying a decoder function g which is optimized to decode some probabilistic information content from the neural population responses:

$$
g(\boldsymbol{r}) \longrightarrow p(\cdot) \quad \text{where } g \text{ is a decoder function}
$$

To establish the expected difference between decoder performances, we consider the limit of perfect decoding where the decoder is expressive enough (e.g. a multi-layer perceptron, MLP), and the data is abundant.

Adopting an information-theoretical approach, given an experimental design, our goal is to derive the *expected difference* between decoder performances when decoding the likelihood function and when decoding the posterior distribution from given neural population responses, a quantity that we termed the *information gap* between the two decoders. We will separately derive the information gap for likelihood-coding and posterior-coding populations, respectively.

A.1.1 Likelihood coding

For a likelihood coding population, the neural population responses $\boldsymbol{r}_{\mathrm{L}}$ encode the likelihood function of the sensory stimulus, and are not modulated and hence independent of the context prior.

$$
\boldsymbol{r}_{\mathrm{L}} \sim f(p(x|\theta)), \text{ where } f \text{ is some neural encoding function.}
$$

Note The decoders are optimized under cross-entropy loss:

$$H(p, q) = -\mathbb{E}_p[\log q] = H(p) + D_{KL}(p \,||\, q)$$

when $q^* = p \Leftrightarrow H(p, q^*)$ is minimized.

Applying a perfect likelihood decoder g_L Applying a likelihood decoder g_L to a likelihood-coding population $\boldsymbol{r}_{\mathrm{L}}$, we want

$$g_L(\boldsymbol{r}_{\mathrm{L}}) \longrightarrow p(x|\theta)$$

Let us consider

$$\forall x_i, c : \boldsymbol{r}^c_{\mathrm{L},i} = \boldsymbol{r}_{\mathrm{L},i} \sim f(p(x|\theta))$$

Since $\boldsymbol{r}_{\mathrm{L},i}$ is context-independent, let us denote the likelihood decoder output $g_L(\boldsymbol{r}^c_{\mathrm{L},i}) = g_L(\boldsymbol{r}_{\mathrm{L},i})$. With the likelihood decoder output and the corresponding context prior, the context-dependent decoded posterior distribution $q^c_{L,i}(\theta)$ is given by:

$$q^c_{L,i}(\theta) = \eta^c_{L,i} \cdot g_L(\boldsymbol{r}_{\mathrm{L},i}) \cdot p^c(\theta), \text{ where } \eta^c_{L,i} \text{ is a normalization constant.}$$

The cross-entropy loss for data samples associated with x_i, c, $H(p^c(\theta|x_i), q^c_{L,i}(\theta))$, is minimized when:

$$\begin{aligned}
& q^{c*}_{L,i}(\theta) = p^c(\theta|x_i) \\
\Rightarrow\ & \eta^c_{L,i} \cdot g^*_L(\boldsymbol{r}_{\mathrm{L},i}) \cdot p^c(\theta) = \frac{p(x_i|\theta) \cdot p^c(\theta)}{p(x_i)} \\
\Rightarrow\ & g^*_L(\boldsymbol{r}_{\mathrm{L},i}) = \alpha^c_{L,i} \cdot p(x_i|\theta), \text{ where } \alpha^c_{L,i} \text{ is a constant}
\end{aligned}$$

That is, after training, the likelihood decoder output $g_L(\boldsymbol{r}_{\mathrm{L},i})$ will converge to $g^*_L(\boldsymbol{r}_{\mathrm{L},i}) \propto p(x_i|\theta)$ given enough samples.

Marginalizing over all x_i, c, the expected cross-entropy loss for a perfect likelihood decoder can be computed as:

$$\begin{aligned}
\mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i), q^{c*}_{L,i}(\theta))] &= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i)) + D_{\mathrm{KL}}(p^c(\theta|x_i) \,||\, q^{c*}_{L,i}(\theta))] \\
&= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i))] \\
&= \sum_{x_i,c} H(p^c(\theta|x_i)) \cdot p(x_i, c) \\
&= \sum_{x_i,c} H(p^c(\theta|x_i)) \cdot p(c) \cdot \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta) \Big] \\
&= \sum_{x_i} \sum_{c} p(c) H(p^c(\theta|x_i)) \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta) \Big] \\
&= \sum_{x_i} \Big\{ H(p^A(\theta|x_i)) \cdot p(c = A) \Big[\sum_{\theta} p(x_i|\theta) p^A(\theta) \Big] + \\
&\qquad H(p^B(\theta|x_i)) \cdot p(c = B) \Big[\sum_{\theta} p(x_i|\theta) p^B(\theta) \Big] \Big\}
\end{aligned}$$

