Playing it safe: information constrains collective betting strategies

Philipp Fleig
Department of Cellular Biophysics, Max Planck Institute for Medical Research, Jahnstraße 29, 69120 Heidelberg, Germany and Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

Vijay Balasubramanian
Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA and Santa Fe Institute, Santa Fe, NM 87501, USA

(Dated: May 28, 2023)

Every interaction of a living organism with its environment involves the placement of a bet. Armed with partial knowledge about a stochastic world, the organism must decide its next step or near-term strategy, an act that implicitly or explicitly involves the assumption of a model of the world. Better information about environmental statistics can improve the bet quality, but in practice resources for information gathering are always limited. We argue that theories of optimal inference dictate that "complex" models are harder to infer with bounded information and lead to larger prediction errors. Thus, we propose a principle of *playing it safe* where, given finite information gathering capacity, biological systems should be biased towards simpler models of the world, and thereby to less risky betting strategies. In the framework of Bayesian inference, we show that there is an optimally safe adaptation strategy determined by the Bayesian prior. We then demonstrate that, in the context of stochastic phenotypic switching by bacteria, implementation of our principle of "playing it safe" increases fitness (population growth rate) of the bacterial collective. We suggest that the principle applies broadly to problems of adaptation, learning and evolution, and illuminates the types of environments in which organisms are able to thrive.

INTRODUCTION

Risk is an inherent part of life. Whether in pathogen detection [1], phenotype selection [2, 3], biochemical and evolutionary mechanisms [4, 5], foraging and exploration-exploitation strategies [6–9], or reproduction [10, 11], biological functions are partly shaped by the need to reduce risk. Broadly, risk arises from uncertain interactions of an organism with the world around it [12]. First, both the environment and the typical sensory apparatus are intrinsically noisy. Thus an organism only has probabilistic information about the state of the world. Second, any finite system can only observe some aspects of the world, while others, perhaps most, remain hidden from observation. Thus, an organism inherently functions with partial information, bounding its capacity for rational decision-making. In other words, living systems must necessarily play betting games, producing responses, e.g. expressing a phenotype, detecting an odor, showing an immune response, that lead to outcomes that will probably be beneficial given the limited and uncertain information available for decision-making.

Every betting game has an associated risk. To improve the betting strategy, it is vital to know the odds of the game. An organism can mitigate the risk of its strategy, and minimize the likelihood of wrong bets, by gathering information to assess the odds and adapting to the learned statistics of the environment. Such learning and adaptation is particularly important if rare, but potentially catastrophic, events can occur.

Living systems can adapt in this way to the environment over many timescales, over the course of evolution, in response to environmental cues, and through ongoing life-long learning. Given the stochastic nature of the environment from the perspective of an agent inhabiting it, essentially all adaptation processes can be viewed as schemes for inferring a probability distribution [13, 14]. The ability to infer this distribution is limited by physical and temporal constraints of information gathering. In turn, this limits the capacity for adaptation, and thus increases risk for the system. There are several probabilistic inference frameworks that explicitly or implicitly account for the limitations of information gathering processes guiding model choice [15–22]. Here we argue that a key idea arising from these frameworks is that "complex" models, in a sense defined in information geometry [22, 23], are difficult to infer robustly from limited data, and that a small inaccuracy in inference of such models can lead to large deviations from the optimal strategy.

Because of this, if information-gathering is limited, it can pay to bias inference towards simple models. Thus, we propose a principle of "playing it safe", which dictates that given finite information, biological systems should effectively bias themselves towards simpler models. We derive this principle for a large class of probabilistic models using Bayesian probability theory and information geometry, and show how the bias towards simplicity can be tuned by the choice of prior. We then illustrate the efficacy of the principle in the classic example of Kelly betting [24, 25], and demonstrate a biological realization in the phenomenon of stochastic phenotypic switching by bacteria. We conclude with a discussion of the implications of our finding for the types of environments that are "learnable", and in which living systems can prosper.

RESULTS

Optimising betting success is hard for complex models

Betting games are unpredictable by nature. However, bet placement can be optimised by knowing the probabilities controlling events in the game, which often requires estimation of a probability distribution determined by d real parameters, $P(\mathbf{x};\boldsymbol{\theta}=\{\theta_1,\ldots,\theta_d\})$, denoted here by $\boldsymbol{\theta}$. In real-world settings, the information required to infer parameters is typically limited. For example, we may have N independent observations $\mathbf{x}_i$, $\mathbf{n}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ of a random event which can be used for the inference of parameters. However, finite data leads to statistical fluctuations in the inference process and we need to understand how these affect betting success.

Success in betting games can be quantified through a loss function. In many games the loss is a function of the game's true probability distribution $P_{\boldsymbol{\theta}^*}$ and an inferred distribution $Q_{\boldsymbol{\phi}}$, adopted by the bettor. The true distribution can lie outside the manifold parameterized by $\boldsymbol{\theta}$ within which the inference is performed, but we will nevertheless use the notation $\boldsymbol{\theta}^*$ for the data generating distribution, because we will often consider cases where this distribution lies within the considered model family. We expect the loss to have a global minimum when $\boldsymbol{\phi}=\boldsymbol{\theta}^*$ if the truth lies within the model family under consideration, since exact knowledge of the game's odds allows the bettor to place bets optimally and thus minimise the loss. If the truth lies outside the model family, the loss should be minimized when $\boldsymbol{\phi}$ comes "as close as possible" in terms of the loss function. Here, we adopt the Kullback-Leibler (KL) divergence

$$\mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\phi}) \equiv D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||Q_{\boldsymbol{\phi}}) = \sum_{\mathbf{x}} P_{\boldsymbol{\theta}^*}(\mathbf{x}) \log \frac{P_{\boldsymbol{\theta}^*}(\mathbf{x})}{Q_{\boldsymbol{\phi}}(\mathbf{x})} \quad (1)$$

to measure the difference between the true probabilities and the adopted model. The closer the adopted model is to the truth, the smaller the loss. The KL divergence is the loss function of the well-known Kelly betting setup. An important property of the KL divergence is that for many natural parametrizations of distributions it diverges on the boundaries of the parameter space. Changing the parameter by a small amount in such regions leads to large changes in the KL divergence.

A way to understand the origin of this behaviour is through model complexity. Consider the Jeffreys prior

$$w(\boldsymbol{\theta}) = \frac{\sqrt{\det \mathcal{I}(\boldsymbol{\theta})}}{\int \mathrm{d}^d\boldsymbol{\psi} \sqrt{\det \mathcal{I}(\boldsymbol{\psi})}} \quad (2)$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information metric (FIM), obtained by Taylor expansion from the KL divergence between models in the parametric family [26, 27] (*Materials and Methods*). In information geometry, this quantity is understood as the density of distinguishable models, and is regarded as a local measure of complexity in a parametric model family [22, 27].

In regions where Jeffreys prior is small, i.e., where the model family is *simple*, changing a parameter by a small amount does not move the distribution much in model space [22]. However, Jeffreys prior can become large and even diverge. In such regions, i.e., where the model family is *complex*, even models which are close in parameter space are far from each other in model space. In many natural parametrizations, divergences occur at the boundaries of parameter space (see Figure 1 and Eq. (25) in *Materials and Methods* for the multinomial distribution, and *Appendix*, Section A for the Poisson and Gaussian distributions).

The multinomial distribution in $d=2$ variables is described by a single parameter, typically the event probability θ. The corresponding Jeffreys prior in Figure 1A takes the form $w(\theta) = 1/\pi\sqrt{\theta(1-\theta)}$. Equivalently $w(\theta)\,\delta\theta^2$ measures the KL divergence between multinomial models parametrized by θ and $\theta+\delta\theta$. We see that the density of distinguishable models has a minimum at $\theta=1/2$ and diverges at the boundaries in these particular coordinates. The inset in Figure 1A shows that to keep the KL divergence $D_{\mathrm{KL}}(P_\theta||P_\phi)$ between two nearby multinomial models below some value ϵ, the parameter $\phi\equiv\theta+\delta\theta$ must lie closer to θ, when θ itself lies closer to the boundary of the one-dimensional parameter space at 0. By symmetry, the same holds near the other boundary at 1. Thus, we see that in this case when the true model is close to a boundary of parameter space, the parameter ϕ must be inferred with higher precision than when the true model lies near a value around 1/2, in order to achieve the same degree of statistical proximity.

