Scalars Are All You Need for Multimodal Inference

Kyle Cranmer

April 1, 2026

Abstract

Multimodal foundation models integrate heterogeneous data sources into unified representations. This is particularly powerful in scientific settings as these unified representations can be used as the input to downstream statistical inference tasks such as hypothesis testing, parameter estimation, or other forms of inference of underlying phenomena. Typically, contrastive learning or masked objectives are used to train task-agnostic representations followed by task-specific fine-tuning or the addition of task-specific heads. We propose an alternative framework for multimodal inference in which the task-specific training is applied to each modality individually. We show that the output representation for each modality can be reduced to a scalar function of the parameters and that these scalars can be aggregated additively without degrading performance. We show that, under conditional independence across modalities, this architecture enables permutation-invariant fusion, independent training of modality-specific encoders, and append-only extensibility. For regression tasks, predictive estimates can be extracted by passing this scalar function to a differentiable optimization layer. We show that in the limit of many observations, the resulting estimators achieve optimal properties. These results suggest that when representations are conditioned on the inferential context, high-dimensional multimodal embeddings are unnecessary: one scalar function per modality suffices.

1 Introduction

Multimodal foundation models are increasingly viewed as a pathway toward scientific discovery, where diverse data sources are combined to infer underlying laws governing complex systems. Notationally, we will use θ to represent a discrete index or the continuous parameters of a family of scientifically-motivated generative models for the data. In many such settings, this shared parameter vector θ determines the distribution of multiple observational modalities $x_1, \ldots, x_M$, each providing a complementary view of the same phenomenon.

A central challenge is how to combine information across modalities in a way that is both efficient and scalable. Existing approaches typically rely on high-dimensional shared embeddings or aggregation strategies, which must retain information relevant across many

potential downstream tasks (or relevant for predicting the parameter values θ that correspond to various hypotheses for the data generating process). While these approaches have demonstrated strong empirical performance, they raise fundamental questions about the nature and dimensionality of representations required for optimal inference.

In this work, we propose an alternative perspective in which multimodal inference is reduced to the aggregation of scalar, context-dependent evidence contributions. Rather than learning fixed high-dimensional representations that are independent of the downstream inferential task, we embrace the opposite extreme in which each modality has a context-dependent encoder that depends explicitly on the parameter θ under consideration. This leads to a formulation in which each modality contributes a single scalar-valued function, and these functions are combined additively. We show that this formulation leads to estimators $\hat{\theta}$ that achieve the best possible precision allowed by the data, while also providing a simple and extensible architecture for multimodal learning.

The main contributions of this work are:

- A formulation of multimodal inference based on *contextual scalar bottlenecks*
- A demonstration that additive scalar aggregation is sufficient for optimal inference under conditional independence
- A modular training strategy enabling independent optimization of modality-specific encoders
- An append-only architecture supporting seamless integration of new modalities and additional data

2 Multimodal Inference Architectures

2.1 Aggregation Task-Independent Embeddings

A common approach to multimodal learning proceeds in two stages [1, 2]. First, one constructs a latent representations for each modality

$$h_m = \phi_m(x_m),$$

where each h_m lives in a shared embedding space $\mathbb{R}^d$. Next, these representations are aggregated with a permutation-invariant function such as a sum in the case of DeepSets [3] or an attention mechanism in the case of a transformer [4]:

$$h_{\text{tot}} = \text{Aggregate}(\{h_1, \ldots, h_M\}).$$

Ref. [5] provides a unifying view of these approaches. Note that in models such as 4M, individual instances x_i are mapped into a set or sequence of tokens [6] instead of a single