where the second equality holds because $D_{\mathrm{KL}}(p^c(\theta|x_i) \,||\, q^{c*}_{L,i}(\theta)) = 0$ for a perfect likelihood decoder. That is, the expected cross-entropy loss for a perfect likelihood decoder should approach the expected posterior entropy as determined by the context distribution.

Applying the best possible posterior decoder g_P Applying a posterior decoder g_P to a likelihood-coding population $\boldsymbol{r}_{\mathrm{L}}$, we want:

$$g_P(\boldsymbol{r}_{\mathrm{L}}) \longrightarrow p^c(\theta|x)$$

However, since there is no context information encoded in the population responses, the posterior decoder cannot achieve the same performance as the likelihood decoder as there are scenarios where the same inputs ($\boldsymbol{r}_{\mathrm{L}}$) are trained to map to different outputs ($p^c(\theta|x)$) depending on the inaccessible ground-truth context information.

Let us consider:

$$\forall x_i, c : \boldsymbol{r}^c_{\mathrm{L},i} = \boldsymbol{r}_{\mathrm{L},i} \sim f(p(x|\theta))$$

First consider the frequency of a context given observing data samples associated with x_i:

$$\begin{aligned} p(c|x = x_i) &= \frac{p(c, x_i)}{p(x_i)} \\ &= \frac{p(c) \cdot p(x_i|c)}{\sum_{c'} p(c') \cdot p(x_i|c')} \\ &= \frac{p(c) \cdot \sum_\theta p^c(\theta) \cdot p(x_i|\theta)}{\sum_{c'} p(c') \sum_\theta p^{c'}(\theta) \cdot p(x_i|\theta)} \end{aligned}$$

Let us denote

$$\begin{aligned} S^A_i &\coloneqq p(c = A) \sum_\theta p^A(\theta) p(x_i|\theta) \\ S^B_i &\coloneqq p(c = B) \sum_\theta p^B(\theta) p(x_i|\theta) \end{aligned}$$

Hence we can define the observation-dependent context frequency for a given x_i:

$$\begin{aligned} \rho^A_i &\coloneqq p(c = A|x = x_i) = S^A_i/(S^A_i + S^B_i) \\ \rho^B_i &\coloneqq p(c = B|x = x_i) = S^B_i/(S^A_i + S^B_i) \end{aligned}$$

Now, let us denote the posterior decoder output $g_P(\boldsymbol{r}_{\mathrm{L},i}) \equiv q_{P,i}(\theta)$, as the posterior decoder output can be interpreted directly as the posterior distribution over the hidden state θ. Since the posterior decoder output is agnostic to the specific context prior, under cross-entropy loss, $q_{P,i}(\theta)$ is trained to minimize:

$$\begin{aligned} &\min_{q_{P,i}(\theta)} \Big\{ \mathbb{E}_{p(c|x_i)} \big[H(p^c(\theta|x_i), q_{P,i}(\theta)) \big] \Big\} \\ = &\min_{q_{P,i}(\theta)} \Big\{ \rho^A_i H(p^A(\theta|x_i), q_{P,i}(\theta)) + \rho^B_i H(p^B(\theta|x_i), q_{P,i}(\theta)) \Big\} \\ = &\min_{q_{P,i}(\theta)} \Big\{ -\sum_\theta \big[\rho^A_i p^A(\theta|x_i) + \rho^B_i p^B(\theta|x_i) \big] \cdot \log q_{P,i}(\theta) \Big\} \end{aligned}$$

