Effect of Geometric Complexity on Intuitive Model Selection

Effect of Geometric Complexity on Intuitive Model Selection

Eugenio Piasini[†], Vijay Balasubramanian, and Joshua I. Gold

Computational Neuroscience Initiative, University of Pennsylvania

Abstract

Occam's razor is the principle stating that, all else being equal, simpler explanations for a set of observations are to be preferred to more complex ones. This idea can be made precise in the context of statistical inference, where the same quantitative notion of complexity of a statistical model emerges naturally from different approaches based on Bayesian model selection and information theory. The broad applicability of this mathematical formulation suggests a normative model of decision-making under uncertainty: complex explanations should be penalized according to this common measure of complexity. However, little is known about if and how humans intuitively quantify the relative complexity of competing interpretations of noisy data. Here we measure the sensitivity of naive human subjects to statistical model complexity. Our data show that human subjects bias their decisions in favor of simple explanations based not only on the dimensionality of the alternatives (number of model parameters), but also on finer-grained aspects of their geometry. In particular, as predicted by the theory, models intuitively judged as more complex are not only those with more parameters, but also those with larger volume and prominent curvature or boundaries. Our results imply that principled notions of statistical model complexity have direct quantitative relevance to human decision-making.

1 Introduction

Occam's razor is a philosophical prescription to keep our models of the world as simple as possible. But does naive human decision making under uncertainty follow this prescription, and if it does, how strong is the preference for simple models? To ask these questions we must first provide a normative reference point for human behavior, by understanding from first principles what it means for a model to be simple or complex and how strongly should an optimal decision-making process be affected by a simplicity bias.

[†]Correspondence: epiasini@sas.upenn.edu

Many statistical learning techniques are governed by hyperparameters that can be tuned to control some notion of complexity of the underlying model. For instance, in regression, regularization techniques such as LASSO enforce an adjustable-strength constraint on the space of desired solutions [13]. In unsupervised learning, clustering methods often allow to control the desired level of granularity by specifying in advance the number of clusters [6]. More generally, in many probabilistic model selection settings, one compares several models with different number of parameters. In this case, the well known Akaike and Bayesian Information Criteria (AIC, BIC) state that the best model for a certain set of observations is the one that maximizes the log likelihood of the data minus some penalty that depends on the complexity of the model, measured by the number of parameters it contains [13]. A common thread among these examples is the idea of trading off some goodness of fit on the training data in exchange for model simplicity. A bias towards simplicity is desirable because it improves the performance of the model on unseen data, or because it makes it more interpretable, instantiating Occam's razor in concrete statistical practice. However, the examples above highlight that there are multiple possible definitions of complexity, some of which may be applicable only in relatively narrow contexts.

To overcome this difficulty, we draw on the theory of Bayesian model selection [16, 12, 20, 3]. This framework offers a principled definition of model complexity that is applicable across multiple settings, and makes complexity commensurable with goodness of fit by placing the two quantities on the same scale. Here, "model" always refers to a parametric family of probability distributions. For instance, the set of all Binomial probability distributions with n fixed to a certain value and p unknown, $0 \leq p \leq 1$, is a one-parameter model. Starting from a set of observations $X = \{x_n\}$ and a (finite) set of models, with a choice of prior probability over models $p(\mathcal{M})$ and over the parameters ϑ characterising each of them $p(\vartheta|\mathcal{M})$, by applying Bayes' theorem and marginalising over model parameters one can invert the likelihood function $p(X|\mathcal{M}, \vartheta)$ to yield a posterior distribution over models given the data, $p(\mathcal{M}|X)$. One can then select the model that maximises the posterior. It can be shown [3] that assuming an uninformative prior for the model parameters ϑ leads to an expression for the model posterior that generalizes the BIC. When the number of data points N is large enough, the (log) posterior probability of a model can be approximated by an expression consisting of the maximum log likelihood of the data under that model, plus a number of penalty factors which posses an elegant geometrical interpretation. The expression, known as Fisher Information Approximation

(FIA) is

$$
\begin{aligned}
-\log p(\mathcal{M}|X) = &-\log p(X|\hat{\vartheta}) + \frac{d}{2}\log\frac{N}{2\pi} \\
&+ \log \int \mathrm{d}^d\,\vartheta \sqrt{\det g(\vartheta)} \\
&+ \frac{1}{2}\log\left[\frac{\det h(X;\hat{\vartheta})}{\det g(\hat{\vartheta})}\right] + \dots \\
=: &L + D + V + R + \dots
\end{aligned}
\tag{1}
$$

where $\hat{\vartheta}$ is the parameter value that maximises the likelihood of the data under $\mathcal{M}$, d is the dimensionality of $\mathcal{M}$ (number of parameters), g_{ab} and h_{ab} are respectively the Fisher Information and the Observed Fisher Information [7], and the remainder $(\dots)$ collects terms that get smaller when N grows larger. We will call the terms of the FIA *likelihood* (L), *dimensionality* (D), *volume* (V) and *robustness* (R), respectively. It can be shown [3] that the volume term actually measures the volume of the model, seen as a statistical manifold in the sense of information geometry [2]. The robustness term is related to the shape of the statistical manifold in the vicinity of the maximum likelihood point, and more specifically to its embedding curvature in data space [24].

By direct application of the rules of Bayesian statistics, one then arrives at the conclusion that more complex models should be penalized, and the correct measure of complexity and its exchange rate with goodness of fit depends not only on the dimensionality of the model (as in the BIC, which corresponds to only using the first two terms of the FIA), but also on its finer geometrical properties. Interestingly, analogous expressions can be obtained by distinct arguments based on information theory, using the Minimum Description Length principle [25, 11] or the Predictive Information framework [5].

The elegance of this result, and the fact that the same prescription emerges from distinct approaches in information theory, make it a good candidate for a general notion of statistical complexity upon which to build a normative model of decision-making under uncertainty in rational observers. It is natural to ask if human subjects exhibit a preference for simpler models, and if they do, to quantitatively compare their intuitive measurement of complexity to the prescriptions of the theory.

1.1 Related work

Some evidence for a simplicity bias in human decision-making can be found in the existing literature. Johnson, Jin, and Keil [17] showed that, in a model selection task, subjects prefer simpler models (characterised as those with fewer parameters) when the likelihood of the data is approximately the same across the models being compared. Genewein and Braun [10] also studied a model selection task, providing more solid theoretical grounding in Bayesian model

selection theory. However, that study also focused primarily on qualitative preferences in equal-likelihood conditions (showing that indeed subjects possess a bias towards simple models), stopping short of a quantitative evaluation of the strength of the bias. To our knowledge, our work is the first attempt to: 1) precisely quantify the tradeoff between simplicity and goodness of fit in human decision-making; 2) investigate the behavioral relevance of geometrical complexity; and 3) consider the individual impact of the model features captured by the terms of the FIA, including the effect of a novel form of penalty that can emerge for models with boundaries.

2 Methods

2.1 Psychophysics

We designed a visual psychophysics experiment to probe human subjects' sensitivity to statistical model complexity. The experiment is based on a two-alternative forced-choice task designed as described below. Detailed preregistration documents for the experiments, including design, sampling and analysis rationale, code for running the task, experimental stimuli, and a snapshot of the core libraries developed to analyze the data are available at [22, 23].

The subjects were shown two curves and 10 dots on a screen (see examples in Figure 1). One curve was located in the upper half of the screen, the other in the bottom half. The curves represent two parametric statistical models of the form

$$p(x|t) = \frac{1}{\sqrt{2\pi}} \exp\left[-(x-\mu(t))^2/2\right]$$

where x is a location on the 2D plane visualized on the screen and $\mu(t)$ is a parametrization of the curve. In other words, the curves represent Gaussians of unit isotropic variance whose mean μ can be located at any point along them. The dots shown to the subjects were sampled iid from one of the two models, selected at random with uniform probability. The location of the true mean of the Gaussian generating the dots (i.e., the value of t in the expression above) was randomly sampled from Jeffrey's prior for the selected model [15]. All dots shown within a trial come from the same distribution (same model and same true mean). The subjects had to report which curve (model) the dots are more likely to come from. They did so by pressing the "up" or "down" keys on their keyboard to select the curve in the upper or lower part of the screen.