It is hard to optimise the KL divergence in regions of parameter space that are complex in the sense described above. For example, we may decide to rely on the maximum-likelihood estimate of parameters $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ and define the expected loss as

$$\left\langle \mathcal{L}_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) \right\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} \equiv \int \mathrm{d}^d\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\, p(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\, \mathcal{L}_{\boldsymbol{\theta}^*}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\,, \quad (3)$$

where p is the distribution of maximum-likelihood values under the true distribution $\boldsymbol{\theta}^*$ given N observations. In the limit that $N\to\infty$, this distribution becomes sharply peaked around $\boldsymbol{\theta}^*$. However, for finite N, p has finite width and the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is subject to statistical fluctuations due to finite sampling, thus limiting the precision with which the parameters can be determined. In *Appendix*, Section B we show for the Bernoulli and Poisson models that the KL divergence between the true model and the expected maximum-likelihood model plus/minus the standard error diverges as the truth approaches the boundary, even if the standard error on the maximum-likelihood estimate decreases to zero with increasing amounts of data. In addition, apart from fluctuations due to finite sampling, in real-world data the observations may additionally be corrupted, effectively introducing an error floor in the

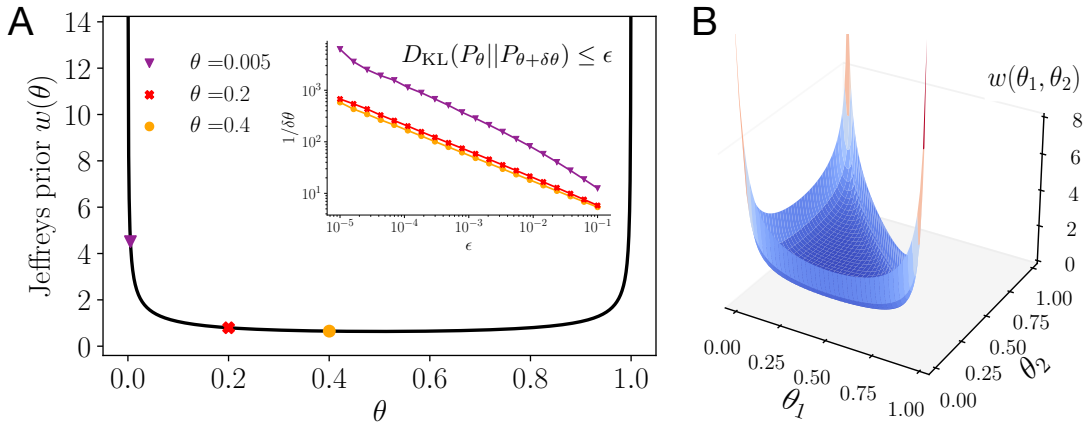

FIG. 1. Landscape of model complexity dictates the required inference precision. Inferring a model close to a boundary of parameter space requires higher inference precision. (A) Jeffreys prior (black curve) of the Bernoulli model with event probability θ measures model complexity and diverges at the boundaries of the one-dimensional parameter space. Three example models are indicated by the orange, red, and purple markers. Inset: precision with which the models $\phi = \theta + \delta\theta$ need to be determined to keep the Kullback-Leibler divergence, below some value ϵ. (B) Jeffreys prior for the multinomial model in $d = 3$ dimensions. The parameter space is two-dimensional, showing the typical divergences of probabilistic models on boundaries of the space.

empirical estimate.

The question then is what probabilistic model should be selected to control the effects of fluctuations while optimising the placement of bets with finite data?

The playing it safe principle

We seek a model with the following properties. In the large data limit (large N), the optimal model has to converge to the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ which itself converges to the true model $\boldsymbol{\theta}^*$. In the opposite limit of small data (small N), we should select a model that is less affected by statistical fluctuations. We will parametrize this in terms of a deterministic bias $\boldsymbol{\theta}_{\mathrm{bias}}$ and show below how to select it. For finite data size, the optimal model should represent a compromise between the maximum-likelihood model and the bias. Overall, we make the following Ansatz for the model

$$\boldsymbol{\theta} = \kappa\, \boldsymbol{\theta}_{\mathrm{bias}} + (1-\kappa)\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\,, \tag{4}$$

where the parameter $\kappa \in [0,1]$ is used to interpolate between the bias and maximum-likelihood term. In the next section we show that this Ansatz appears naturally in the inference of a large class of probability distributions (exponential families with a conjugate prior). *A priori*, the values of the parameters κ and $\boldsymbol{\theta}_{\mathrm{bias}}$ can be chosen freely, as long as $\boldsymbol{\theta}$ lies in the parameter space of the probabilistic model. We determine the optimal choice of these parameters through minimisation of the expected loss, $\langle \mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}}$. There are different ways to optimise with respect to these parameters. For instance, we may keep κ fixed and optimise for the bias, or vice versa. These lead to different optimal models and which way we pick depends on the setup of the inference we want to solve. We focus on the former approach to optimisation and later discuss the second option in an example. Thus, we optimise the expected loss with respect to the bias:

$$\boldsymbol{\theta}^{\mathrm{opt}}_{\mathrm{bias}} = \arg\min_{\boldsymbol{\theta}_{\mathrm{bias}}} \langle \mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}}\,. \tag{5}$$

To evaluate the condition, it is convenient to write it as a gradient of the expected loss

$$\frac{1}{\kappa}\nabla_{\boldsymbol{\theta}_{\mathrm{bias}}}\langle \mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \nabla_{\boldsymbol{\theta}}\langle \mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0\,. \tag{6}$$

The gradient field is defined over parameter space and we are looking for optimal points where the gradient vanishes. As the minima of gradient fields are invariant under coordinate changes, the solution of the optimization problem is invariant to changes in parameterization of the underlying probabilistic model. However, next we also show that for a large class of probability distributions, there exists a particular parameterization that allows us to understand in detail how the optimum arises from an interplay of the amount of data available about the true model and the geometry of model space. This interplay constrains the regions of parameter space that yield safe betting models, and we thus refer to it as the "playing it safe" principle.

To make analytic progress, we specialise to exponential families which form a large class of probability distributions, which include many well-known distributions such as the Multinomial, Poisson, and Gaussian distributions etc. [28, 29]. An exponential family is a collection of probability densities that can be written in the canonical form

$$P_{\boldsymbol{\eta}}(x) = \exp\left(\boldsymbol{\eta}^T \cdot \mathbf{t}(x) - F(\boldsymbol{\eta}) + k(x)\right)\,, \tag{7}$$

where $\boldsymbol{\eta}$ is the *canonical parameter* associated with the *sufficient statistic* $\mathbf{t}(x)$, F is the log-partition function, and $k(x)$ is referred to as the auxiliary carrier term. Alternatively, we can use the *mean parameters* $\boldsymbol{\theta}$ which are related to the canonical parameters by a one-to-one map (see *Materials and Methods*). The mean parameters often also have a simple relation to commonly used parameterizations of well-known distributions. Examples are the event probabilities in the Multinomial distribution, the average number of counts in the Poisson distribution, and the variance of the Gaussian distribution with zero mean. We give examples of these relations below and in *Appendix*, Sections D and E.