Since $p^A(\theta|x_i)$ and $p^B(\theta|x_i)$ are probability distributions over θ, and $\rho^A_i + \rho^B_i = 1$, hence $\rho^A_i p^A(\theta|x_i) + \rho^B_i p^B(\theta|x_i)$ is also a probability distribution over θ. Therefore the above minimization happens when:

$$\begin{aligned} q^*_{P,i}(\theta) &= \rho^A_i p^A(\theta|x_i) + \rho^B_i p^B(\theta|x_i) \\ &= \frac{S^A_i}{S^A_i + S^B_i} \frac{p^A(\theta) p(x_i|\theta)}{\sum_\theta p^A(\theta) p(x_i|\theta)} + \frac{S^B_i}{S^A_i + S^B_i} \frac{p^B(\theta) p(x_i|\theta)}{\sum_\theta p^B(\theta) p(x_i|\theta)} \\ &= \frac{[p(c = A)p^A(\theta) + p(c = B)p^B(\theta)] \cdot p(x_i|\theta)}{S^A_i + S^B_i} \\ &= \frac{[p(c = A)p^A(\theta) + p(c = B)p^B(\theta)] \cdot p(x_i|\theta)}{\sum_{\theta'} \{[p(c = A)p^A(\theta') + p(c = B)p^B(\theta')] \cdot p(x_i|\theta')\}} \end{aligned}$$

That is, after training, the best possible posterior decoder output for data samples associated with x_i is as if the decoder was using a surrogate prior $\tilde{p}_i(\theta) = p(c = A)p^A(\theta) + p(c = B)p^B(\theta)$, which is the Bayes-optimal estimator of the prior distributions across contexts, and in this case equals the marginalized prior distribution over θ across contexts.

Marginalizing over all x_i, c, the expected cross-entropy loss for the best possible posterior decoder can be computed as:

$$\begin{aligned}\mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i), q^*_{P,i}(\theta))] &= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i)) + D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta))] \\ &= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i))] + \mathbb{E}_{p(x_i,c)}\big[D_{KL}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta))\big] \\ &= \text{CE loss for the perfect likelihood decoder} \\ &\quad + \mathbb{E}_{p(x_i,c)}\big[D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta))\big] \end{aligned} \tag{6}$$

Information gap for a likelihood-coding population $\Delta_{\text{L}}^{\text{info}}$ From equation 6, let us define $\Delta_{\text{L}}^{\text{info}}$, information gap for a likelihood-coding population between a perfect likelihood decoder g^*_L and the best possible posterior decoder g^*_P, as the difference in the expected cross-entropy loss of the two decoders:

$$\begin{aligned}\Delta_{\text{L}}^{\text{info}} &\coloneqq \mathbb{E}_{p(x_i,c)}\big[D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta))\big] \\ &= \sum_{x_i,c} D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta)) \cdot p(x_i, c) \\ &= \sum_{x_i,c} D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta)) \cdot p(c) \cdot \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\ &= \sum_{x_i} \sum_{c} p(c) D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^*_{P,i}(\theta)) \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\ &= \sum_{x_i} \Big\{ D_{\text{KL}}(p^A(\theta|x_i) \mid\mid q^*_{P,i}(\theta)) \cdot p(c = A) \Big[\sum_{\theta} p(x_i|\theta) p^A(\theta)\Big] + \\ &\qquad D_{\text{KL}}(p^B(\theta|x_i) \mid\mid q^*_{P,i}(\theta)) \cdot p(c = B) \Big[\sum_{\theta} p(x_i|\theta) p^B(\theta)\Big] \Big\}\end{aligned}$$

A.1.2 Posterior coding

For a posterior coding population, the neural population responses $\boldsymbol{r}^c_{\text{P}}$ encode the posterior distribution over θ given x under a context c, and are therefore modulated by and dependent on the context prior.

$$\boldsymbol{r}^c_{\text{P}} \sim f(p^c(\theta|x)), \text{ where } f \text{ is some neural encoding function.}$$