We designed four variants of the task, each of which asked the subjects to make a selection between two models. The model pairings differed across task variants, and are illustrated in Figure 1. Each model pairing is designed to study primarily a different term of the FIA: dimensionality for the "point" pairing, boundary for "vertical" (we defer the formal introduction of the boundary term until the next section), volume for "horizontal", and robustness for "rounded". The models in the "point" task variant have different dimensionality ($d = 0$ for

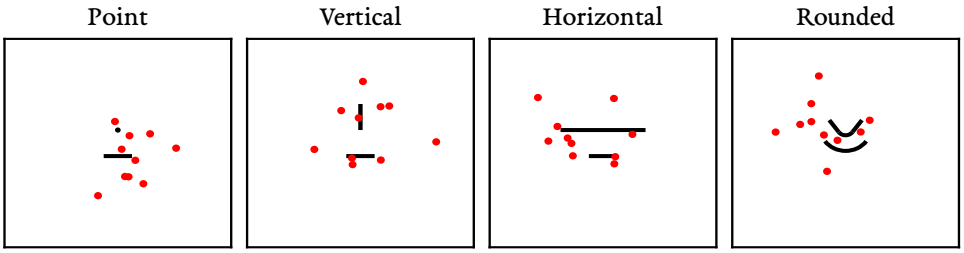

Figure 1: Task types with corresponding names. Each panel shows an example trial for one of the four task types in the experiment. The model manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ are drawn in black, and 10 points x sampled from a probability distribution contained in either $\mathcal{M}_1$ or $\mathcal{M}_2$ are shown in red. Given a visual stimulus similar to one of these panels, the subjects have to report which model ($\mathcal{M}_2$ or $\mathcal{M}_2$) is more likely to have generated the data.

the point and $d = 1$ for the line)[1]. In the "horizontal" variant, the models have the same dimensionality but different volume (length). In the "rounded" variant, the models have the same dimensionality and volume, but their curvature is such that one of them bends away from the region of data space which is more likely to contain ambiguous stimuli, whereas the other bends around it (and therefore the robustness term for these models has opposite sign for data points that fall in that region). Finally, in the "vertical" variant, the models have the same dimensionality and volume, and are both flat so that their robustness terms are always identically zero; however, they are oriented such that the lower endpoint (boundary) of the vertically oriented model is the closest point on the model to the ambiguous (equal-likelihood) region of data space located at the midpoint between the two models. Therefore, if some data falls within that region, assigning that data to the vertically oriented model will incur a penalty due to the boundary effect.

A single run of the task consisted in a brief tutorial followed by 500 trials, divided in 5 blocks of 100 trials each. In each trial, the chosen curve pairing was presented, randomly flipped vertically. At the end of each block, the subject received feedback on their overall performance during that block. Subjects received a fixed compensation for taking part in the experiment.

We ran the experiment on the online platform Pavlovia (`pavlovia.org`). For each task type we collected data from at least 50 subjects who passed a pre-established performance threshold: 60% correct for the "rounded" task variant and 70% correct for the other variants, as reported on the preregistration docu-

[1]In a similar way, one could define a two-dimensional model represented by a 2D area on the screen. This approach would be useful to provide an additional evaluation point for the dependence of the simplicity bias on model dimensionality. However, unlike a 0D or 1D model, a 2D model in a 2D data space will always suffer from boundary effects for data falling anywhere outside the model manifold. Therefore, because one primary goal of this study was to disentangle the distinct contributions of the models' different geometrical features to the simplicity bias, we only use 0D and 1D models.

ments [22, 23]. We discarded the data collected from all other subjects. These exclusion rules led to a final dataset containing 52 subjects for the "rounded" task variant, and exactly 50 subjects for each of the other task variants.

2.2 Penalty term for model boundaries

Most of the models used in the experiments have bounded parameter spaces. For instance, the base model is parametrized by one parameter t that is subject to the constraint $0 \leq t \leq 1$. The conditions $t = 0$ and $t = 1$ are mapped to the endpoints of the segment representing the model in data space in Figure 1. Having models with such boundaries is an issue for the applicability of the FIA, because one of the hypotheses underlying the derivation of Equation 1 is that $\hat{\vartheta}$ must be in the interior of the parameter space, and this assumption can easily break down in presence of models with bounded parameter spaces. To solve this issue, we extended the FIA to deal with the simple case of a linear boundary in parameter space (see Appendix A). When the maximum-likelihood point is on the edge of the parameter space, an additional term, which we indicate with the symbol S, appears in the FIA:

$$S = \frac{1}{2} \log \frac{N}{2\pi} + \log \left[2\pi \|l\|_{\Delta} \right] \tag{2}$$

where

$$l_a = -\frac{1}{N} \sum_i \frac{\partial}{\partial \vartheta_a} \log p(x_i|\vartheta)$$

is minus the empirical average of the score vector (log-likelihood gradient), and Δ is the inverse of the observed Fisher information:

$$\Delta = h^{-1} \qquad , \qquad h_{ab} = -\frac{1}{N} \sum_i \frac{\partial^2}{\partial \vartheta_a \partial \vartheta_b} \log p(x_i|\vartheta)$$

Equation 2 shows that the penalty associated to being at the boundary of parameter space corresponds to increasing the parameter dimensionality by one, plus a term that depends on the norm of the log-likelihood gradient (the gradient is not zero at the maximum likelihood point, precisely because we are on the boundary of the optimization domain). For a broad class of models, the second term can be shown to measure the degree of model misspecification induced by the existence of the boundary [24].

2.3 Comparison between subject behavior and Bayesian ideal observer

In our experimental scenario, the theory of Bayesian model selection applies directly. Given two models $\mathcal{M}_1$ and $\mathcal{M}_2$, assuming a flat prior over models

$p(\mathcal{M}_1) = p(\mathcal{M}_2) = 1/2$ and an uninformative (Jeffrey's) prior over the parameters of each model, when N is sufficiently large the log posterior ratio for $\mathcal{M}_1$ over $\mathcal{M}_2$ can be written

$$\begin{aligned}\log \frac{p(\mathcal{M}_1|X)}{p(\mathcal{M}_2|X)} &= \log \frac{p(\mathcal{M}_1|X)}{1-p(\mathcal{M}_1|X)} \\ &\simeq (L_2 - L_1) + (D_2 - D_1) + (S_2 - S_1) + (V_2 - V_1) + (R_2 - R_1)\end{aligned} \tag{3}$$

where L_i, D_i, etc represent the FIA terms for model i.

This expression suggests a very simple normative model for subject behavior. Equation 3 determines the probability of reporting $\mathcal{M}_1$ for an ideal Bayesian observer performing probability matching. We can then compare subject behavior to the normative prescription by allowing subjects to have distinct sensitivities to the various terms of the FIA:

$$\begin{aligned}\log \frac{p(\text{report } \mathcal{M}_1|X)}{p(\text{report } \mathcal{M}_2|X)} &= \alpha + \beta_L(L_2 - L_1) + \beta_D(D_2 - D_1) + \\ &\quad + \beta_S(S_2 - S_1) + \beta_V(V_2 - V_1) + \beta_R(R_2 - R_1)\end{aligned} \tag{4}$$

where α and β are free parameters: α captures any fixed bias, β_L the sensitivity to differences in maximum likelihood, β_D the sensitivity to differences in dimensionality, and so on.

2.4 Data analysis

We fitted the model expressed by Equation 4 to subject behavior using a hierarchical, Bayesian logistic regression scheme:

$$\nu_\alpha, \nu_L, \ldots, \nu_R \sim 1 + \text{Exponential}(29) \tag{5}$$

$$\mu_\alpha, \mu_L, \ldots, \mu_R \sim \text{Normal}(0, 3) \tag{6}$$

$$\sigma_\alpha, \sigma_L, \ldots, \sigma_R \sim \text{Exponential}(3) \tag{7}$$

$$\alpha_i \sim \text{StudentT}(\nu_\alpha, \mu_\alpha, \sigma_\alpha) \tag{8}$$

$$\beta_{L,i} \sim \text{StudentT}(\nu_L, \mu_L, \sigma_L) \tag{9}$$

$$\vdots \tag{10}$$

$$\beta_{R,i} \sim \text{StudentT}(\nu_R, \mu_R, \sigma_R) \tag{11}$$

$$C_{i,t} \sim \text{Bernoulli}\Big(\text{logit}^{-1}\Big(\text{lpr}\big(\alpha_i, \beta_{L,i}, \beta_{D,i}, \beta_{S,i}, \beta_{V,i}, \beta_{R,i}, X_{i,t}\big)\Big)\Big) \tag{12}$$

where $C_{i,t}$ is the choice made by subject i on trial t, $X_{i,t}$ is the sensory stimulus on that same trial, lpr is the log posterior ratio defined by Equation 4, α_i is the bias for subject i, $\beta_{L,i}$ is the likelihood sensitivity of that same subject, and so on for the other sensitivity parameters. The bias and sensitivity parameters describing each subject are modeled as independent samples from a population-level Student-T probability distribution characterized by a certain shape (ν),