The KL divergence between two densities $P_{\boldsymbol{\eta}^*}$ and $P_{\boldsymbol{\eta}}$ of the same exponential family, can be written as a Bregman divergence [30, 31] between the associated mean parameters:

$$D_{\mathrm{KL}}(P_{\boldsymbol{\eta}^*(\boldsymbol{\theta}^*)}||P_{\boldsymbol{\eta}(\boldsymbol{\theta})}) = B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}) \,, \tag{8}$$

where F^* is the dual of the log-partition function F (see *Materials and Methods*). Bregman divergences form a large class of statistical distances, including many commonly known measures, such as the squared Euclidean distance, the Mahalanobis distance, or the Itakura–Saito distance. We note that in our context, the particular choice of the probabilistic model $P_{\boldsymbol{\eta}}$, implies a particular Bregman divergence through the dual function F^*. In *Appendix*, Section C1 we show that for exponential families in mean parameterisation, the optimality condition Eq. (6) takes the simple form

$$\left\langle \mathcal{I}(\boldsymbol{\theta})\delta\boldsymbol{\theta} \right\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0 \,, \tag{9}$$

which holds even for differences $\delta\boldsymbol{\theta} \equiv \boldsymbol{\theta}^* - \boldsymbol{\theta}$ of finite size. In general, the condition provides d non-linear constraints on the parameters κ and $\boldsymbol{\theta}_{\mathrm{bias}}$ such that the expected loss is optimised. We give an example illustrating the derivation of this condition for the Bernoulli model in *Appendix*, Section C2.

We now illustrate that this condition, which holds for the mean parameterisation, implies a bias towards models of lower complexity in the sense described in the previous section. For this we focus on the case when the exponential family depends on a single parameter and write out the condition in Eq. (9) explicitly using the definitions of $\delta\theta$ and θ in Eq. (4):

$$\left\langle \mathcal{I}(\theta)\left[\theta^* - \theta\right] \right\rangle_{\hat{\theta}_{\mathrm{ML}}} = \left\langle \mathcal{I}(\theta)\left[\theta^* - \hat{\theta}_{\mathrm{ML}} - \kappa(\theta_{\mathrm{bias}} - \hat{\theta}_{\mathrm{ML}})\right] \right\rangle_{\hat{\theta}_{\mathrm{ML}}} = 0 \,. \tag{10}$$

To illustrate the implications of this condition, we assume an infinite parameter space with a single boundary. For example, the Poisson distribution with parameter $\lambda \in (0, \infty)$ has such a parameter space. We consider the scenario shown in Figure 2, where the Fisher information metric $\mathcal{I}$ diverges to one side of the true model and falls off to small values to the other side. We then analyse the large and small data limits as illustrated in Figure 2, and optimise θ_{bias} while κ is fixed to a finite value:

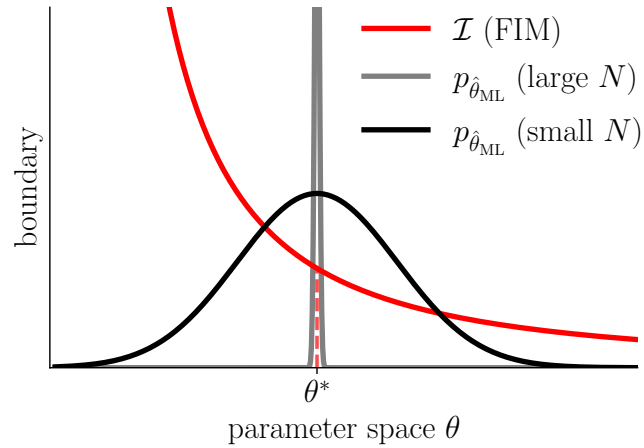

FIG. 2. Diverging Fisher information at the boundary leads to an asymmetric complexity weight $\mathcal{I}(\theta)$ around the true model θ^*. We illustrate this for a one-dimensional parameter space with a single boundary. The black and grey curves are the distributions of the empirically observed maximum-likelihood value $\hat{\theta}_{\mathrm{ML}}$ for small and large N, respectively. For small N, the distribution is wide and models to the left of the true model θ^* (dashed red line), receive a higher complexity weight when evaluating $\langle \mathcal{I}(\theta)(\theta^* - \theta)\rangle_{\hat{\theta}_{\mathrm{ML}}}$ (left-hand side of Eq. (9)), than models to the right of the true model.

Large N: The distribution of the maximum-likelihood value becomes sharply peaked around θ^* and we see that in order to make the bracket inside the expectation value vanish, we require $\theta_{\mathrm{bias}} \to \theta^*$. We note that the local complexity $\mathcal{I}$ can be arbitrarily large in this limit as it is multiplied by the bracket with value zero. Empirical parameters $\hat{\theta}_{\mathrm{ML}}$ for which $\theta^* - \hat{\theta}_{\mathrm{ML}}$ is not zero are improbable and multiplied by very small probability when taking the expectation value. Thus the entire expectation value will vanish as required.

Small N: The distribution of the ML parameter has finite spread around the true model θ^*. For simplicity of argument suppose that the distribution is symmetric around θ^*. To understand the optimization, we go through the possible cases for choosing the bias:

- $\theta_{\mathrm{bias}} = \theta^*$: The expectation value becomes non-zero (either positive or negative), because the ML values on one side of the true model θ^* receive a higher complexity weight $\mathcal{I}(\theta)$ than the ML values on the opposite side.

- θ_{bias} on the side of θ^* with diverging complexity: The expectation value assumes even larger (negative or positive) values.

- θ_{bias} on the side of θ^* with diminishing complexity: For the right magnitude of θ_{bias} the expectation value is adjusted to zero, because the models

$\theta(\theta_{\rm bias}, \hat{\theta}_{\rm ML})$ are biased to receive a smaller complexity weight. This demonstrates the "playing it safe" principle for this case.

This argument in the one-dimensional case immediately generalises to the case of higher-dimensional parameter spaces. Equation (9) provides an independent condition for each of the d dimensions of parameter space and the components of the bias $\theta_{\rm bias}$ can thus be adjusted individually to satisfy all conditions.

Optimal Bayesian inference for exponential families

We now show how the "playing it safe" principle introduced in the previous section, and in particular the Ansatz Eq. (4), arises for Bayesian inference of exponential families with conjugate priors, for which we can provide an analytical treatment. In short, the information contained in the prior determines the function κ and the bias $\boldsymbol{\theta}_{\rm bias}$ in Eq. (4). We will show that by tuning the conjugate prior through its hyperparameters we can control the strength of the bias relative to the maximum-likelihood term.

We now specialise to exponential families Eq. (7) in canonical coordinates $\boldsymbol{\eta}$. If the true model $\boldsymbol{\eta}^*$ is unknown, we can at best use the N independent observations $\mathbf{x}_i$, $\mathbf{n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, to determine a Bayesian posterior distribution for $\boldsymbol{\eta}$,

$$P(\boldsymbol{\eta}|\mathbf{n}) = \frac{P(\mathbf{n}|\boldsymbol{\eta})P(\boldsymbol{\eta})}{\int \mathrm{d}^d\boldsymbol{\psi}\, P(\mathbf{n}|\boldsymbol{\psi})P(\boldsymbol{\psi})}\,, \tag{11}$$

where $P(\boldsymbol{\eta})$ is a prior distribution, and $P(\mathbf{n}|\boldsymbol{\eta})$ is the likelihood of seeing data $\mathbf{n}$ given the model parameterized by $\boldsymbol{\eta}$. Given the posterior, the goal is to minimise the posterior expected loss

$$\left\langle \mathcal{L} \right\rangle_{\rm post} = \int \mathrm{d}\tilde{\boldsymbol{\eta}} P(\tilde{\boldsymbol{\eta}}|\mathbf{n}) D_{\rm KL}(Q_{\tilde{\boldsymbol{\eta}}}||Q_{\boldsymbol{\eta}}) \tag{12}$$

$$= \int d\tilde{\boldsymbol{\theta}}\, P(\tilde{\boldsymbol{\theta}}|\mathbf{n}) B_{F^*}(\tilde{\boldsymbol{\theta}} : \boldsymbol{\theta})\,, \tag{13}$$

where in the second line we have used relation Eq. (8) and transformed variables from canonical to mean parameters also in the posterior density and integral. We want to determine the minimum value of this expression with respect to $\boldsymbol{\theta}$

$$\left\langle \mathcal{L} \right\rangle_{\rm post}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left\langle \mathcal{L} \right\rangle_{\rm post}(\boldsymbol{\theta})\,. \tag{14}$$

This optimal value is also known as the Bregman information. In [32] it was shown that the conditional optimal parameter $\boldsymbol{\theta}$ is given by the posterior expectation value

$$\boldsymbol{\theta} = \int d\tilde{\boldsymbol{\theta}} P(\tilde{\boldsymbol{\theta}}|\mathbf{n})\tilde{\boldsymbol{\theta}}\,. \tag{15}$$

Thus, the optimal parameter depends both on the data and the choice of prior. Our next goal is to understand how the prior choice in relation to the available data size affects the optimal parameter.