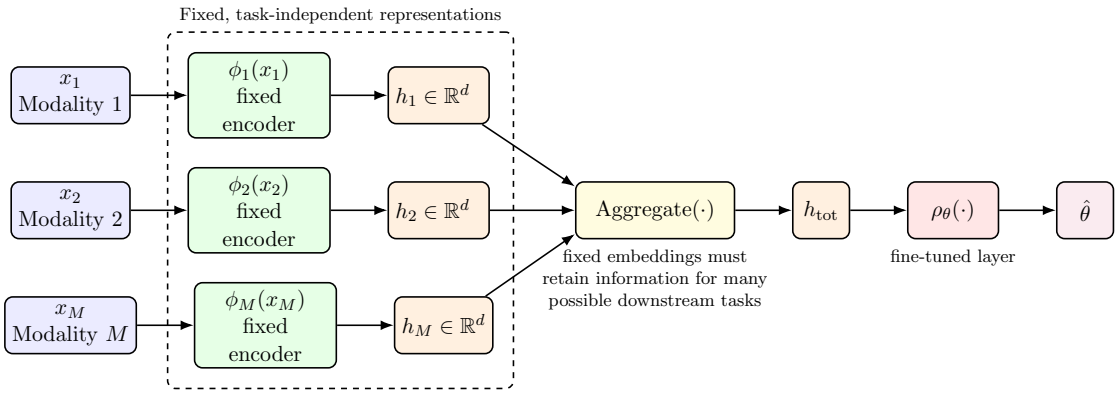

Figure 1: Multimodal inference with fixed, task-independent embeddings.

embedding vector, but this does not change the fact that the aggregation operation acts over a set of representations living in a shared embedding space.

Finally, when addressing a downstream task such as regression on the parameter θ, one either fine tunes or adds a dedicated task-specific head ρ_θ (the subscript θ is used to indicate that this function is task-specific). While the embeddings may be trained with self-supervised learning (either contrastive [7, 8, 9, 10] or masked modeling objecties [11, 12, 13, 14]), the task-specific head or fine tuning is usually trained in a supervised way using a loss function that has access to pairs $\{(x_{m,i}, \theta_i)\}$. The shared representation is then used to produce an estimate:

$$\hat{\theta} = \rho_\theta(h_{\text{tot}}).$$

Figure 1 illustrates the full pipeline in the case of fixed backbone and a task-specific head.

These approaches aim to produce general-purpose representations that are useful across a wide range of tasks. However, because the representation must be reusable, it may contain substantially more information than is required for a specific inferential objective. At the same time, there is no guarantee that the self-supervised pre-training actually retains all the information that is relevant to any given downstream task [15].

2.2 Context-Dependent Scalar Functions as Representations

Now we consider an alternative perspective and embrace the opposite extreme in which multimodal inference is reduced to the aggregation of context-dependent scalar functions. Rather than learning fixed, high-dimensional representations that are independent of the downstream inferential task, each modality has a fine-tuned, context-dependent encoder that depends explicitly on the unknown parameter θ under consideration. These encoders are trained or fine-tuned in a supervised way with access to pairs $(x_{m,i}, \theta_i)$, which may be obtained from simulators or labeled datasets. During training θ is known, but after

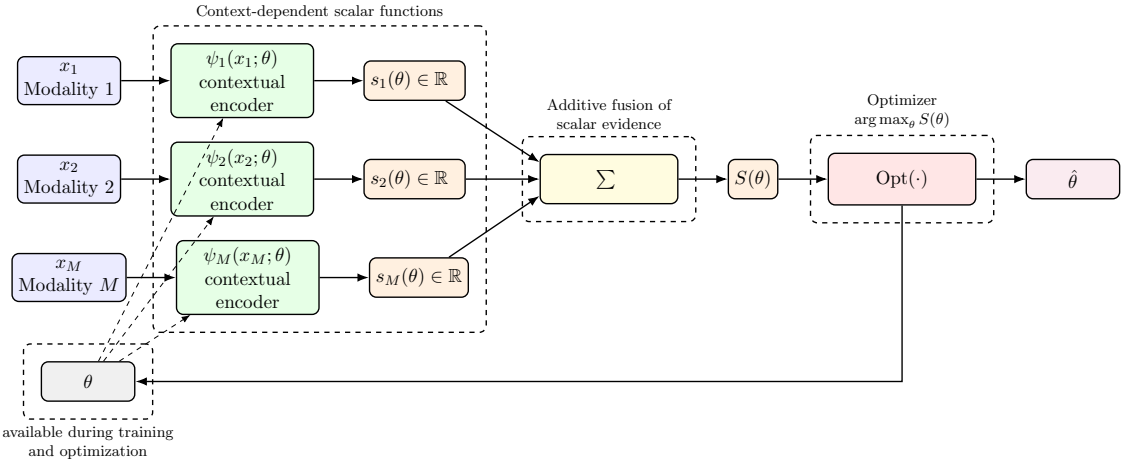

Figure 2: Multimodal inference with task-specific scalar functions as representations.