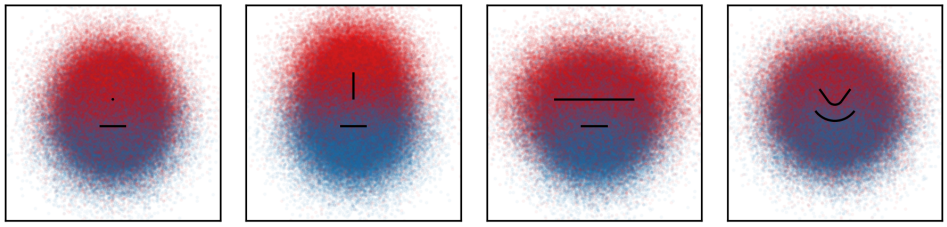

Figure 2: Overview of experimental data. Each panel overlays all stimuli shown to all subjects that performed a given task type. All 10 dots shown on a given trials are colored red if the subject reported "up" or blue if they reported "down" on that trial. Note that the actual location of the model manifolds and the stimuli were flipped vertically in roughly 50% of trials (see Methods), and have been counter-flipped in this plot for visualization purposes.

location (μ) and scale (σ). The priors assumed over these population-level parameters are standard weakly informative priors [9, 18], and broader or flat priors lead to similar results to those presented below. The model was implemented in PyMC3 [26], and inference was performed by sampling from the posterior for the parameters given the experimental data $\{C_{i,t}, X_{i,t}\}$ using the No-U-Turn Sampler algorithm [14, 4]. Further technical details on the inference procedure can be found in Appendix B.

3 Results

In our experiment, a simplicity bias would manifest by shifting the psychometric indifference point away from the simpler alternative. In other words, given a sensory stimulus such as those in Figure 1, a subject with simplicity bias would not always assign the red dots simply to the model that is, on average, closer to the dot cloud. They would instead trade off some of the goodness of fit of the models (in this case the geometrical distance) against some measure of simplicity. For instance, in the "point" task type (Figure 1, left), for the subject to choose the 1-dimensional model (the line) over the 0-dimensional one (the point), it would not be enough for the dot cloud to be on average closer to the line than to the point, but the difference in distance would have to be larger than a certain nonzero amount. The value of this critical difference is controlled by the exact nature of the tradeoff operated by the subject between simplicity and goodness of fit, or in other words the "exchange rate" between these two desirable objectives.

An overview of the experimental data collected is shown in Figure 2. A qualitative inspection of the figure already suggests the existence of a simplicity bias like the one just described. For instance, in the first panel on the left ("point" task type), the transition from red to blue is located further down than the vertical midpoint between the two models, suggesting that subjects tended to choose

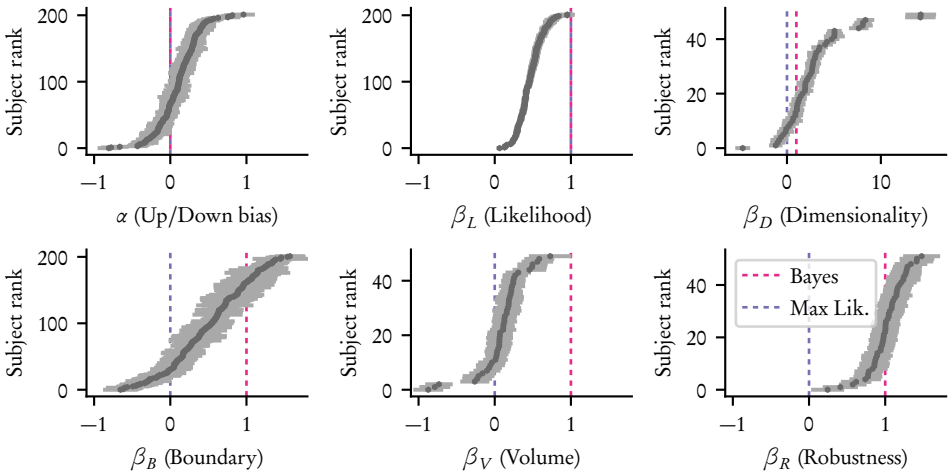

Figure 3: Subject-level estimates of sensitivity to the terms in Equation 4. Top left panel, dark gray dots: posterior mean $\mathbb{E}[\alpha_i]_{p(\alpha_i|X,C)}$ for the up/down bias of individual subjects. Light gray bars: standard deviation of the posterior distribution for the same parameters. Subjects are ranked based on the mean posterior. All other panels: same as the top left panel, for the β sensitivity parameters in Equation 4. Dashed lines: reference value of the parameters for the ideal Bayesian observer described by Equation 3 (magenta) and a "maximum likelihood" observer that disregards model complexity and selects models only based on distance from the data (purple). Note that number of dots (subjects) differs across panels because three of the regression parameters (β_D, β_V and β_R) can only be estimated for the subjects that performed a specific variant of the task (the "point", "horizontal", and "rounded" variant respectively). By contrast, α, β_L and β_S can be estimated for all subjects.

the point more often than the line for stimuli that were roughly equidistant from either.

We quantified these effects using the formal framework of Bayesian model selection and compared them to those predicted by the ideal observer. In Figure 3 we report the mean and standard deviation of the posterior estimates for the sensitivity of individual subjects to the FIA terms (the α_i and β_i parameters in Equation 4). These estimates show that most subjects possess a bias in favor of simple models, even though the strength of the bias is fairly heterogeneous across the population (this hypothesis was also tested with a formal model comparison procedure, using the Widely Applicable Information Criterion [21] — see Appendix B.3). We also note that the strength of the bias exhibited systematic differences in scale between the different terms of the FIA: for instance, the bias towards models with smaller dimensionality (Figure 3, top right panel) can be much stronger than the bias towards models with a smaller volume (bottom middle panel).

We can get a better idea of these global properties of the estimated parameters by studying the population level parameters μ_α, μ_L, μ_D,... (Equation 6), which

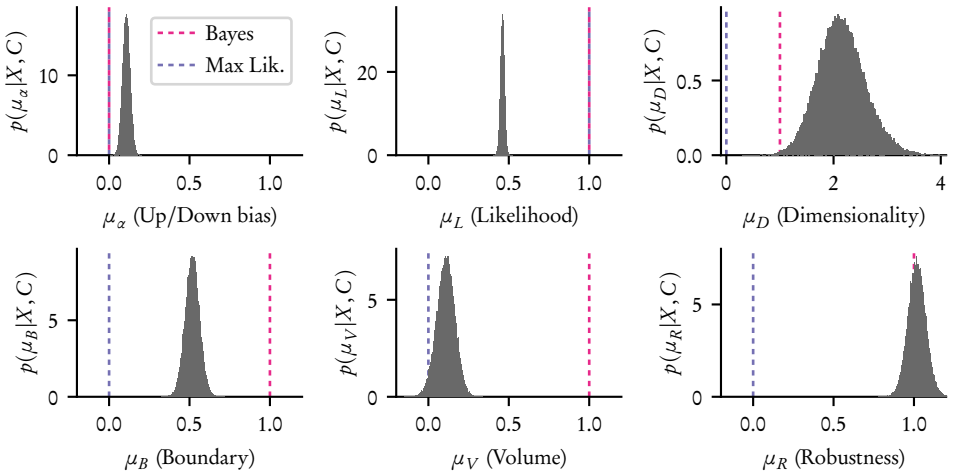

Figure 4: Population-level estimates of sensitivity to the terms in Equation 4, reported as the full posterior distribution for the μ_α and μ_β parameters, conditional on the observed experimental data $\{X_{i,t}, C_{i,t}\}$. Dashed magenta and purple lines are the reference values for the ideal Bayesian and maximum-likelihood observers, respectively, as in Figure 3.

parametrize the location (the mean) of the distributions from which the subject-level sensitivities are sampled. We report the full posterior distribution of the μ parameters in Figure 4. These analyses indicate that the subjects were sensitive to model complexity in general as well as to all terms of the FIA taken individually, and that some model features contributing to the FIA (dimensionality and shape) seemed to affect subject behavior more strongly than others (volume and presence of boundaries).

4 Discussion

Occam's razor is a ubiquitous principle in statistics and learning theory that we can express in a rigorous and elegant way using Bayesian model-selection theory. We sought to build on this solid theoretical grounding by using it to understand if and how Occam's razor applies to human decision-making under uncertainty.

Specifically, we have formulated a class of psychophysical tasks that allowed us to probe this hypothesis directly and quantitatively. A critical technical step in doing so was the extension of the existing theory surrounding the Fisher Information Approximation (Equation 1) to deal with the case of parametric models with bounded parameter spaces. We have shown that, when the maximum likelihood solution lies on the boundary of the statistical manifold, a novel term appears in the approximation (Equation 2). This novel boundary term can be seen as describing an aspect of the geometrical complexity of the model [3], but unlike the previously known geometric complexity terms describing the

model volume and shape (V and R) it scales logarithmically with the sample size N. This scaling property suggests that, when it is not zero, the boundary term may be the dominant contribution to geometric complexity in all but the most undersampled regimes.