Applying a perfect posterior decoder g_P Applying a posterior decoder g_P to a posterior-coding population $\boldsymbol{r}_{\text{P}}$, we want

$$g_P(\boldsymbol{r}_{\text{P}}) \longrightarrow p^c(\theta|x)$$

Let us consider

$$\forall x_i, c : \boldsymbol{r}^c_{\text{P},i} \sim f(p^c(\theta|x_i))$$

Denote the output of a posterior decoder as $g_P(\boldsymbol{r}^c_{\text{P},i}) \equiv q^c_{P,i}(\theta)$, which is context-dependent as $\boldsymbol{r}^c_{\text{P},i}$ is context-dependent. The cross-entropy loss for data samples associated with x_i, c, $H(p^c(\theta|x_i), q^c_{P,i}(\theta))$, is minimized when:

$$q^{c*}_{P,i}(\theta) = p^c(\theta|x_i)$$

That is, after training, the posterior decoder output $g_P(\boldsymbol{r}^c_{\text{P},i})$ will converge to $q^{c*}_{P,i}(\theta) = p^c(\theta|x_i)$ given enough samples. Marginalizing over all x_i, c, the expected cross-entropy loss for a perfect

posterior decoder can be computed as:

$$\begin{aligned}
\mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i), q^{c*}_{P,i}(\theta))] &= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i)) + D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^{c*}_{P,i}(\theta))] \\
&= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i))] \\
&= \sum_{x_i,c} H(p^c(\theta|x_i)) \cdot p(x_i, c) \\
&= \sum_{x_i,c} H(p^c(\theta|x_i)) \cdot p(c) \cdot \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\
&= \sum_{x_i} \sum_{c} p(c) H(p^c(\theta|x_i)) \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\
&= \sum_{x_i} \Big\{ H(p^A(\theta|x_i)) \cdot p(c=A) \Big[\sum_{\theta} p(x_i|\theta) p^A(\theta)\Big] + \\
&\qquad H(p^B(\theta|x_i)) \cdot p(c=B) \Big[\sum_{\theta} p(x_i|\theta) p^B(\theta)\Big] \Big\}
\end{aligned}$$

where the second equality holds because $D_{\text{KL}}(p^c(\theta|x_i) \mid\mid q^{c*}_{P,i}(\theta)) = 0$ for a perfect posterior decoder. That is, the expected cross-entropy loss for a perfect posterior decoder on a posterior-coding population should approach the expected posterior entropy as determined by the context distribution. Note this is the same as the expected cross-entropy loss for a perfect likelihood decoder on a likelihood-coding population as derived previously.

Applying the best possible likelihood decoder g_L Applying a likelihood decoder g_L to a posterior-coding population $\boldsymbol{r}_{\text{P}}$, we want

$$g_L(\boldsymbol{r}_{\text{P}}) \longrightarrow p(x|\theta)$$

Let us consider some pair (x_j, x_k) such that

$$\begin{aligned}
& p^A(\theta|x_j) = p^B(\theta|x_k),\ \forall\theta \quad \text{(can be measured in terms of KL divergence)} \\
\Leftrightarrow\quad & p^A(\theta) \cdot p(x_j|\theta) \propto p^B(\theta) \cdot p(x_k|\theta),\ \forall\theta \\
\Leftrightarrow\quad & \boldsymbol{r}^A_{\text{P},j} \sim f(p^A(\theta|x_j)) \\
& \cong \boldsymbol{r}^B_{\text{P},k} \sim f(p^B(\theta|x_k))
\end{aligned}$$

That is, we consider the condition $\boldsymbol{r}^A_{\text{P},j} \cong \boldsymbol{r}^B_{\text{P},k}$, where the input ($\boldsymbol{r}^A_{\text{P},j}$ or $\boldsymbol{r}^B_{\text{P},k}$) to the likelihood decoder is approximately the same but the target output is different depending on the context ($p(x_j|\theta)$ or $p(x_k|\theta)$). Under the limit of perfect decoding, $\chi \coloneqq \{(x_j, x_k)\}$ are the scenarios which will cause the likelihood decoder to not be perfect even with large data samples and expressive parametrization. The intuition is that with expressive enough decoder and large enough data samples, the only scenarios where the best possible likelihood decoder g^*_L is still not perfect are those where the same inputs (population responses encoding posterior distributions) need to be decoded into different outputs (likelihood function) depending on the context.