training θ is free to take on any value and can be interpreted as the remaining context needed to specify the downstream inference task. With this in mind it is natural to think of the embedding layer as taking x_m as input and returning a scalar function of θ as output:

$$s_m(\theta) = \psi_m(x_m; \theta) \in \mathbb{R}.$$

It may be useful to interpret the scalar function $s_m(\theta)$ as an infinite dimensional object (e.g. evaluating $s_m(\theta_j)$ for $j = 1, \ldots, d$ can be seen as constructing a representation in $\mathbb{R}^d$, and we can evaluate at arbitrarily many points.). That view helps reconcile how a modest scalar function can convey information for a wide range of tasks such as hypothesis tests between any two candidate data generating processes indexed by θ_1 and θ_2.

In practice, modality-specific encoders can be decomposed into two stages. First, let

$$z_m = f_m(x_m),$$

be a general-purpose representation, which can be learned using large-scale pre-training, including self-supervised approaches with unlabled data used to train models as in Figure 1. The second stage is task-specific and produces the scalar evidence value conditioned on θ:

$$s_m(\theta) = \psi_m(x_m; \theta) = g_m(z_m; \theta).$$

This separation allows the use of large, unlabeled datasets and self-supervised training strategies for representation learning, followed by fine-tuning using labeled data from the data generating processes to tailor the networks to the inference problem of interest.

2.2.1 Training the modality-specific encoders

For this approach to be successful, the scalar functions need to cary useful information for downstream inference tasks related to the parametric family of generative models indexed by θ. Each modality-specific encoder can be trained using pairs $(x_{m,i}, \theta_i)$, which may be obtained from simulators or labeled datasets, but what loss functions will lead to useful representations? It turns out that a few training objectives will result in encoders that have the desired properties. One such training objective is

$$\mathcal{L}_m = \mathbb{E}_{x,\theta}\left[-\log \exp(\psi_m(x_m;\theta))\right].$$

This objective encourages the encoder to assign higher scalar values to parameter values that to regions of the data with high density. In order for the training to converge, ψ_m must be normalized such that

$$\int \exp(\psi_m(x_m;\theta))\,\mathrm{d}x_m = C, \forall\theta$$

where $0 < C < \infty$ is an arbitrary real-valued θ-independent constant. This can be achieved, for instance, by using a conditional normalizing flow $q_m(x_m;\theta)$ that imposes normalization constraints architecturally and identifying $\psi_m(x_m;\theta) = \log q_m(x_m;\theta)$ [1]. In the Appendix, we show that this loss function leads to $\psi_m(x_m;\theta)$ with several important properties. In particular, when evaluated on samples $\{x_{m,i}\}$ resulting from the generative model indexed by θ^*, the function $s_m(\theta)$ peaks near the value θ^*. More formally, the expected value of the gradient of the function is zero when evaluated at θ^*.

An alternate strategy, not detailed here for brevity, is to setup a binary classification problem between two labeled samples of $(x_{m,i}, \theta_i)$ pairs. Minimizing the binary cross-entropy leads to a function that has the same key properties as those described above. More details of this strategy can be found in Ref. [17], though again that work is not in a multimodal setting.

Importantly, each modality can be trained independently, and no joint optimization across modalities is required. This has significant practical advantages.

2.2.2 Aggregation Strategy

Next we consider a permutation-invariant aggregation strategy reminiscent of the DeepSet approach: a sum of scalar functions

$$S(\theta) = \sum_{m=1}^{M} s_m(\theta).$$

The scalar function $S(\theta)$ plays an analogous role to h_{tot} in the task-independent approach. On one hand, the representation $s_m(\theta)$ is an infinite dimensional object that has

[1] a similar strategy was explored in [16], but not in a multimodal setting

the capacity to cary relevant to a family of inference problems indexed by θ. On the other hand, once θ has been specified, it allows for extreme compression of each modality to a one-dimensional representation without loss of task-relevant information.