Our experimental data show that naive human subjects are sensitive to model complexity in general, and to each component of the Fisher Information Approximation individually. The sensitivity is different for distinct model features (dimensionality, volume, shape, and presence of boundary), suggesting that perceptual or resource constraints may play an important role in determining the precise pattern of deviation from the ideal observer. Nevertheless, our study shows how to link principled and abstract notions of statistical model complexity to human decision making under uncertainty.

5 Acknowledgements

We thank Chris Pizzica for help with setting up the web-based version of the experiments, and for managing subject recruitment. We acknowledge support or partial support from R01 NS113241 (EP) and R01 EB026945 (VB and JG).

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. New York: Dover, 1972. ISBN: 0486612724.

[2] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Trans. by Daishi Harada. Translations of Mathematical Monographs. American Mathematical Society, 2000. 206 pp. ISBN: 0821843028.

[3] Vijay Balasubramanian. "Statistical Inference, Occam's Razor, and Statistical Mechanics on the Space of Probability Distributions." In: *Neural Computation* 9.2 (1997), pp. 349–368. DOI: 10.1162/neco.1997.9.2.349.

[4] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018. arXiv: 1701.02434 [stat.ME].

[5] William Bialek, Ilya Nemenman, and Naftali Tishby. "Predictability, Complexity and Learning." In: *Neural Computation* (13 2001), pp. 2409–2463. DOI: 10.1162/089976601753195969.

[6] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006. ISBN: 978-0387-31073-2.

[7] Bradley Efron and David L Hinkley. "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information." In: *Biometrika* 65.3 (1978), pp. 457–483. DOI: 10.1093/biomet/65.3.457.

[8] Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models." In: *Statistics and Computing* 24.6 (Aug. 2013), pp. 997–1016. DOI: 10.1007/s11222-013-9416-2.

[9] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd ed. CRC Press, 2014. ISBN: 9781439840955.

[10] Tim Genewein and Daniel A. Braun. "Occam's Razor in sensorimotor learning." In: *Proceedings of the Royal Society B: Biological Sciences* 281.1783 (May 2014), p. 20132952. DOI: 10.1098/rspb.2013.2952.

[11] Peter D Grünwald. *The Minimum Description Length Principle*. MIT press, 2007. ISBN: 9780262072816.

[12] Stephen F. Gull. "Bayesian Inductive Inference and Maximum Entropy." In: *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Springer Netherlands, 1988, pp. 53–74. DOI: 10.1007/978-94-009-3049-0_4.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2nd ed. Springer-Verlag, 2009. ISBN: 978-0387848570.

[14] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

[15] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Apr. 1, 2003. 753 pp. ISBN: 0521592712.

[16] Harold Jeffreys. *Theory of probability*. Clarendon Press, 1939. 380 pp.

[17] Samuel Johnson, Andy Jin, and Frank Keil. "Simplicity and Goodness-of-Fit in Explanation: The Case of Intuitive Curve-Fitting." In: *Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)*. 2014.

[18] John K Kruschke. *Doing Bayesian Data Analysis*. 2nd ed. Academic Press, 2015. ISBN: 9780124058880.

[19] Ravin Kumar et al. "ArviZ a unified library for exploratory analysis of Bayesian models in Python." In: *Journal of Open Source Software* 4.33 (2019), p. 1143. DOI: 10.21105/joss.01143.

[20] David J. C. MacKay. "Bayesian Interpolation." In: *Neural Computation* 4.3 (May 1992), pp. 415–447. DOI: 10.1162/neco.1992.4.3.415.

[21] Richard McElreath. *Statistical Rethinking*. CRC Press, 2016. ISBN: 9781482253443.

[22] Eugenio Piasini, Vijay Balasubramanian, and Joshua I. Gold. *Preregistration document*. 2020. URL: https://doi.org/10.17605/OSF.IO/2X9H6.

[23] Eugenio Piasini, Vijay Balasubramanian, and Joshua I. Gold. *Preregistration document addendum*. 2021. URL: https://doi.org/10.17605/OSF.IO/5HDQZ.

[24] Eugenio Piasini, Joshua I Gold, and Vijay Balasubramanian. “Information Geometry of Bayesian Model Selection.” Unpublished. 2021.

[25] Jorma Rissanen. “Stochastic Complexity and Modeling.” In: *The Annals of Statistics* 14.3 (Sept. 1986), pp. 1080–1100. URL: https://www.jstor.org/stable/3035559.

[26] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. “Probabilistic programming in Python using PyMC3.” In: *PeerJ Computer Science* 2 (Apr. 2016), e55. DOI: 10.7717/peerj-cs.55.

[27] Aki Vehtari et al. “Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC.” In: *Bayesian Analysis* (July 2020). DOI: 10.1214/20-ba1221.

Appendices

A Derivation of the boundary term in the Fisher Information Approximation

Here we generalize the derivation of the Fisher Information Approximation given by Balasubramanian [3] to the case where the maximum likelihood solution for a model lies on the boundary of the parameter space. Apart from the more general assumptions, the following derivation follows closely the original one, with some minor notational changes.

A.1 Set-up and hypotheses

The problem we consider here is that of selecting between two models (say $\mathcal{M}_1$ and $\mathcal{M}_2$), after observing empirical data $X = \left\{x_i\right\}_{i=1}^N$. N is the sample size and $\mathcal{M}_1$ is assumed to have d parameters, collectively indexed as ϑ taking values in a compact domain Θ. As a prior over ϑ we take Jeffrey's prior:

$$w(\vartheta) = \frac{\sqrt{\det g(\vartheta)}}{\int \mathrm{d}^d \vartheta \sqrt{\det g(\vartheta)}} \tag{13}$$

where g is the (expected) Fisher Information of the model $\mathcal{M}_1$:

$$g_{\mu\nu}(\vartheta) = \mathbb{E}\left[-\frac{\partial^2 \ln p(x|\vartheta)}{\partial \vartheta^\mu \partial \vartheta^\nu} \right]_\vartheta \tag{14}$$

The Bayesian posterior

$$\mathbb{P}(\mathcal{M}_1|X) = \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(X)} \int \mathrm{d}^d \, w(\vartheta) \mathbb{P}(X|\vartheta) \tag{15}$$

then becomes, after assuming a flat prior over models and dropping irrelevant terms,

$$\mathbb{P}(\mathcal{M}_1|X) = \frac{\int_\Theta \mathrm{d}^d \vartheta \sqrt{\det g} \exp\left[-N(-\frac{1}{N} \ln \mathbb{P}(X|\vartheta)) \right]}{\int \mathrm{d}^d \vartheta \sqrt{\det g}} \tag{16}$$

Just as in [3], we now make a number of regularity assumptions: 1. $\ln \mathbb{P}(X|\vartheta)$ is smooth; 2. there is a unique global minimum $\hat{\vartheta}$ for $\ln \mathbb{P}(X|\vartheta)$; 3. $g_{\mu\nu}(\vartheta)$ is smooth; 4. $g_{\mu\nu}(\hat{\vartheta})$ is positive definite; 5. $\Theta \subset \mathbb{R}^d$ is compact; and 6. the values of the local minima of $\ln \mathbb{P}(X|\vartheta)$ are bounded away from the global minimum by some $\epsilon > 0$. Importantly, unlike in [3], we don't assume that $\hat{\vartheta}$ is in the interior of Θ.

The shape of Θ. Because we are specifically interested in understanding what happens at a boundary of the parameter space, we will add a further assumption that, while being not very restrictive in spirit, will allow us to derive a particularly

interpretable result. In particular, we will assume that Θ is specified by a single linear constraint of the form

$$D_\mu \vartheta^\mu + d \geq 0 \tag{17}$$

Without loss of generality, we'll also take the constraint to be expressed in Hessian normal form — namely, $\|D_\mu\| = 1$.

For clarity, note this assumption on the shape of Θ is only used from subsection A.3 onward.