Let us first consider the frequency of a context given observing data samples associated with $\boldsymbol{r}^A_{\text{P},j}$ or $\boldsymbol{r}^B_{\text{P},k}$:

$$\begin{aligned}
p(c=A|\boldsymbol{r} = \boldsymbol{r}^A_{\text{P},j} \vee \boldsymbol{r}^B_{\text{P},k}) &= \frac{p(c=A, \boldsymbol{r} = \boldsymbol{r}^A_{\text{P},j} \vee \boldsymbol{r}^B_{\text{P},k})}{p(\boldsymbol{r} = \boldsymbol{r}^A_{\text{P},j} \vee \boldsymbol{r}^B_{\text{P},k})} \\
&= \frac{p(c=A) \cdot \sum_\theta p^A(\theta) p(x_j|\theta)}{p(c=A) \cdot \sum_\theta p^A(\theta) p(x_j|\theta) + p(c=B) \cdot \sum_\theta p^B(\theta) p(x_k|\theta)}
\end{aligned}$$

Similarly, we have:

$$p(c=B|\boldsymbol{r} = \boldsymbol{r}^A_{\text{P},j} \vee \boldsymbol{r}^B_{\text{P},k}) = \frac{p(c=B) \cdot \sum_\theta p^B(\theta) p(x_k|\theta)}{p(c=A) \cdot \sum_\theta p^A(\theta) p(x_j|\theta) + p(c=B) \cdot \sum_\theta p^B(\theta) p(x_k|\theta)}$$

Denote

$$S_j^A \coloneqq p(c=A)\sum_\theta p^A(\theta)p(x_j|\theta)$$
$$S_k^B \coloneqq p(c=B)\sum_\theta p^B(\theta)p(x_k|\theta)$$

Define the observation-dependent context frequency for observing data samples coming from $\boldsymbol{r}_{\mathrm{P},j}^A$ or $\boldsymbol{r}_{\mathrm{P},k}^B$:

$$\rho_{jk}^A \coloneqq p(c=A|\boldsymbol{r}=\boldsymbol{r}_{\mathrm{P},j}^A \vee \boldsymbol{r}_{\mathrm{P},k}^B) = S_j^A/(S_j^A+S_k^B)$$
$$\rho_{jk}^B \coloneqq p(c=B|\boldsymbol{r}=\boldsymbol{r}_{\mathrm{P},j}^A \vee \boldsymbol{r}_{\mathrm{P},k}^B) = S_k^B/(S_j^A+S_k^B)$$

Now let us denote the context-independent likelihood decoder output as $g_L(\boldsymbol{r}_{\mathrm{P},j}^A) = g_L(\boldsymbol{r}_{\mathrm{P},k}^B) \equiv \ell_{jk}(\theta)$. The context-dependent posterior distribution given the corresponding context prior is given by:

$$q_{L,j}^A(\theta) = \frac{p^A(\theta)\ell_{jk}(\theta)}{\sum_{\theta'}p^A(\theta')\ell_{jk}(\theta')} = \frac{p^A(\theta)\ell_{jk}(\theta)}{Z_j^A[\ell_{jk}(\theta)]}$$
$$q_{L,k}^B(\theta) = \frac{p^B(\theta)\ell_{jk}(\theta)}{\sum_{\theta'}p^B(\theta')\ell_{jk}(\theta')} = \frac{p^B(\theta)\ell_{jk}(\theta)}{Z_k^B[\ell_{jk}(\theta)]}$$

where $Z_j^A[\ell_{jk}(\theta)]$ and $Z_k^B[\ell_{jk}(\theta)]$ are normalization constants dependent on $\ell_{jk}(\theta)$, defined as:

$$Z_j^A[\ell_{jk}(\theta)] \coloneqq \sum_\theta p^A(\theta)\ell_{jk}(\theta)$$
$$Z_k^B[\ell_{jk}(\theta)] \coloneqq \sum_\theta p^B(\theta)\ell_{jk}(\theta)$$