Conjugate priors for exponential families allow us to make general analytic statements in this respect. We denote the conjugate prior of the exponential family by $P(\boldsymbol{\eta}; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a set of hyperparameters. In [33] it was shown for exponential families, that under the conjugate prior the posterior mean is given by

$$\boldsymbol{\theta}(\hat{\boldsymbol{\theta}}_{\rm ML}; \boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha} + N\hat{\boldsymbol{\theta}}_{\rm ML}}{N_0 + N} = \kappa\frac{\boldsymbol{\alpha}}{N_0} + (1-\kappa)\hat{\boldsymbol{\theta}}_{\rm ML}\,, \tag{16}$$

where N_0 is a function of the hyperparameters $\boldsymbol{\alpha}$, and $\kappa \equiv N_0/(N_0 + N)$. From this expression we can understand the role played by the hyperparameters. $N_0(\boldsymbol{\alpha})$ is an threshold number of samples determined by the hyperparameters that sets the scale for whether the actual sample size N is large enough for the evidence to overcome the prior. Likewise, $\boldsymbol{\alpha}/N_0$ functions as effective additional parameters that appear on equal footing as the maximum-likelihood parameter. In fact, the quotient $\boldsymbol{\alpha}/N_0$ is the prior expectation for the parameter $\boldsymbol{\theta}$, and we denote it by $\boldsymbol{\theta}_{\rm prior} \equiv \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \boldsymbol{\alpha}/N_0$. Overall, this leaves us with the following form of the conditional optimal parameter:

$$\boldsymbol{\theta} = \kappa\boldsymbol{\theta}_{\rm prior} + (1-\kappa)\hat{\boldsymbol{\theta}}_{\rm ML}\,, \tag{17}$$

with

$$\kappa = \frac{N_0/N}{N_0/N + 1} = \frac{q}{q+1}\,, \tag{18}$$

where $q \equiv N_0/N$. We find $\kappa \to 0$ for $N_0/N \to 0$, and $\kappa \to 1$ for $N_0/N \to \infty$. Finally, comparison with Eq. (4) shows that $\boldsymbol{\theta}_{\rm prior}$ takes the role of the bias $\boldsymbol{\theta}_{\rm bias}$, such that overall this confirms our initial Ansatz.

In the above derivation, the conjugate prior $P(\boldsymbol{\eta}; \boldsymbol{\alpha})$ was defined on the canonical parameters and lead to the linear convex form of Eq. (17). However, in the next section and in (*Appendix*, Sections D and E), we show for the multinomial, Poisson, and Gaussian model, that defining the conjugate prior on the event probabilities, the mean number of counts, and the variance, respectively, equally results in the optimal parameter having convex linear form. We emphasise that for the development of our theory it is not important on which set of parameters the conjugate prior is defined. What counts for our argument, is that the final expression for the conditional optimal parameter has the linear convex form.

The conditional optimal parameter in convex linear form depends on the data and crucially also on the shape of the prior, set by its hyperparameters. While for a given prior, the loss is minimised by the conditional optimal model, we may further optimise the expected loss with respect to the prior choice which we do by optimising the hyperparameters. Given the convex linear form in Eq. (17), we see that there are two natural ways to optimise the hyperparameters. The first, is to make a

choice for $\boldsymbol{\theta}_{\text{prior}}$, and then optimise the value of κ (or, equivalently q). The second way is to set the value of κ and then optimise $\boldsymbol{\theta}_{\text{prior}}$. We now illustrate both paths of optimisation for the multinomial distribution.

Optimization for the multinomial model

The multinomial model is the probability distribution of N events, each assuming one of d possible outcomes (*Materials and Methods*). The true model is parameterised by the d-dimensional set of probabilities $\boldsymbol{\theta}^*$. We note that these probabilities are related to the mean parameters of the multinomial distribution by rescaling the mean parameters with a factor of $1/N$. For simplicity of notation and to emphasize their close relation, we use the same symbol $\boldsymbol{\theta}$ to denote the probabilities that we used above for the mean parameters. It is natural to collect information on the model by counting event outcomes and computing empirical probabilities $\hat{\boldsymbol{\theta}}$. The data $\mathbf{n}$ consists of the counts of occurrence x_i for each of the d possible outcomes over N observations ($\sum_{i=1}^{d} x_i = N$). Following our earlier discussion, we use Bayesian inference with a conjugate prior and optimise the conditional model to minimise the expected loss. The likelihood function is the multinomial distribution and its conjugate prior is a Dirichlet distribution given by

$$P(\mathbf{n}|\boldsymbol{\theta}) = \frac{1}{B(\mathbf{n})} \prod_{i=1}^{d} \theta_i^{x_i} \,, \quad P(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{d} \theta_i^{\alpha_i - 1} \,. \tag{19}$$

The set of d hyperparameters $\alpha_i > 0$ determines the prior shape. Direct calculation yields, as expected, the linear convex form Eq. (17) of the conditional optimal model

$$\theta_i(\hat{\boldsymbol{\theta}};\boldsymbol{\alpha}) = \frac{\frac{\alpha_i}{N} + \hat{\theta}_i}{\frac{1}{N}\sum_i \alpha_i + 1} \,, \tag{20}$$

where $\hat{\theta}_i = x_i/N$, are the empirical probabilities. The expression in Eq. (20) is also known as a pseudo-Laplacian law [15, 34, 35] with the α_i being "pseudocounts". We now specialise to $d = 2$, i.e. two possible event outcomes. The conditional optimal model depends on two parameters, α_1 and α_2 from which we define

$$N_0 \equiv \alpha_1 + \alpha_2 \,, \; \boldsymbol{\theta}_{\text{bias}} \equiv \frac{\boldsymbol{\alpha}}{N_0} \,. \tag{21}$$

The next step is to optimize these hyperparameters in the two possible ways outlined in the previous section.

First, we consider optimization of the bias $\boldsymbol{\theta}_{\text{bias}}$, while κ is fixed. Since $\theta_2 = 1 - \theta_1$, we are dealing with optimization in one dimension. For simplicity of notation, we suppress component indices in the following. The landscape of the KL loss is shown in Figure 3A for $\theta^* = 0.2$ and $\kappa = 0.2$. The curve of the optimal bias choice and the optimized parameter $\langle\theta_{\text{opt}}\rangle$ (averaged over empirical observations), as a function of N is shown in white. Asymptotically, the curve approaches θ^* for both parameters. In Figure 3B, we show how the local complexity of the model with optimal parameter compares to the complexity of the true model as a function of N and for different values of the true model. Asymptotically, the complexity ratio approaches unity, but the closer the true model lies to the boundary, the more data is required to reach the full complexity of the true model. We also observe in Figs. 3A and B that for small N there is an initial dip in the curves. At first sight this may seem to be in tension with the "playing it safe" principle. However, this is explained and resolved by the existence of the second boundary of the parameter space of the $d = 2$ multinomial model at $\theta = 1$. For small N, the distribution of maximum-likelihood values is so broad that diverging complexity at the second boundary influences the optimization, pushing the parameter away from the boundary.

Next, we consider optimization of the parameter κ, while θ_{bias} is held fixed. We make the choice $\alpha_1 = \alpha_2$ which, by Eq. (21) implies $\theta_{\text{bias}} = 1/2$. This is an appealing choice as it implies equal prior knowledge in all directions of parameter space. The landscape of the KL loss is shown in Figure 3C with $\theta^* = 0.2$. The curve of the optimal choice for κ and the optimized parameter $\langle\theta_{\text{opt}}\rangle$ as a function of N is shown in white. Asymptotically, κ approaches zero and the optimal model approaches θ^*. In Figure 3D, we show for different true models, that the complexity ratio approaches unity in the asymptotic limit as expected, but the closer the true model lies to the boundary, the more data is required.