A.2 Preliminaries

We will now proceed to set up a low-temperature expansion of Equation 16 around the saddle point $\hat{\vartheta}$. We start by rewriting the numerator in Equation 16 as

$$\int_\Theta \mathrm{d}^d \vartheta \exp\left[-N\left(-\frac{1}{2N}\ln\det g - \frac{1}{N}\ln \mathbb{P}(X|\vartheta)\right)\right] \tag{18}$$

The idea of the Fisher Information Approximation is to expand the integrand in Equation 18 in powers of N around the maximum likelihood point $\hat{\vartheta}$. To this end, let's define three useful objects:

$$\tilde{I}_{\mu_1\cdots\mu_i} := -\frac{1}{N}\,\nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\mathbb{P}(X|\vartheta)\bigg|_{\hat{\vartheta}} = -\frac{1}{N}\sum_{j=1}^{N}\nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\mathbb{P}(x_j|\vartheta)\bigg|_{\hat{\vartheta}}$$

$$F_{\mu_1\cdots\mu_i} := \nabla_{\mu_1}\cdots\nabla_{\mu_i}\ln\det g(\vartheta)\bigg|_{\hat{\vartheta}}$$

$$\psi := -\frac{1}{2N}\ln\det g - \frac{1}{N}\ln\mathbb{P}(X|\vartheta)$$

We immediately note that

$$\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\bigg|_{\hat{\vartheta}} = \tilde{I}_{\mu_1\cdots\mu_i} - \frac{1}{2N}F_{\mu_1\cdots\mu_i}$$

which is useful in order to compute

$$\begin{aligned}
\psi(\vartheta) &= \psi(\hat{\vartheta}) + \nabla_\mu\psi\bigg|_{\hat{\vartheta}}(\vartheta^\mu - \hat{\vartheta}^\mu) + \frac{1}{2}\,\nabla_\mu\nabla_\nu\psi\bigg|_{\hat{\vartheta}}(\vartheta^\mu - \hat{\vartheta}^\mu)(\vartheta^\nu - \hat{\vartheta}^\nu) + \ldots \\
&= \sum_{i=0}^{\infty}\frac{1}{i!}\,\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\bigg|_{\hat{\vartheta}}(\vartheta^{\mu_1} - \hat{\vartheta}^{\mu_1})\cdots(\vartheta^{\mu_i} - \hat{\vartheta}^{\mu_i}) \\
&= \sum_{i=0}^{\infty}\frac{1}{i!}\,\nabla_{\mu_1}\cdots\nabla_{\mu_i}\psi\bigg|_{\hat{\vartheta}}\prod_{k=1}^{i}(\vartheta^{\mu_k} - \hat{\vartheta}^{\mu_k})
\end{aligned}$$

It is also useful to center the integration variables by introducing

$$\phi := \sqrt{N}(\vartheta - \hat{\vartheta}) \tag{19}$$

$$\mathrm{d}^d \phi = N^{d/2}\, \mathrm{d}^d \vartheta \tag{20}$$

so that

$$\nabla_{\mu_1} \cdots \nabla_{\mu_i} \psi \Big|_{\hat{\vartheta}} \prod_{k=1}^{i} (\vartheta^{\mu_k} - \hat{\vartheta}^{\mu_k}) = N^{-i/2} \left(\tilde{I}_{\mu_1 \cdots \mu_i} - \frac{1}{2N} F_{\mu_1 \cdots \mu_i} \right) \phi^{\mu_1} \cdots \phi^{\mu_i} \tag{21}$$

and Equation 18 becomes

$$\begin{aligned}
\int \mathrm{d}^d \vartheta \exp\big[-N\psi\big] &= N^{-d/2} \int \mathrm{d}^d \phi \exp\left[-N \sum_{i=0}^{\infty} \frac{1}{i!} N^{-i/2} \left(\tilde{I}_{\mu_1 \cdots \mu_i} - \frac{1}{2N} F_{\mu_1 \cdots \mu_i} \right) \phi^{\mu_1} \cdots \phi^{\mu_i} \right] \\
&= N^{-d/2} \int \mathrm{d}^d \phi \exp\left\{ -N \left(-\frac{1}{N} \ln \mathbb{P}(X|\hat{\vartheta}) - \frac{1}{2N} \ln \det g(\hat{\vartheta}) \right) + \right. \\
&\quad \left. -N \left[\sum_{i=1}^{\infty} \frac{1}{i!} N^{-i/2} \left(\tilde{I}_{\mu_1 \cdots \mu_i} - \frac{1}{2N} F_{\mu_1 \cdots \mu_i} \right) \phi^{\mu_1} \cdots \phi^{\mu_i} \right] \right\} \\
&= N^{-\frac{d}{2}} \exp\left[-\left(-\ln \mathbb{P}(X|\hat{\vartheta}) - \frac{1}{2} \ln \det g(\hat{\vartheta}) \right) \right] \times \\
&\quad \times \int \mathrm{d}^d \phi \exp\left\{ -N \left[\frac{1}{\sqrt{N}} \tilde{I}_{\mu} \phi^{\mu} + \frac{1}{2N} \tilde{I}_{\mu\nu} \phi^{\mu} \phi^{\nu} + \right. \right. \\
&\quad \left. \left. + \frac{1}{N} \sum_{i=1}^{\infty} N^{-\frac{i}{2}} \left(\frac{1}{(i+2)!} \tilde{I}_{\mu_1 \cdots \mu_{i+2}} \phi^{\mu_1} \cdots \phi^{\mu_{i+2}} - \frac{1}{2i!} F_{\mu_1 \cdots \mu_i} \phi^{\mu_1} \cdots \phi^{\mu_i} \right) \right] \right\}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\mathcal{M}_1|X) &= N^{-\frac{d}{2}} \exp\left[-\left(-\ln \mathbb{P}(X|\hat{\vartheta}) - \frac{1}{2} \ln \det g(\hat{\vartheta}) + \ln \int \mathrm{d}^d \vartheta \sqrt{\det g} \right) \right] \times \\
&\quad \times \int \mathrm{d}^d \phi \exp\left[-\sqrt{N} \tilde{I}_{\mu} \phi^{\mu} - \frac{1}{2} \tilde{I}_{\mu\nu} \phi^{\mu} \phi^{\nu} + \right. \\
&\quad \left. - \sum_{i=1}^{\infty} N^{-\frac{i}{2}} \left(\frac{1}{(i+2)!} \tilde{I}_{\mu_1 \cdots \mu_{i+2}} \phi^{\mu_1} \cdots \phi^{\mu_{i+2}} - \frac{1}{2i!} F_{\mu_1 \cdots \mu_i} \phi^{\mu_1} \cdots \phi^{\mu_i} \right) \right] \Bigg\} \\
&= N^{-\frac{d}{2}} \exp\left[-\left(-\ln \mathbb{P}(X|\hat{\vartheta}) - \frac{1}{2} \ln \det g(\hat{\vartheta}) + \ln \int_{\Theta} \mathrm{d}^d \vartheta \sqrt{\det g} \right) \right] \cdot Q
\end{aligned} \tag{22}$$

where

$$Q = \int_{\Phi} \mathrm{d}^d \phi \exp\left[-\sqrt{N}\tilde{I}_\mu \phi^\mu - \frac{1}{2}\tilde{I}_{\mu\nu}\phi^\mu \phi^\nu - G(\phi)\right] \tag{23}$$

and

$$G(\phi) = \sum_{i=1}^{\infty} N^{-\frac{i}{2}} \left(\frac{1}{(i+2)!}\tilde{I}_{\mu_1 \cdots \mu_{i+2}} \phi^{\mu_1} \cdots \phi^{\mu_{i+2}} - \frac{1}{2i!} F_{\mu_1 \cdots \mu_i} \phi^{\mu_1} \cdots \phi^{\mu_i} \right) \tag{24}$$

where $G(\phi)$ collects the terms that are suppressed by powers of N.

Our problem has been now reduced to computing Q by performing the integral in Equation 23. Now our assumptions come into play for the key approximation step. For the sake of simplicity, assuming that N is large we drop $G(\phi)$ from the expression above, so that Q becomes a simple Gaussian integral with a linear term:

$$Q = \int_{\Phi} \mathrm{d}^d \phi \exp\left[-\sqrt{N}\tilde{I}_\mu \phi^\mu - \frac{1}{2}\phi^\mu \tilde{I}_{\mu\nu}\phi^\nu\right] \tag{25}$$

A.3 Choosing a good system of coordinates

Consider now the Observed Fisher Information at maximum likelihood, $\tilde{I}_{\mu\nu}$. As long as it is not singular, we can define its inverse $\Delta^{\mu\nu} = (\tilde{I}_{\mu\nu})^{-1}$. If $\tilde{I}_{\mu\nu}$ is positive definite, then the matrix representation of $\tilde{I}_{\mu\nu}$ will have a set of d positive eigenvalues which we will denote by $\{\sigma_{(1)}^{-2}, \sigma_{(2)}^{-2}, \ldots, \sigma_{(d)}^{-2}\}$. The matrix representation of $\Delta^{\mu\nu}$ will have eigenvalues $\{\sigma_{(1)}^{2}, \sigma_{(2)}^{2}, \ldots, \sigma_{(d)}^{2}\}$, and will be diagonal in the same choice of coordinates as $\tilde{I}_{\mu\nu}$. Denote by U the (orthogonal) diagonalizing matrix, i.e., U is such that

$$U\Delta U^\intercal = \begin{bmatrix} \sigma_{(1)}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{(2)}^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_{(d)}^2 \end{bmatrix} \quad , \quad U^\intercal U = U U^\intercal = \mathbb{I} \tag{26}$$