Under cross-entropy loss, we want $\ell_{jk}(\theta)$ (and its associated posteriors $q_{L,j}^A(\theta)$ and $q_{L,k}^B(\theta)$) to minimize:

$$\begin{aligned}
&\min_{\ell_{jk}(\theta)} \Big\{\rho_j^A H(p^A(\theta|x_j), q_{L,j}^A(\theta)) + \rho_k^B H(p^B(\theta|x_k), q_{L,k}^B(\theta))\Big\} \\
&= \min_{\ell_{jk}(\theta)} \Big\{ -\sum_\theta \Big[\rho_j^A \frac{p^A(\theta)p(x_j|\theta)}{\sum_{\theta'}p^A(\theta')p(x_j|\theta')} \log \frac{p^A(\theta)\ell_{jk}(\theta)}{Z_j^A[\ell_{jk}]} + \\
&\qquad \rho_k^B \frac{p^B(\theta)p(x_k|\theta)}{\sum_{\theta'}p^B(\theta')p(x_k|\theta')} \log \frac{p^B(\theta)\ell_{jk}(\theta)}{Z_k^B[\ell_{jk}]}\Big]\Big\}
\end{aligned} \tag{7}$$

Define

$$\mu_j^A(\theta) \coloneqq \rho_j^A p^A(\theta|x_j) = \rho_j^A \frac{p^A(\theta)p(x_j|\theta)}{\sum_{\theta'}p^A(\theta')p(x_j|\theta')} = \frac{p(c=A)p^A(\theta)p(x_j|\theta)}{S_j^A+S_k^B}$$
$$\mu_k^B(\theta) \coloneqq \rho_k^B p^B(\theta|x_k) = \rho_k^B \frac{p^B(\theta)p(x_k|\theta)}{\sum_{\theta'}p^B(\theta')p(x_k|\theta')} = \frac{p(c=B)p^B(\theta)p(x_k|\theta)}{S_j^A+S_k^B}$$

Note

$$\sum_\theta \mu_j^A(\theta) = \frac{p(c=A)\sum_\theta p^A(\theta)p(x_j|\theta)}{S_j^A+S_k^B} = \rho_j^A$$
$$\sum_\theta \mu_k^B(\theta) = \frac{p(c=B)\sum_\theta p^B(\theta)p(x_k|\theta)}{S_j^A+S_k^B} = \rho_k^B$$

The cross-entropy loss in Eq. 7 becomes:

$$\begin{aligned}
L(\ell_{jk}(\theta)) = -\sum_{\theta} \Big[& \mu_j^A(\theta) \cdot \Big(\log p^A(\theta) + \log \ell_{jk}(\theta) - \log Z_j^A[\ell_{jk}(\theta)] \Big) + \\
& \mu_k^B(\theta) \cdot \Big(\log p^B(\theta) + \log \ell_{jk}(\theta) - \log Z_k^B[\ell_{jk}(\theta)] \Big) \Big] \\
= -\Big\{ & \sum_{\theta} \Big[\mu_j^A(\theta) \log p^A(\theta) + \mu_k^B(\theta) \log p^B(\theta) \Big] \\
& + \sum_{\theta} \Big[\big(\mu_j^A(\theta) + \mu_k^B(\theta) \big) \cdot \log \ell_{jk}(\theta) \Big] \\
& - \Big[\sum_{\theta} \mu_j^A(\theta) \Big] \cdot \log Z_j^A[\ell_{jk}(\theta)] - \Big[\sum_{\theta} \mu_k^B(\theta) \Big] \cdot \log Z_k^B[\ell_{jk}(\theta)] \Big\}
\end{aligned}$$

Note $L(\alpha \ell) = L(\ell),\ \forall \alpha > 0$. Therefore ℓ^* that minimizes L is determined up to a multiplicative constant. The above minimization happens at the critical point $\ell_{jk}^*(\theta)$ where $\frac{\partial L}{\partial \ell_{jk}^*(\theta)} = 0,\ \forall \theta$.

Before proceeding to find the minimum, let us first find

$$\begin{aligned}
\frac{\partial}{\partial \ell_{jk}(\theta)} Z_j^A[\ell_{jk}(\theta)] &= \frac{\partial}{\partial \ell_{jk}(\theta)} \Big\{ \sum_{\theta'} p^A(\theta') \ell_{jk}(\theta') \Big\} = p^A(\theta) \\
\frac{\partial}{\partial \ell_{jk}(\theta)} Z_k^B[\ell_{jk}(\theta)] &= \frac{\partial}{\partial \ell_{jk}(\theta)} \Big\{ \sum_{\theta'} p^B(\theta') \ell_{jk}(\theta') \Big\} = p^B(\theta)
\end{aligned}$$