We thus see that either way of optimising implies "playing it safe" as strategy. In (*Appendix*, Sections D and E), we demonstrate the optimization of the bias for the Poisson model with unknown average number of counts, and the Gaussian model with unknown variance parameter.

Playing it safe in bacterial phenotypic switching

Betting strategies in biology are often found in collective systems where bets can be distributed (or hedged), across the possible outcomes of random events. There are many cases of bet hedging systems in biology [10, 36, 37]; a particularly well-known example is stochastic phenotypic switching in bacterial populations [2, 38]. In the simplest setup, the population lives in an environment which can assume one of two different states: a nutrient-rich state, allowing the bacterial population to grow, or an antibiotic, hazardous state that kills off bacteria. Each bacterium switches stochastically and independently from the others between a growth and a resistant phenotype. The growth phenotype prospers in the growth state of the environment, but is instantly killed if the environment is in the antibiotic state. Conversely, the resistant phenotype does not grow in the growth environment, but survives antibiotic treatment. It has been

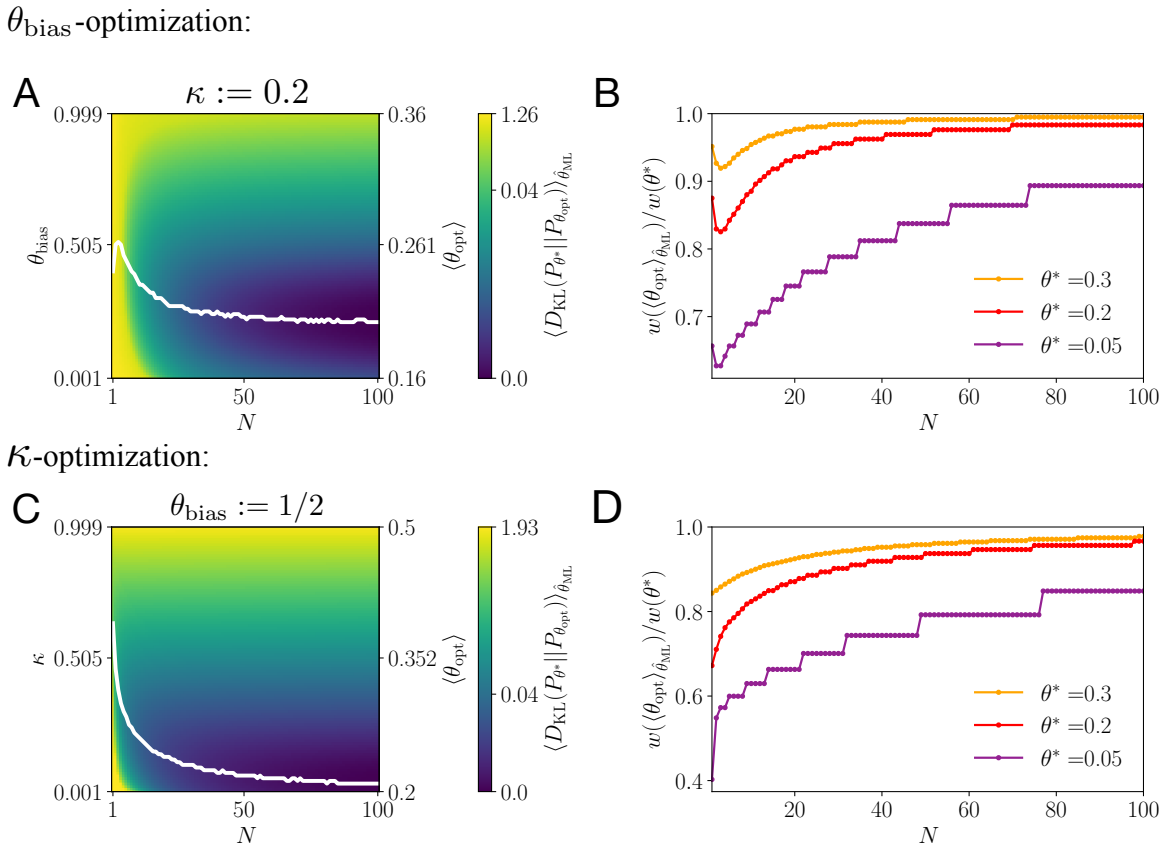

FIG. 3. Optimization of the expected loss with respect to the prior shows a bias to low complexity models. The true model lies at $\theta^* = 0.2$. (A) The bias θ_{bias} is optimised while $\kappa = 0.2$ ($q = 0.25$) is kept fixed. The plot shows the landscape of the expected loss as a function of the bias and the number of observations. We are using a non-linear colormap based on the cumulative distribution of the loss. (B) We show the ratio of the local complexity as measured by Jeffreys prior of the expected optimal to that of the true model for three different true models θ^*. (C) The parameter κ is optimised while the bias $\theta_{\text{bias}} = 1/2$ (imposing isotropy on hyperparameter space). The plot shows the landscape of the expected loss as a function of the parameter κ and the number of observations. (D) We show the ratio of the local complexity as measured by Jeffreys prior of the expected optimal to that of the true model for three different true models θ^*.

shown that in order to maximise population growth, the bacteria should perform Kelly betting (also known as proportional betting) [2, 24]. According to Kelly's theory, the bacteria should stochastically select each phenotype with a probability that is proportional to the probability of occurrence of the corresponding environmental state. We consider the case where the bacteria only have imprecise information about environmental probabilities, show how the "playing it safe" principle implies safer betting strategies in this case, and how these choices translate into optimisation of growth of the bacterial population.

In our toy model of stochastic phenotypic switching, successive discrete environmental states are chosen independently according to a Bernoulli model with a probability θ^* for the nutrient state, and probability $1-\theta^*$ for the antibiotic states. The bacteria can maximise population growth by betting according to Kelly's theory, namely by matching their probabilistic model to the environment. In Kelly betting any mismatch is penalised and the loss is captured by the KL divergence between the true probabilistic model P_{θ^*} and the model P_θ assumed by the bacterial population (the bettor). Instead of the loss, it is also instructive to consider the closely related growth rate function derived by Kelly in [24, 25]:

$$G_{\theta^*}(\theta) = [H(P_{1/2}) - H(P_{\theta^*})] - D_{\text{KL}}(P_{\theta^*}||P_\theta)\,, \quad (22)$$

where $H(\cdot)$ is the entropy. This function captures the long-term population growth rate, i.e., the growth rate after a long succession of environmental states. The term in brackets is the maximal achievable growth rate from which the KL loss is subtracted. In Figure 4A we show the long-term growth rate G_{θ^*} of Kelly betting. The growth rate has a single maximum when the assumed model is equal to the truth, $\theta = \theta^*$ for which the loss is zero. The further away the adopted model lies from the truth, the smaller the growth rate; it can even become negative if the assumed model lies too far away from the truth. We will see (Figures 4A,B,C and S7) that there is a region of *safe* model choices. This region is given by models θ, for which the long-term population growth rate