Define also the matrix K as the product of the diagonal matrix with elements $1/\sigma_{(k)}$ along the diagonal and U:

$$K = \begin{bmatrix} 1/\sigma_{(1)} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{(2)} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & 1/\sigma_{(d)} \end{bmatrix} U \tag{27}$$

Note that

$$\det K = \left(\det \Delta^{\mu\nu}\right)^{-1/2} = \sqrt{\det \tilde{I}_{\mu\nu}}$$

and that K corresponds to a sphering transformation, in the sense that

$$K\Delta K^{\intercal} = \mathbb{I} \quad \text{or} \quad K^{\mu}{}_{\kappa}\Delta^{\kappa\lambda}K^{\nu}{}_{\lambda} = \delta^{\mu\nu} \tag{28}$$

and therefore, if we define the inverse

$$P = K^{-1}$$

we have

$$P^{\intercal}(\tilde{I}_{\mu\nu})P = \mathbb{I} \quad \text{or} \quad P^{\kappa}{}_{\mu}\tilde{I}_{\kappa\lambda}P^{\lambda}{}_{\nu} = \delta_{\mu\nu} \tag{29}$$

We can now define a new set of coordinates by centering and sphering, as follows:

$$\xi^{\mu} = K^{\mu}{}_{\nu}\left(\phi^{\nu} + \sqrt{N}\Delta^{\nu\kappa}\tilde{I}_{\kappa}\right) \tag{30}$$

Then,

$$\mathrm{d}^d\,\xi = \sqrt{\det \tilde{I}_{\mu\nu}}\,\mathrm{d}^d\,\phi \tag{31}$$

and

$$\phi^{\mu} = P^{\mu}{}_{\nu}\xi^{\nu} - \sqrt{N}\Delta^{\mu\nu}\tilde{I}_{\nu} \tag{32}$$

In this new set of coordinates,

$$\begin{aligned}
-\sqrt{N}\tilde{I}_{\nu}\phi^{\nu} - \frac{1}{2}\phi^{\mu}\tilde{I}_{\mu\nu}\phi^{\nu} = \\
&= -\left(\sqrt{N}\tilde{I}_{\nu} + \frac{1}{2}\phi^{\mu}\tilde{I}_{\mu\nu}\right)\phi^{\nu} \\
&= -\left(\sqrt{N}\tilde{I}_{\nu} + \frac{1}{2}P^{\mu}{}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}\frac{1}{2}\sqrt{N}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}\right)\phi^{\nu} \\
= -\sqrt{N}\tilde{I}_{\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} + N\Delta^{\nu\lambda}\tilde{I}_{\lambda}\tilde{I}_{\nu} - \frac{1}{2}P^{\mu}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} + \frac{\sqrt{N}}{2}P^{\mu}{}_{\kappa}\xi^{\kappa}\tilde{I}_{\mu\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} + \\
&+ \frac{\sqrt{N}}{2}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}P^{\nu}{}_{\lambda}\xi^{\lambda} - \frac{N}{2}\Delta^{\mu\kappa}\tilde{I}_{\kappa}\tilde{I}_{\mu\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} \\
&= \frac{N}{2}\tilde{I}_{\nu}\Delta^{\nu\lambda}\tilde{I}_{\lambda} - \frac{1}{2}\xi^{\kappa}\delta_{\kappa\lambda}\xi^{\lambda}
\end{aligned} \tag{33}$$

where we have used Equation 29 as well as the fact that $\Delta^{\mu\nu} = \Delta^{\nu\mu}$ and that $\Delta^{\mu\kappa}\tilde{I}_{\kappa\nu} = \delta^{\mu}{}_{\nu}$ by definition.

Therefore, putting Equation 31 and Equation 33 together, Equation 25 becomes

$$Q = \frac{\exp\left[\frac{N}{2}\tilde{I}_{\mu}\Delta^{\mu\nu}\tilde{I}_{\nu}\right]}{\sqrt{\det \tilde{I}_{\mu\nu}}}\int_{\Xi}\mathrm{d}^d\,\xi\exp\left[-\frac{1}{2}\xi_{\mu}\delta^{\mu\nu}\xi_{\nu}\right] \tag{34}$$

The problem is reduced to a (truncated) spherical gaussian integral, where the domain of integration Ξ will depend on the original domain Θ but also on $\tilde{I}_{\mu}$, $\tilde{I}_{\mu\nu}$ and $\hat{\vartheta}$. To complete the calculation, we now need to make this dependence explicit.

A.4 Determining the domain of integration

We start by combining Equation 19 and Equation 32 to yield

$$\vartheta^\mu = \frac{1}{\sqrt{N}} P^\mu{}_\nu \xi^\nu - \Delta^{\mu\nu}\tilde{I}_\nu + \hat{\vartheta}^\mu \tag{35}$$

By substituting Equation 35 into Equation 17 we get

$$D_\mu \left(\frac{P^\mu{}_\nu \xi^\nu}{\sqrt{N}} - \Delta^{\mu\nu}\tilde{I}_\nu + \hat{\vartheta}^\mu \right) + d \geq 0$$

which we can rewrite as

$$\tilde{D}_\mu \xi^\mu + \tilde{d} \geq 0 \tag{36}$$

with

$$\tilde{D}_\mu \coloneqq \frac{1}{\sqrt{N}} D_\nu P^\nu{}_\mu \tag{37}$$

and

$$\begin{aligned} \tilde{d} &\coloneqq d + D_\mu \hat{\vartheta}^\mu - D_\mu \Delta^{\mu\nu}\tilde{I}_\nu \\ &= d + D_\mu \hat{\vartheta}^\mu - \langle D_\mu, \tilde{I}_\mu \rangle_\Delta \end{aligned} \tag{38}$$

where by $\langle \cdot, \cdot \rangle_\Delta$ we mean the inner product in the inverse observed Fisher information metric. Now, note that whenever $\tilde{I}_\mu$ is not zero it will be parallel to D_μ. Indeed, by construction of the maximum likelihood point $\hat{\vartheta}$, the gradient of the log likelihood can only be orthogonal to the boundary at $\hat{\vartheta}$, and pointing towards the outside of the domain; therefore $\tilde{I}_\mu$, which is defined as minus the gradient, will point inward. At the same time, D_μ will also always point toward the interior of the domain because of the form of the constraint we have chosen in Equation 17. Because by assumption $\|D_\mu\| = 1$, we have that

$$\tilde{I}_\mu = \|\tilde{I}_\nu\| D_\mu$$

and

$$\langle D_\mu, \tilde{I}_\mu \rangle_\Delta = \|D_\nu\|_\Delta \cdot \|\tilde{I}_\nu\|_\Delta$$

so that

$$\tilde{d} = d + D_\mu \hat{\vartheta}^\mu - \|D_\mu\|_\Delta \cdot \|\tilde{I}_\mu\|_\Delta \tag{39}$$

Now, the signed distance of the boundary to the origin in ξ-space is

$$l = -\frac{\tilde{d}}{\|\tilde{D}_\mu\|}$$

where the sign is taken such that l is negative when the origin is included in the integration domain. But noting that

$$K^\mu{}_\kappa \Delta^{\kappa\lambda} K^\nu{}_\lambda = \delta^{\mu\nu} \quad \Rightarrow \quad \Delta^{\mu\nu} = P^\mu{}_\kappa \delta^{\kappa\lambda} P^\nu{}_\lambda$$

we have

$$\begin{aligned}\|\tilde{D}_\mu\| &= \sqrt{\tilde{D}_\mu \delta^{\mu\nu} \tilde{D}_\nu} = \sqrt{\frac{1}{N} D_\kappa \left(P^\kappa{}_\mu \delta^{\mu\nu} P^\lambda{}_\nu \right) D_\lambda} \\ &= \sqrt{\frac{1}{N} D_\kappa \Delta^{\kappa\lambda} D_\lambda} = \frac{\|D_\mu\|_\Delta}{\sqrt{N}}\end{aligned}$$

and therefore

$$l = -\sqrt{N} \frac{\tilde{d}}{\|D_\mu\|} \tag{40}$$

Finally, by plugging Equation 39 into Equation 40 we obtain

$$\begin{aligned} l &= -\sqrt{N} \left[\frac{d + D_\mu \hat{\vartheta}^\mu}{\|D_\mu\|_\Delta} - \|\tilde{I}_\mu\|_\Delta \right] \\ &=: \sqrt{2}(s - m) \end{aligned} \tag{41}$$

where m and s are defined for convenience like so:

$$m := \sqrt{\frac{N}{2}} \frac{d + D_\mu \hat{\vartheta}^\mu}{\|D_\mu\|_\Delta} \quad (\geq 0) \tag{42}$$

$$s := \sqrt{\frac{N}{2}} \|\tilde{I}_\mu\|_\Delta \quad (\geq 0) \tag{43}$$

We note that m is a rescaled version of the margin defined by the constraint on the parameters (and therefore is never negative by assumption), and s is a rescaled version of the norm of the gradient of the log likelihood in the inverse observed Fisher metric (and therefore is nonnegative by construction).