To find the minimum, let us take the derivative with respect to $\ell_{jk}(\theta)$ and set it to zero:

$$\begin{aligned}
0 &= \frac{\partial L(\ell_{jk}(\theta))}{\partial \ell_{jk}(\theta)} \\
&= -\frac{\partial}{\partial \ell_{jk}(\theta)} \Big\{ \sum_{\theta} \Big[\mu_j^A(\theta) \log p^A(\theta) + \mu_k^B(\theta) \log p^B(\theta) \Big] \\
&\qquad + \sum_{\theta} \Big[\big(\mu_j^A(\theta) + \mu_k^B(\theta) \big) \cdot \log \ell_{jk}(\theta) \Big] \\
&\qquad - \Big[\sum_{\theta} \mu_j^A(\theta) \Big] \cdot \log Z_j^A[\ell_{jk}(\theta)] - \Big[\sum_{\theta} \mu_k^B(\theta) \Big] \cdot \log Z_k^B[\ell_{jk}(\theta)] \Big\} \\
&= -\Big\{ \frac{\mu_j^A(\theta) + \mu_k^B(\theta)}{\ell_{jk}(\theta)} - \frac{\Big[\sum_{\theta} \mu_j^A(\theta) \Big]}{Z_j^A[\ell_{jk}(\theta)]} \frac{\partial Z_j^A[\ell_{jk}(\theta)]}{\partial \ell_{jk}(\theta)} - \frac{\Big[\sum_{\theta} \mu_k^B(\theta) \Big]}{Z_k^B[\ell_{jk}(\theta)]} \frac{\partial Z_k^B[\ell_{jk}(\theta)]}{\partial \ell_{jk}(\theta)} \Big\} \\
&= -\Big\{ \frac{\mu_j^A(\theta) + \mu_k^B(\theta)}{\ell_{jk}(\theta)} - \frac{\rho_j^A}{Z_j^A[\ell_{jk}(\theta)]} p^A(\theta) - \frac{\rho_k^B}{Z_k^B[\ell_{jk}(\theta)]} p^B(\theta) \Big\}
\end{aligned}$$

Therefore minimization happens when (determined up to a multiplicative constant):

$$\begin{aligned}
\ell_{jk}^*(\theta) &= \frac{\mu_j^A(\theta) + \mu_k^B(\theta)}{\frac{\rho_j^A}{Z_j^A[\ell_{jk}^*]} p^A(\theta) + \frac{\rho_k^B}{Z_k^B[\ell_{jk}^*]} p^B(\theta)} \\
&= \frac{\rho_j^A p^A(\theta | x_j) + \rho_k^B p^B(\theta | x_k)}{\frac{\rho_j^A}{Z_j^A[\ell_{jk}^*]} p^A(\theta) + \frac{\rho_k^B}{Z_k^B[\ell_{jk}^*]} p^B(\theta)}
\end{aligned} \tag{8}$$

Since both $Z_j^A[\ell_{jk}^*]$ and $Z_k^B[\ell_{jk}^*]$ depend on $\ell_{jk}^*(\theta)$, Eq. 8 gives an implicit expression for $\ell_{jk}^*(\theta)$. The equation can be solved using fixed-point iteration starting with some initial guess for $\ell_{jk}^{(0)}(\theta) > 0$.

For instance:

$$
\begin{aligned}
&\text{Initialize } \ell_{jk}^{(0)}(\theta) \propto 1 \\
&\text{for } t = 0, 1, 2, \ldots : \\
&\qquad \text{compute } Z_j^{A,(t)}[\ell_{jk}^{(t)}] = \sum_\theta \ell_{jk}^{(t)}(\theta) p^A(\theta) \\
&\qquad\qquad\qquad Z_k^{B,(t)}[\ell_{jk}^{(t)}] = \sum_\theta \ell_{jk}^{(t)}(\theta) p^B(\theta) \\
&\qquad \text{update } \ell_{jk}^{(t+1)}(\theta) = \frac{\rho_j^A p^A(\theta|x_j) + \rho_k^B p^B(\theta|x_k)}{\frac{\rho_j^A}{Z_j^{A,(t)}[\ell_{jk}^{(t)}]} p^A(\theta) + \frac{\rho_k^B}{Z_k^{B,(t)}[\ell_{jk}^{(t)}]} p^B(\theta)} \\
&\text{Stop when } \ell_{jk}^{(t)}(\theta) \text{ converges (up to a multiplicative constant).}
\end{aligned}
$$