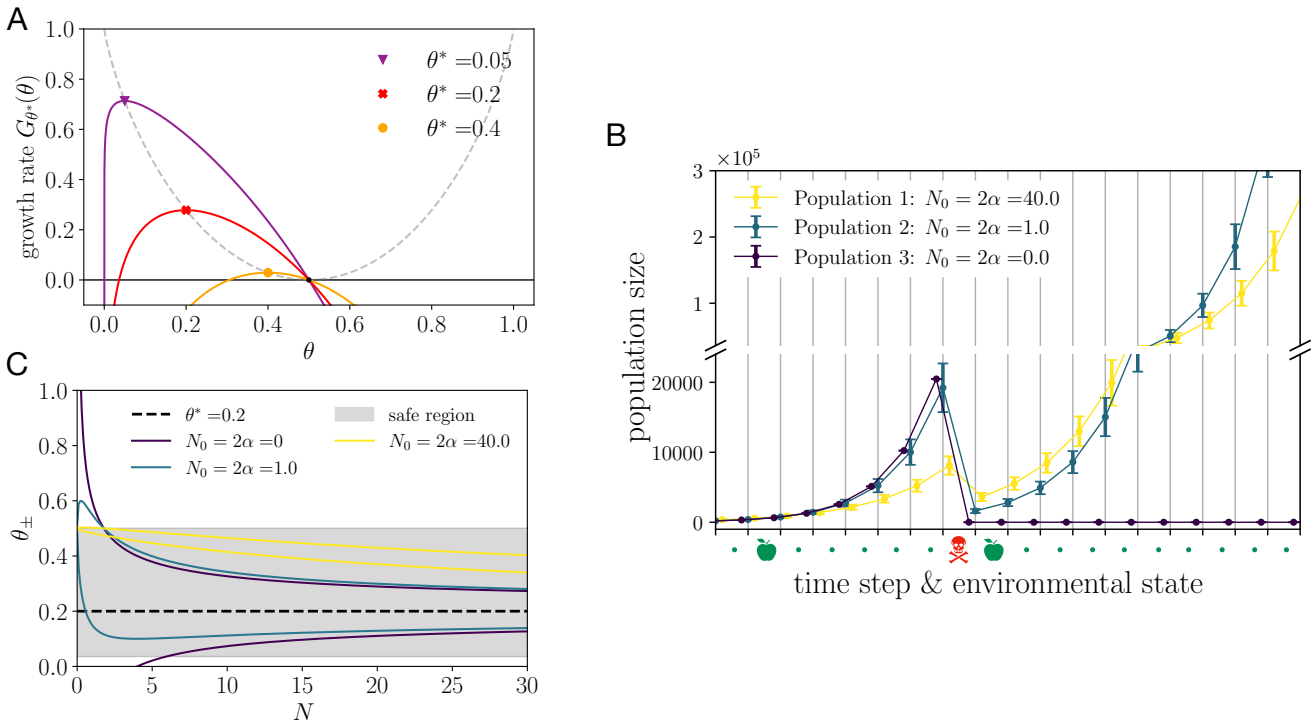

FIG. 4. Bacteria that place bets according to the "playing it safe" principle grow more safely. (A) Long-term growth rate of the bettor's capital for Kelly betting on independent Bernoulli events as a function of the bettor's assumed model θ for three values of the true model θ^* (orange, red, and purple). The maximum is reached when $\theta = \theta^*$. (B) Three bacterial populations, each of initial size 10, are exposed to a sequence of binary environmental states (growth or antibiotic) and the development of population sizes is followed. The individuals of each population stochastically switch between a growth and resistant phenotype independently from one and another. All populations keep a memory of the last $N = 5$ past environmental states, but each population uses a different learning strategy, set by the hyperparameters $N_0 = 2\alpha$ and $\theta_{\text{bias}} = 1/2$. The environmental state sequence contains a single antibiotic state (skull), while all other states are growth states (apple). For each population we show the mean population size and the standard deviation. The population with the least safe strategy (dark blue) grows the fastest, but is extinguished by the antibiotic environment. The population with the safest strategy (yellow) grows more slowly, but shows the smallest loss due to antibiotic treatment. The population with intermediate safety (teal), experiences larger loss after antibiotic treatment, but outgrows the population with highest safety. The learning strategy of intermediate safety represents a compromise between too much risk and too much safety. Population curves are computed from fifty independent trials and data points have a small horizontal offset to increase legibility. (C) The magnitude of fluctuations of the Bayesian conditional optimal model $\theta(\hat{\theta}; N_0, \theta_{\text{bias}} = 1/2)$ is shown as the plus and minus standard error of the conditional optimal model, $\theta_\pm = \theta(\theta^* \pm \sigma_N; N_0, \theta_{\text{bias}} = 1/2)$, as a function of the number of observations N and the prior choice. Three prior choices from the family of the conjugate prior, set by different values of the hyperparameter N_0 (dark blue, teal and yellow) and the fixed bias $\theta_{\text{bias}} = 1/2$. The choice $N_0 = 0$ corresponds to maximum-likelihood estimation. For large number N of observations, the fluctuations diminish ($q \to 0$ and $\kappa \to 0$) and the conditional optimal model θ converges to the true model $\theta^* = 0.2$ (dotted line). The safe region (gray) constitutes models θ for which the growth rate function $G_{\theta^*}(\theta)$ is non-negative. For larger values of N_0, the standard error of the conditional optimal model θ is pulled into the safe region, such that a smaller loss in the growth rate is expected.

is non-negative. Models in this region are safe, because while they might not be optimal, they guarantee that there are no long-term losses.

We now assess, for Kelly betting, how the amount of information available about the true model and the prior choice affect the magnitude of fluctuations in the assumed model. The empirical observations are subject to statistical fluctuations and as discussed above translate to the conditional optimal model. To quantify the magnitude of fluctuations, we consider the *standard error* of the empirical estimate $\hat{\theta}$. For N observations the standard error is given by $\sigma_N = \sqrt{\theta^*(1-\theta^*)/N}$. The fluctuations of the empirical estimate, determine the fluctuations of the conditional optimal model. The magnitude of fluctuations of the conditional optimal strategy Eq. (17) within one standard deviation, is given by the bounds $\theta_\pm \equiv \theta(\hat{\theta}^\pm_{\text{ML}}; N_0, \theta_{\text{bias}})$, where $\hat{\theta}^\pm_{\text{ML}} = \theta^* \pm \sigma_N$.

We will examine a toy model of bacterial phenotypic switching, in which environmental states are chosen and phenotypes expressed simultaneously at discrete time steps. The bacterial population then performs collective betting. We know that with perfect information

about the environmental probabilities the optimal strategy would be to perform proportional betting. This means matching the probability of expressing the growth phenotype with that of the occurrence of the growth state of the environment. However, the bacteria do not know the probability of environmental states. We will suppose that instead they keep an implicit memory of the previous N environmental states, which they use to compute empirical probabilities.

To illustrate the effect of safe learning, we compare the growth of three types of bacterial colonies distinguished only by their learning strategy, as they respond to a sequence of environmental states which include a single antibiotic episode. In Figure 4B we show how the size of each population develops over time. All bacteria have a memory over the past $N = 5$ environmental states. In the extreme case of maximum-likelihood estimation ($N_0 = 0$, dark blue curve), the population is killed completely by the antibiotic state. In contrast, the safer the learning strategy (larger N_0), the smaller the loss a bacterial population suffers due to antibiotic treatment. After the antibiotic treatment, bacteria with the safest strategy recover their numbers most quickly (yellow curve), but they are eventually overtaken by bacteria with less safe learning strategies (teal curve). The observed trade-off between the ability to survive antibiotic treatment and achieve high growth rates is the reason for the existence of an optimally safe learning strategy. The precise choice of optimal strategy depends on the statistics of the environment, as well as potentially other factors such as a future horizon over which safe growth needs to be guaranteed.

In environments which fluctuate more the optimal learning strategy will need to be safer, because the probabilities inferred from a finite memory of past events will have a larger variance. However, with a maximally safe learning strategy ($\theta = 0.5$), the bacterial population will not be able to achieve any long-term growth, but only keep its population size constant as shown in Figure 4A. In turn, this means that long-term physical growth is only possible in environments which are not too random. In fact, the more environments are suitable for life to prosper, the closer they are to being deterministic. Near-deterministic environments are thus livable, but as we have shown, rare, yet potentially catastrophic events require organisms to "play it safe" to survive in even in such worlds.

Finally, in Figure 4C we contrast different prior choices to show how these affect model safety and plot the standard error bounds of the conditional optimal model $\theta_\pm$ as function of the observations N. In this example we fix $\theta_{\text{bias}} = 1/2$ which, by Eq. (21), implies $N_0 = 2\alpha$. We observe that a large value of N_0 reduces the magnitude of statistical fluctuations and biases the conditional optimal model θ_{opt} towards a probability of $1/2$, i.e. minimal complexity. Overall, the inferred model is pulled into the safe growth region and larger values of N_0 thus result in a safer learning strategy. From (*Appendix*, Section F and Figure S6) we see that this effect is the more pronounced the closer the true model lies to the boundary of parameter space. Notably, maximum-likelihood estimation ($N_0 = 2\alpha = 0$) is the least safe strategy. For a large number of observations N, the fluctuations diminish and the inferred model converges to the true model. However, we observe that safer learning strategies (larger N_0) converge to the true model more slowly than less safe strategies. Furthermore, in the limit of maximum safety $q = N_0/N \to \infty$, the optimal model $\theta_{\text{opt}} \to 1/2$ for which the growth rate vanishes. This illustrates the trade-off between safe learning and growth maximisation.