A.5 Computing the penalty

We can now perform a final change of variables in the integral in Equation 34. We rotate our coordinates to align them to the boundary, so that

$$\tilde{D}_\mu = (\|\tilde{D}_\mu\|, 0, 0, \ldots, 0)$$

Note that we can always do this as our integrand is invariant under rotation. In this coordinate system, Equation 34 factorizes:

$$\begin{aligned} Q &= \frac{\exp\left[\frac{N}{2}\tilde{I}_\mu \Delta^{\mu\nu}\tilde{I}_\nu\right]}{\sqrt{\det \tilde{I}_{\mu\nu}}} \int_{\mathbb{R}^{d-1}} \mathrm{d}^{d-1}\xi \exp\left[-\frac{\xi_\mu \delta^{\mu\nu}\xi_\nu}{2}\right] \int_l^\infty \mathrm{d}\zeta \exp\left[-\frac{\zeta^2}{2}\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}} \exp\left[\frac{N}{2}\|\tilde{I}\|_\Delta^2\right] \frac{1}{\sqrt{\pi}} \int_l^\infty \frac{\mathrm{d}\zeta}{\sqrt{2}} \exp\left[-\frac{\zeta^2}{2}\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}} \exp\left(s^2\right) \frac{1}{\sqrt{\pi}} \int_{l/\sqrt{2}}^\infty \mathrm{d}\zeta \exp\left[-\zeta^2\right] \\ &= \sqrt{\frac{(2\pi)^d}{\det \tilde{I}_{\mu\nu}}} \exp\left(s^2\right) \frac{\operatorname{erfc}(s-m)}{2} \end{aligned} \tag{44}$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function [1, section 7.1.2].

Finally, plugging Equation 44 into Equation 22 and taking the log, we obtain the extended FIA:

$$-\ln \mathbb{P}(\mathcal{M}_1|E) \simeq \ln \mathbb{P}(E|\hat{\vartheta}) + \frac{d}{2}\ln\frac{N}{2\pi} + \ln \int_\Theta \mathrm{d}^d\vartheta \sqrt{\det g} + \frac{1}{2}\ln\left[\frac{\det \tilde{I}_{\mu\nu}}{\det g_{\mu\nu}}\right] + S \tag{45}$$

where

$$S \coloneqq \ln(2) - \ln\left[\exp\left(s^2\right)\operatorname{erfc}(s-m)\right] \tag{46}$$

can be interpreted as a penalty arising from the presence of the boundary in parameter space.

A.6 Interpreting the penalty

We will now take a closer look at Equation 46. To do this, one key observation we will use is that, by construction, at most one of m and s is ever nonzero. This is because in the interior of the manifold, $m > 0$ by definition, but $s = 0$ because the gradient of the likelihood is zero at $\hat{\vartheta}$; and on the boundary, $m = 0$ by definition, and s can be either zero or positive.

Interior of the manifold When $\hat{\vartheta}$ is in the interior of the parameter space Θ, then $\tilde{I}_\mu = 0 \Rightarrow s = 0$ and Equation 46 simplifies to

$$S = \ln(2) - \ln\left(\operatorname{erfc}(-m)\right) \tag{47}$$

but since N is large we have $m \gg 0$, $\operatorname{erfc}(-m) \to 2$ and $S \to 0$, so our result passes the first sanity check: we recover the expression in [3].

Boundary of the manifold When $\hat{\vartheta}$ is on the boundary of Θ, $m = 0$ and $s \geq 0$. Equation 46 becomes

$$S = \ln(2) - \ln\Big[\exp\big(s^2\big)\operatorname{erfc}(s)\Big] = \ln(2) - \ln\big(w(is)\big) \tag{48}$$

where w is the Feddeeva function [1, p. 7.1.3]:

$$w(z) = e^{-z^2}\operatorname{erfc}(-iz)$$

This function is tabulated and can be computed efficiently. However, it is interesting to analyze its limiting behavior.

As a consistency check, when s is small we have at fixed N, to first order:

$$\begin{aligned} S &\simeq \ln(2) - \ln\left(1 - \frac{2s}{\sqrt{\pi}}\right) \\ &\simeq \ln(2) + \frac{2s}{\sqrt{\pi}} = \ln(2) + \sqrt{\frac{2N}{\pi}}\|\tilde{I}_\mu\|_\Delta \end{aligned} \tag{49}$$

and $S = \ln(2)$ when $\tilde{I}_\mu = 0$, as expected.

However, the real case of interest is the behavior of the penalty when N is assumed to be large, as this is consistent with the fact that we derived Equation 44 as an asymptotic expansion of Equation 23. In this case, using the asymptotic expansion for the Feddeeva function [1, section 7.1.23]:

$$\exp\big[s^2\big]\operatorname{erfc}(s) \sim \frac{1}{s\sqrt{\pi}}\left[1 + \sum_{m=1}^{\infty}(-1)^m \frac{1\cdot 3\cdots(2m-1)}{(2s^2)^m}\right]$$

To leading order we obtain

$$\begin{aligned} S &\simeq \ln(2) + \ln\big(s\sqrt{\pi}\big) \\ &= \ln(2) + \ln\left(\sqrt{\frac{N\pi}{2}}\|\tilde{I}_\mu\|_\Delta\right) \end{aligned}$$

which we can rewrite as

$$\boxed{S \simeq \frac{1}{2}\ln\frac{N}{2\pi} + \ln\Big[2\pi\|\tilde{I}_\mu\|_\Delta\Big]} \tag{50}$$

We can summarize the above by saying that a new penalty term of order $\ln N$ arose due to the presence of the boundary. Interestingly, comparing Equation 50 with Equation 45 we see that the first term in Equation 50 is analogous to counting an extra parameter dimension in the original Fisher Information Approximation.