That is, after training, the best possible likelihood decoder output for data samples associated with $\boldsymbol{r}_{\mathrm{P},j}^A$ and $\boldsymbol{r}_{\mathrm{P},k}^B$ is as if the decoder was dividing a surrogate posterior utilizing the Bayes-optimal estimator of the context-dependent likelihood functions across context, in this case equals a weighted sum of ground-truth posteriors $\rho_j^A p^A(\theta|x_j) + \rho_k^B p^B(\theta|x_k)$ by a weighted sum of ground-truth context priors $\frac{\rho_j^A}{Z_j^A[\ell_{jk}^*]} p^A(\theta) + \frac{\rho_k^B}{Z_k^B[\ell_{jk}^*]} p^B(\theta)$.

The posterior of the best likelihood decoder output g_L^* given the corresponding context prior for $\boldsymbol{r}_{\mathrm{P},j}^A$ and $\boldsymbol{r}_{\mathrm{P},k}^B$ is:

$$
\begin{aligned}
q_{L,j}^{A*}(\theta) &= \frac{\ell_{jk}^*(\theta) p^A(\theta)}{Z_j^A[\ell_{jk}^*]} \\
q_{L,k}^{B*}(\theta) &= \frac{\ell_{jk}^*(\theta) p^B(\theta)}{Z_k^B[\ell_{jk}^*]}
\end{aligned}
$$

Hence, the cross-entropy loss for the best possible likelihood decoder can be computed as:

$$
\begin{aligned}
\mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i), q_{L,i}^{c*}(\theta))] &= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i)) + D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta))] \\
&= \mathbb{E}_{p(x_i,c)}[H(p^c(\theta|x_i))] + \mathbb{E}_{p(x_i,c)}\big[D_{KL}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta))\big] \\
&= \text{CE loss for the perfect posterior decoder} \\
&\quad + \mathbb{E}_{p(x_i,c)}\big[D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta))\big] \qquad (9)
\end{aligned}
$$

Information gap for a posterior-coding population $\Delta_{\mathrm{P}}^{\mathbf{info}}$ From equation 9, let us define $\Delta_{\mathrm{P}}^{\mathrm{info}}$, information gap for a posterior-coding population between a perfect posterior decoder and the best

possible likelihood decoder, as the difference in the expected cross-entropy loss of the two decoders:

$$
\begin{aligned}
\Delta_{\mathrm{P}}^{\text{info}} &:= \mathbb{E}_{p(x_i,c)}\big[D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta))\big] \\
&= \sum_{x_i,c} D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta)) \cdot p(x_i, c), \\
&\quad \text{since only } x_i \in \{(x_j, x_k)\} \text{ terms are nonzero, denote } \chi = \{(x_j, x_k)\} \\
&= \sum_{x_i \in \chi, c} D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta)) \cdot p(x_i, c) \\
&= \sum_{x_i \in \chi, c} D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{*}(\theta)) \cdot p(c) \cdot \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\
&= \sum_{x_i \in \chi} \sum_{c} p(c) D_{\mathrm{KL}}(p^c(\theta|x_i) \mid\mid q_{L,i}^{c*}(\theta)) \Big[\sum_{\theta} p(x_i|\theta) p^c(\theta)\Big] \\
&= \sum_{(x_j,x_k)} \Big\{ D_{\mathrm{KL}}(p^A(\theta|x_j) \mid\mid q_{L,j}^{A*}(\theta)) \cdot p(c = A) \Big[\sum_{\theta} p(x_j|\theta) p^A(\theta)\Big] \\
&\qquad\qquad + D_{\mathrm{KL}}(p^B(\theta|x_k) \mid\mid q_{L,k}^{B*}(\theta)) \cdot p(c = B) \Big[\sum_{\theta} p(x_k|\theta) p^B(\theta)\Big] \Big\}
\end{aligned}
$$