DISCUSSION

We have proposed that biological systems acting with bounded information should "play it safe" by biasing themselves towards models of the environment that are less complex, in the sense of being easier to infer despite statistical fluctuations in limited observed data. We derived this result analytically for exponential families, a very general class of probability distributions, including, e.g., the widely employed multinomial, Poisson, and Gaussian distributions. In the context of collective betting, our results imply that optimal adaptation involves a balance between risk created by uncertainty in inference of the latent distribution, and making full use of the potential for growth or reward.

We illustrated these ideas in the example of bacteria in an environment which mostly supports growth, but occasionally manifests an antibiotic. In this situation, it pays to maintain diversity between high-growth and low-growth but resistant phenotypes, as a bet-hedging mechanism [2, 38]. Indeed, wild type *Escherichia coli* show a resistant fraction of $\sim 10^{-6} - 10^{-5}$ [39]. We showed how this sort of strategy arises in terms of "playing it safe" in an uncertain world. It will be interesting in the future to apply our analysis to other situations where populations of organisms must make decisions with limited information, for example in bacterial chemotaxis [40]

We also cast our analysis in the language of Bayesian inference, and noted that the bias towards simpler models can be implemented by an appropriate choice of prior. The optimal prior depends on knowledge of the true model, or rather its distribution in cases where the generative process in the world itself changes stochastically. It would be interesting to determine the sorts of evolutionary dynamics that would allow systems to adjust their priors to "play it safe" across a given distribution of generative processes.

Another interesting question is to understand how our considerations operate if we use the information maximizing discrete priors discussed in [41, 42]. We instead used the continuous conjugate priors because it seemed natural that a biological agent might have ways of tuning over a continous family of choices, e.g., by adjusting concentration levels of a protein, or activity levels of a

circuit. By contrast, it is harder to imagine biological methods for selecting over a discrete set of values. Neverthless, it is important and interesting to consider how our "playing it safe" principle might interact with the discrete priors of [41, 42] because the authors of these articles are precisely concerned with the limitations imposed on inference by having a finite amount of data. Indeed, their discrete priors depend on the number of data points N, and are designed to maximize the amount of information gained about the parameters from each observation. Perhaps there is some way for an organism to incorporate an understanding of the amount of data it is able to accumulate from the past, to effectively discretize its prior on the parameters of the world. It would be useful to understand whether and how some variant of our "playing it safe" principle could be realized in that context. Models of biological systems often depend on a very large number of parameters which have been shown to be close to impossible to determine with high precision [43]. This important observation may suggest that we should define a "playing it safe" principle on effective parameters.

Finally, in a time-varying world, inferences are only useful up to some future time horizon. This restricted future utility should limit the value and precision required of the optimal model, while also increasing the imperative for "safety" in light of possible future changes in the world. We can expect trade-offs associated with keeping past memory to play a role in such situations [44, 45]. Perhaps this observation also has a bearing on the widespread observation that Occam's Razor, understood as a preference for simplicity or parsimony, seems to be an organizing principle for mental functions [46–51], and that biases towards simple models are helpful in many statistical settings that require inference and prediction [22, 41, 52–55].

MATERIALS AND METHODS

Information geometry basics

We define the FIM by Taylor expansion of the KL divergence between two parametric distributions $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\phi}}$. The Taylor expansion between two models which are close in parameter space $\boldsymbol{\phi} \equiv \boldsymbol{\theta}^* - \delta\boldsymbol{\theta}$ is given by

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}^*-\delta\boldsymbol{\theta}}) = \frac{1}{2}\mathcal{I}_{ij}\delta\theta^i\delta\theta^j + \dots\,, \tag{23}$$

where summation over repeated indices is implied and the FIM $\mathcal{I}$ is given by the coefficient of the quadratic term

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \partial_{\delta\theta^i}\partial_{\delta\theta^j} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}||P_{\boldsymbol{\theta}-\delta\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=\mathbf{0}}\,. \tag{24}$$

As an important example we consider the multinomial distribution

$$P(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{B(\mathbf{x})}\prod_{i=1}^{d}\theta_i^{x_i}\,,\ \text{with}\ \ \sum_{i=1}^{d}\theta_i = 1\,,\ x_i \in \{0,...,N\}\,, \tag{25}$$

$$\sum_{i=1}^{d} x_i = N\,,\text{and multivariate Beta } B(\mathbf{x}) \equiv \frac{\prod_{i=1}^{d}\Gamma(x_i)}{\sum_{i=1}^{d}\Gamma(x_i)}$$

where the parameters θ_i are probabilities and the mean is $N\boldsymbol{\theta}$. For the multinomial distribution Jeffreys prior takes the form of a Dirichlet distribution

$$w(\boldsymbol{\theta}) = \frac{\Gamma(d/2)}{\sqrt{\pi}^d}\prod_{i=1}^{d}\frac{1}{\sqrt{\theta_i}} \tag{26}$$

with $\theta_d = 1 - \sum_{i=1}^{d-1}\theta_i$. From this form we see that Jeffreys prior has a minimum at $\theta_i = 1/d$, but diverges when any of the parameters tend to zero, $\theta_i \to 0$. Analogously, in (*Appendix*, Sections A) we show for the Poisson and Gaussian distribution, respectively, that Jeffreys prior diverges on the boundaries of the standard parameterization of these distributions.

Exponential families

An *exponential family* consists of a collection of densities

$$P_{\boldsymbol{\eta}}(x) = \exp\left(\boldsymbol{\eta}^T\cdot\mathbf{t}(x) - F(\boldsymbol{\eta}) + k(x)\right)\,, \tag{27}$$

where $\boldsymbol{\eta} = [\eta_1\ \eta_2 \dots\ \eta_d]^T$ is the *canonical parameter* associated with the *sufficient statistic* $\mathbf{t}(x)$, $k(x)$ is the auxiliary carrier term, and F is the log-partition function:

$$F(\boldsymbol{\eta}) = \log\left(\int \exp(\boldsymbol{\eta}^T\cdot\mathbf{t}(x) + k(x))\mathrm{d}\nu(x)\right)\,, \tag{28}$$

where $\nu(x)$ the Lebesgue or counting measure. The log-partition function is convex. If the sufficient statistic is a collection of d functions, the canonical parameter $\boldsymbol{\eta}$ takes values in the convex set

$$\Omega = \{\boldsymbol{\eta}\in\mathbb{R}^d \,|\, F(\boldsymbol{\eta}) < +\infty\}\,. \tag{29}$$

We consider exponential families where Ω is an open set, in which case the exponential family is called *regular*. Since the log-partition function is a convex function, the set Ω is also convex. We restrict ourselves to minimal exponential families, which are defined by not having linear constraints amongst the parameters and also not amongst the components of the sufficient statistic. Three examples appear in the table below.