2.2.3 Aggregation over Multiple Independent Observations

In many scientific settings, multiple i.i.d. observations are available for a given modality:

$$x_{m,1}, \ldots, x_{m,N_m}.$$

The same encoder can be applied to each observation:

$$S_m(\theta) = \sum_{i=1}^{N_m} \psi_m(x_{m,i}; \theta).$$

This leads to a combined scalar function representation

$$S(\theta) = \sum_{m} \sum_{i=1}^{N_m} \psi_m(x_{m,i}; \theta),$$

which naturally scales to arbitrary dataset sizes. This property is particularly important in scientific applications, where the number of observations may not be known in advance. In contrast, fixed-representation approaches often require careful architectural design to accommodate variable-sized inputs or large context windows lead to quadratic scaling for transformer based architectures.

2.2.4 Example usage: Regression

How is this scalar function used for downstream tasks? Let us consider the regression task of estimating θ based on data $\{x_1, \ldots, x_M\}$. If we assume the data result from the data generating process for true value of the parameter θ^*, then we know that $S(\theta)$ should peaks near θ^*. Thus in a regression setting, a natural prediction for θ is simply to find the maximizer:

$$\hat{\theta} = \arg\max_{\theta} S(\theta).$$

This optimization can be implemented using differentiable procedures, including unrolled optimization or implicit differentiation. Importantly, the modality-specific encoders do not require gradients through this optimization step, allowing representation learning and inference to be decoupled. The full pipeline is visualized in Figure 2.

In the Appendix, we show that in the case of many i.i.d. observations that the estimator $\hat{\theta}$ concentrates around the true value θ^*. We sketch the outline of a proof that the variance of this estimator saturates the Hammersley-Chapman-Robbins bound, making it optimal in that sense. [2]

[2]There is an analogous result in the case of hypothesis testing between two competing hypothetical

2.3 Modularity and Extensibility

The additive structure of the model enables a modular architecture in which new modalities can be incorporated without modifying existing components. If a new modality is introduced, its contribution can be added directly: $S_{M+1}(\theta) = S_M(\theta) + s_{M+1}(\theta)$.

This append-only property allows the system to grow as new data sources become available, while preserving previously learned components. This provides enormous flexibility that is practically relevant, enabling different scientific groups that specialize in individual data modalities to collaborate with a loose coupling and minimal coordination.

3 Conclusion

We have introduced a framework for multimodal inference based on contextual scalar function representations. [3] By conditioning modality-specific encoders on the parameter of interest, we reduce multimodal fusion to the additive aggregation of scalar evidence contributions. This perspective provides a simple and unified view of multimodal inference that contrasts with the prevailing emphasis on fixed, task-agnostic, high-dimensional vector embeddings.

The proposed approach offers several advantages. It is inherently permutation-invariant and accommodates variable numbers of modalities and observations. It supports independent training of modality-specific encoders, enabling scalable and modular learning. It also admits a natural append-only extension, allowing new modalities and data sources to be incorporated without retraining existing components. Through a two-stage encoder design, it further enables the combination of large-scale pre-training with task-specific fine-tuning.

Beyond these architectural benefits, the framework achieves strong statistical guarantees. The aggregated evidence function concentrates around the true parameter as more observations are incorporated, and the variance of the resulting estimator shrinks until it saturates the Hammersley-Chapman-Robbins bound, making it optimal in that sense. Taken together, these results suggest that when representations are conditioned on the inferential context, multimodal inference can be achieved using far simpler structures than those typically employed. In particular, one scalar function per modality is sufficient for optimal performance.

---

data generating processes labeled by θ_1 and θ_2 where one can simply consider the differences: $r(\theta_1, \theta_2) = S(\theta_1) - S(\theta_2)$. We leave the derivation of the properties of this approach to future work, but conjecture that coincides with the Bayes optimal classifier.

[3] In the final stages of preparing this manuscript, we became aware of related work that has a similar flavor, but does not employ neural networks [18, 19, 20, 21, 22, 23].

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[2] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey, 2024.

[3] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[5] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021.

[6] David Mizrahi, Roman Bachmann, Oguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[11] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[14] Tao Huang, Yanxiang Ma, Shan You, and Chang Xu. Learning mask invariant mutual information for masked image modeling. *arXiv preprint arXiv:2502.19718*, 2025.

[15] Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress- self-supervised learning and information theory: A review, 2023.

[16] K Cranmer and G Louppe. Unifying generative models and exact likelihood-free inference with conditional bijections. *J. Brief Ideas*, 2016.