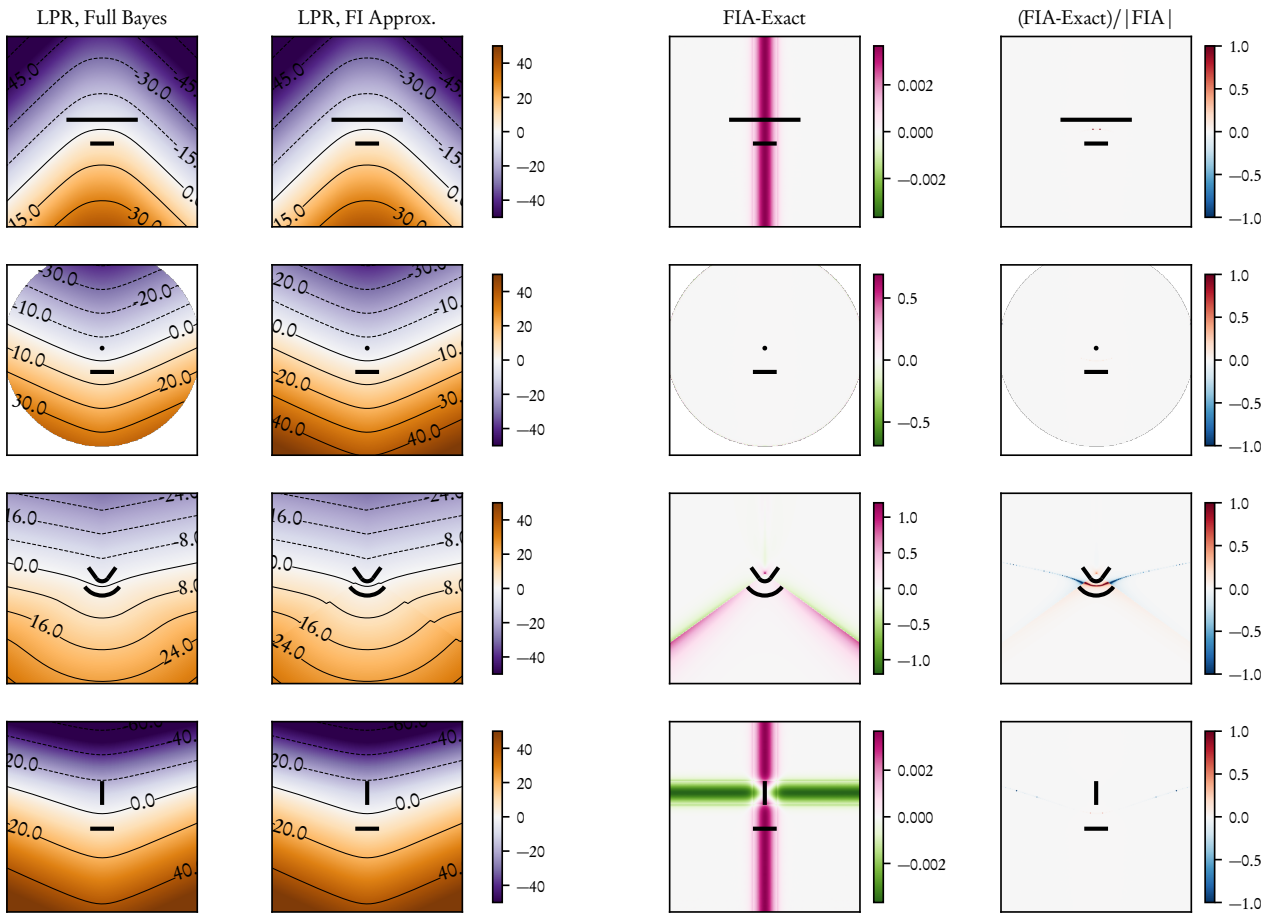

Figure 5: Comparison of Fisher Information Approximation and full Bayes computation of the log posterior ratio (LPR) for the model pairs used in our psychophysics tasks ($N = 10$). Each row corresponds to one task type (from top to bottom, "horizontal", "point", "rounded", "vertical"). First column from the left: full Bayesian LPR, computed by numerical integration. Second column: LPR computed with the Fisher Information Approximation. Third column: difference between FIA and exact LPR. Fourth column: relative difference (difference divided by the absolute value of the FIA LPR).

A.7 Numerical comparison of the extended FIA vs exact Bayes

Figure 5 shows that the FIA computed with the expressions given above provides a very good approximation to the exact Bayesian log posterior ratio (LPR) for the model pairs used in the psychophysics experiments, and for the chosen sample size ($N = 10$). As highlighted in the panels in the rightmost column, the discrepancies between the exact and the approximated LPR are generally small in relative terms, and therefore are not very important for the purpose of model fitting and interpretation. Note that here, as well as for the results in the main text, the S term in the FIA is computed using Equation 46 rather than Equation 50 in order to avoid infinities (that for finite N can arise when the likelihood gradient is very small) and discontinuities (that for finite N can arise on the interior of the manifold, in proximity to the boundary, where the value of S goes from zero when $\hat{\vartheta}$ is in the interior to $\log(2)$ when $\hat{\vartheta}$ is exactly on the boundary).

Even though overall the agreement between the approximation is good, it is interesting to look more closely at where it is the least so. The task type for which the discrepancies are the largest (both in absolute and relative terms) is the "rounded" type (third row in Figure 5). This is because the FIA hypotheses are not fully satisfied everywhere for one of the models. More specifically, the models in that task variant are a circular arc (the bottom model in Figure 5, third row) and a smaller circular arc, concentric with the first, with a straight segment attached to either side (the top model). The log-likelihood function for this second model is only smooth to first order, but its second derivative (and therefore its Fisher Information and its observed Fisher Information) are not continuous at the points where the circular arc is joined with the straight segments, locally breaking hypothesis number 3 in subsection A.1. Geometrically, this is analogous to saying that the curvature of the manifold changes abruptly at the joints. It is likely that the FIA for a model with a smoother transition between the circular arc and the straight arms would have been even closer to the exact value for all points on the 2D plane (the data space). More generally, this line of reasoning suggests that it would be interesting to investigate the features of a model that affect the quality of the Fisher Information Approximation.

B Supplementary information on the analysis of the psychophysics data

B.1 Technical details of the inference procedure

Posterior sampling was performed with PyMC3 [26] version 3.9.3, using the NUTS Hamiltonian Monte Carlo algorithm [14], with target acceptance probability set to 0.9. The posterior distributions reported in the main text are built by sampling 8 independent Markov chains for 10000 draws each. No divergence occurred in any of the chains. Effective sample size and $\hat{R}$ diagnostics for some

Parameter	ESS	$\hat{R}$
μ_α	3214	1.00
μ_L	1068	1.01
μ_S	2017	1.00
μ_D	2047	1.00
μ_V	3737	1.00
μ_R	6181	1.00

Table 1: $\hat{R}$ statistic and effective sample size (ESS) for 8 Markov Chain traces run as described in the main text. See [9, sections 11.4–11.5] and [27] for in-depth discussion of chain quality diagnostics. Briefly, $\hat{R}$ depends on the relationship between the variance of the draws estimated within and between contiguous draw sequences. $\hat{R}$ is close to 1 when the chains have successfully converged. The effective sample size estimates how many independent samples one would need to extract the same amount of information as that contained in the (correlated) MCMC draws. Note that here, for computational convenience, we report diagnostics for 8 chains with 1000 draws each, while the results reported in the main text have been obtained with 10 times as many draws (8 chains × 10000 draws per chain), run with identical settings.

of the key parameters are given in table Table 1 for a shorter run of the same procedure.

B.2 Posterior predictive checks

We performed a simple posterior predictive check [18] to ensure that the Bayesian hierarchical model described in the main text captures the main pattern of behavior across our subjects. In Figure 6, the behavioral performance of the subjects is compared with its posterior predictive distribution under the model. As can be seen from the figure, the performance of each subject is correctly captured by the model, across systematic differences between task types (with subjects performing better in the "vertical" task than the "rounded" task, for instance) as well as individual differences between subjects that performed the same task variant.

B.3 Formal model comparison

We compared the Bayesian hierarchical model described in the main text to a simpler model, where subjects were assumed to only be sensitive to likelihood differences, or in other words to choose $\mathcal{M}_1$ over $\mathcal{M}_2$ only based on which model was on average closer to the dot cloud constituting the stimulus on a given trial. Mathematically, this "likelihood only" model was equivalent to fixing all β parameters to zero except for β_L in the model described in the main text. All other details of the model were the same, and in particular the model still had a hierarchical structure with adaptive shrinkage (the subject-level parameters

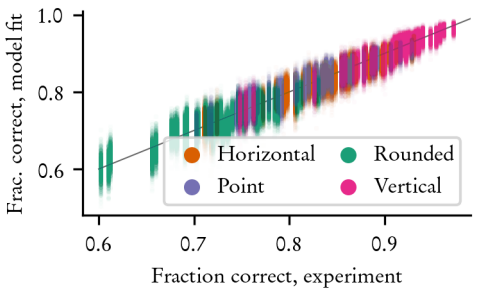

Figure 6: Simple posterior predictive check, looking at subject performance. A random sample of all subject-level parameters (α_i and β_i) is taken at random from the MCMC chains used for model inference. Using those parameter values, a simulation of the experiment is run using the actual stimuli shown to the subjects, and the resulting performance of all 202 simulated subjects is recorded. This procedure is repeated 2000 times, yielding 2000 samples of the joint posterior-predictive distribution of task performance over all experimental subjects. To visualize this distribution, for each subject we plotted a cloud of 2000 dots where the y coordinate of each dot is the simulated performance of that subject in one of the simulations, and the x coordinate is the true performance of that subject in the experiment plus a small random jitter (for ease of visualization). The gray line is the identity, showing that our inference procedure captures well the behavioral patterns in the experimental data.

α and β_L were modeled as coming from Student T distributions controlled by population-level parameters). We compared the full model and the likelihood-only using the Widely Applicable Information Criterion [8]. This comparison, shown in Table 2, reveals strong evidence in favor of the full model.

Model	Rank	WAIC	pWAIC	dWAIC	SE	dSE
Full	0	-34823.9	640.856	0	188.421	0
Likelihood only	1	-37524.2	369.713	2700.3	190.453	69.3959

Table 2: WAIC comparison of the full model and the likelihood-only model for the experimental data, reported in the standard format used by [21, section 6.4.2]. Briefly, WAIC is the value of the criterion (log-score scale — higher is better); pWAIC is the estimated effective number of parameters; dWAIC is the difference between the WAIC of the given model and the highest-ranked one; SE is the standard error of the WAIC estimate; and dSE is the standard error of the difference in WAIC. These estimates were produced with the `compare` function provided by ArviZ [19], using 8 MCMC chains with 1000 samples each for each model (in total, 8000 samples for each model).