[17] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

[18] Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

[19] C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

[20] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

[21] Alan Stuart, J Keith Ord, and Steven Arnold. Kendall's advanced theory of statistics. vol. 2a: Classical inference and the linear model. *Kendall's advanced theory of statistics. Vol. 2A: Classical inference and the linear model*, 1999.

[22] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 02 1933.

[23] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2):1554, 2011.

A Appendix A: Proof of Optimality

We sketch the argument that the proposed training procedure produces an evidence function whose maximizer achieves optimal precision in the limit of many observations.

We begin by considering the training objective for a single modality. Let x denote an observation and θ a parameter. The encoder produces a scalar function $s(x;\theta)$ and is trained using a conditional cross-entropy objective:

$$\mathcal{L} = \mathbb{E}_{(x,\theta)} \left[-\log \exp(s(x;\theta)) \right].$$

The minimizer of this objective over sufficiently expressive function classes is given (up to an additive function independent of θ) by

$$s^*(x;\theta) = \log p(x \mid \theta) + C(x),$$

where $C(x)$ does not depend on θ. This follows from the standard characterization of cross-entropy minimizers. Since $C(x)$ is independent of θ, it does not affect optimization over θ, and we may identify

$$s^*(x;\theta) \equiv \log p(x \mid \theta).$$

We now establish a key property of this function. Define the gradient

$$\nabla_\theta s^*(x;\theta) = \nabla_\theta \log p(x \mid \theta).$$

We show that its expectation vanishes at every θ:

$$\mathbb{E}_\theta \left[\nabla_\theta s^*(x;\theta) \right] = 0.$$

This follows from normalization of the probability density. Since

$$\int p(x \mid \theta)\, dx = 1,$$

differentiating with respect to θ gives

$$\int \nabla_\theta p(x \mid \theta)\, dx = 0.$$

Using

$$\nabla_\theta p(x \mid \theta) = p(x \mid \theta)\, \nabla_\theta \log p(x \mid \theta),$$

we obtain

$$\int p(x \mid \theta)\, \nabla_\theta \log p(x \mid \theta)\, dx = 0,$$

which implies

$$\mathbb{E}_\theta \left[\nabla_\theta s^*(x;\theta) \right] = 0.$$

Thus, at the true parameter, the expected slope of the scalar evidence function is zero. There is no systematic directional bias.

We now consider N independent observations. The aggregated evidence function is

$$S(\theta) = \sum_{i=1}^{N} s^*(x_i; \theta).$$

Its gradient is

$$\nabla_\theta S(\theta) = \sum_{i=1}^{N} \nabla_\theta s^*(x_i; \theta).$$

Taking expectations at the true parameter θ^* yields

$$\mathbb{E}_{\theta^*} \left[\nabla_\theta S(\theta) \right] \Big|_{\theta=\theta^*} = 0.$$

Thus, θ^* is a stationary point of the expected evidence function.

Next, we examine local behavior around θ^*. A second-order expansion gives

$$S(\theta) \approx S(\theta^*) - \frac{1}{2}(\theta - \theta^*)^\top H (\theta - \theta^*),$$

where

$$H = -\sum_{i=1}^{N} \nabla_\theta^2 \log p(x_i \mid \theta^*).$$

The curvature scales linearly with N, implying that the evidence function becomes increasingly sharply peaked as more observations are incorporated.

We now relate this behavior to fundamental limits on estimation. The Hammersley–Chapman–Robbins bound states that for any estimator $\hat{g}$,

$$\mathrm{Var}_\theta[\hat{g}] \geq \sup_{\theta' \neq \theta} \frac{\left(\mathbb{E}_{\theta'}[\hat{g}] - \mathbb{E}_\theta[\hat{g}] \right)^2}{\chi^2(\mu_{\theta'}; \mu_\theta)}.$$

This bound quantifies how distinguishable nearby parameter values are.

In the present setting, the additive structure of $S(\theta)$ implies:

- Differences $S(\theta') - S(\theta)$ grow linearly with N
- Curvature grows linearly with N

- The spread of the maximizer $\hat{\theta}$ shrinks at rate $1/N$

These properties match those required to saturate the bound in the large-sample limit.

In essence, the estimator obtained by maximizing the aggregated evidence function asymptotically achieves the best possible precision permitted by the data.

□