Distribution	$\mathbf{t}(x)$	$F(\boldsymbol{\eta})$	$k(x)$
Multinomial	$[x_1\ x_2\ \dots\ x_{d-1}]^T$	$N\log\left(1+\sum_{i=1}^{d-1}e^{\eta_i}\right)$	$\log\frac{N!}{\prod_{i=1}^{d}x_i!}$
Poisson	x	e^{η}	$-\log x!$
Gaussian	$[x\ x^2]^T$	$-\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log(-\pi/\eta_2)$	0

An alternative set of parameters is given by the *mean parameters* $\boldsymbol{\theta}$ defined as the expectation of the sufficient statistic

$$\boldsymbol{\theta} = \mathbb{E}[\mathbf{t}(x)] = \nabla_{\boldsymbol{\eta}} F(\boldsymbol{\eta}) \,, \tag{30}$$

where the last relation can be verified by direct calculation. The distribution in mean parameterisation is obtained by inverting this defining equation of the mean parameter and substituting for $\boldsymbol{\eta}$ in the canonical form of the exponential family. The analogue of the log-partition function F for the mean parameters is the dual function of F defined by the Legendre-Fenchel transform

$$F^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{\eta}\in\Omega} \left\{ \boldsymbol{\eta}^T \cdot \boldsymbol{\theta} - F(\boldsymbol{\eta}) \right\} \,. \tag{31}$$

The dual function F^* can thus be obtained by explicit substitution for $\boldsymbol{\eta}$ in terms of $\boldsymbol{\theta}$, after inverting the defining equation of the mean parameter. Finally, we note a relation between F^* and negative entropy $-H(P_{\boldsymbol{\eta}(\boldsymbol{\theta})})$ of the distribution

$$\begin{aligned} -H(P_{\boldsymbol{\eta}(\boldsymbol{\theta})}) &= \int P_{\boldsymbol{\eta}(\boldsymbol{\theta})} \log P_{\boldsymbol{\eta}(\boldsymbol{\theta})} \mathrm{d}\nu(x) && (32)\\ &= \mathbb{E}[\boldsymbol{\eta}^T \cdot \mathbf{t}(x) - F(\boldsymbol{\eta})] + \mathbb{E}[k(x)] && (33)\\ &= F^*(\boldsymbol{\theta}) + \mathbb{E}[k(x)] \,, && (34) \end{aligned}$$

where we have used the definition of the mean parameters and the defining relation of the dual function. For distributions which afford a vanishing auxiliary carrier term $k(x)$, the dual function is thus given by the negative entropy. This is for example the case for the Gaussian distribution.

ACKNOWLEDGMENTS

This work was supported in part by the Simons Foundation MMLS Grant 400425 and by the NIH grant R01EB026945. VB thanks the Galileo Galileo Institute where he was a Simons Visiting Scientist, and the Aspen Center for Physics (supported by NSF grant PHY-1607611), for hospitality as this work was completed.

Appendix A: Jeffreys prior for discrete and continuous distributions

We compute Jeffreys prior for common examples of discrete and continuous distributions. To this end, we define the Fisher information metric (FIM) as the coefficient of the second order term in the Taylor expansion of the Kullback-Leibler (KL) divergence Eq. (1). If two parameterised distributions, are defined by discrete sets of parameters $\boldsymbol{\theta} = \{\theta_1 \ldots, \theta_d\}$ and $\boldsymbol{\phi} = \{\phi_1 \ldots, \phi_d\}$, respectively, we consider the Taylor expansion of the KL divergence between two models which are close in parameter space $\boldsymbol{\phi} \equiv \boldsymbol{\theta} - \delta\boldsymbol{\theta}$:

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}} || P_{\boldsymbol{\theta}-\delta\boldsymbol{\theta}}) = \frac{1}{2}\mathcal{I}_{ij}\delta\theta^i\delta\theta^j + \ldots \,, \tag{A1}$$

where summation over repeated indices is implied and $\mathcal{I}$ is the FIM, given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \partial_{\delta\theta^i}\partial_{\delta\theta^j} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}} || P_{\boldsymbol{\theta}-\delta\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=\mathbf{0}} \,. \tag{A2}$$

The normalised Jeffreys prior is computed from the FIM through the relation

$$w(\boldsymbol{\theta}) = \frac{\sqrt{\det \mathcal{I}(\boldsymbol{\theta})}}{\int \mathrm{d}^d\boldsymbol{\psi} \sqrt{\det \mathcal{I}(\boldsymbol{\psi})}} \,. \tag{A3}$$

1. Multinomial distribution

We define the multinomial distributions with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ as:

$$P_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{B(\mathbf{x})} \prod_{i=1}^{d} \theta_i^{x_i} \,,\; Q_{\boldsymbol{\theta}'}(\mathbf{x}) = \frac{1}{B(\mathbf{x})} \prod_{i=1}^{d} \theta'^{x_i}_i \tag{A4}$$

with

$$\sum_{i=1}^{d}\theta_i = \sum_{i=1}^{d}\theta'_i = 1\,,\ x_i \in \{0,...,N\} \text{ and } \sum_{i=1}^{d} x_i = N\,. \tag{A5}$$

The KL divergence between the two distributions is given by

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}||Q_{\boldsymbol{\theta}'}) = \sum_{\mathbf{x}} \frac{1}{B(\mathbf{x})}\prod_{i=1}^{d}\theta_i^{x_i}\log\left(\frac{\prod_j \theta_j^{x_j}}{\prod_j \theta'^{x_j}_j}\right) \tag{A6}$$

$$= \sum_j (\log\theta_j - \log\theta'_j)\left(\sum_{\mathbf{x}}\frac{x_j}{B(\mathbf{x})}\prod_{i=1}^{d}\theta_i^{x_i}\right) \tag{A7}$$

$$= \sum_j (\log\theta_j - \log\theta'_j) N\theta_j\,, \tag{A8}$$

where in the last step we have evaluated the mean of the multinomial distribution. The next step is to Taylor expand the KL divergence around $\boldsymbol{\theta}$, by setting $\boldsymbol{\theta}' = \boldsymbol{\theta} - \delta\boldsymbol{\theta}$. For the Taylor expansion we find

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}||Q_{\boldsymbol{\theta}'}) = \sum_j \left(\frac{\delta\theta_j}{\theta_j} + \frac{\delta\theta_j^2}{2\theta_j^2} + O(\delta\theta_j^3)\right) N\theta_j = \sum_j \left(\frac{\delta\theta_j^2}{2\theta_j^2} + O(\delta\theta_j^3)\right) N\theta_j\,, \tag{A9}$$

where in the last step the linear term vanishes since $\sum_i \theta_i = \sum_i \theta'_i = 1$ and $\theta'_i = \theta_i - \delta\theta_i$. From the quadratic term we read off the form of the FIM

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = N\frac{\delta_{ij}}{\theta_i}\,, \tag{A10}$$

where δ_{ij} is the Kronecker delta. Taking the square root of the determinant and computing the normalisation factor finally gives us Jeffreys prior:

$$w(\boldsymbol{\theta}) = Dir(\boldsymbol{\theta}; 1/2,\ldots,1/2) = \frac{\Gamma(d/2)}{\sqrt{\pi}^d}\frac{1}{\sqrt{1-\sum_{i=1}^{d-1}\theta_i}}\prod_{i=1}^{d-1}\frac{1}{\sqrt{\theta_i}}\,, \tag{A11}$$

where Dir is the Dirichlet distribution.

2. Poisson distribution

We consider two Poisson distributions with parameters λ' and λ:

$$P_{\lambda'}(n) = \frac{\lambda'^n e^{-\lambda'}}{n!}\,,\ Q_\lambda(n) = \frac{\lambda^n e^{-\lambda}}{n!}\,. \tag{A12}$$

The KL divergence between the two distributions is given by

$$D_{\mathrm{KL}}(P_{\lambda'}||Q_\lambda) = \sum_n \frac{\lambda'^n e^{-\lambda'}}{n!}\log\left(\frac{\lambda'^n e^{-\lambda'}}{n!}\frac{n!}{\lambda^n e^{-\lambda}}\right) = \lambda'\log\frac{\lambda'}{\lambda} - \lambda' + \lambda\,. \tag{A13}$$

The next step is to Taylor expand the KL divergence around λ', by setting $\lambda = \lambda' + \delta\lambda$. For the Taylor expansion we find

$$D_{\mathrm{KL}}(P_{\lambda'}||Q_\lambda) = \frac{\delta\lambda^2}{2\lambda'} + O(\delta\lambda^3)\,. \tag{A14}$$

From the quadratic term we read off the form of the FIM

$$\mathcal{I}(\lambda') = \frac{1}{\lambda'}\,. \tag{A15}$$

Taking the square root gives us Jeffreys prior:

$$w(\lambda') = \frac{1}{\sqrt{\lambda'}}\,, \tag{A16}$$

Jeffreys prior is shown in Figure S1. It diverges on the boundary of parameter space for $\lambda' \to 0$ and approach zero as for $\lambda' \to \infty$.

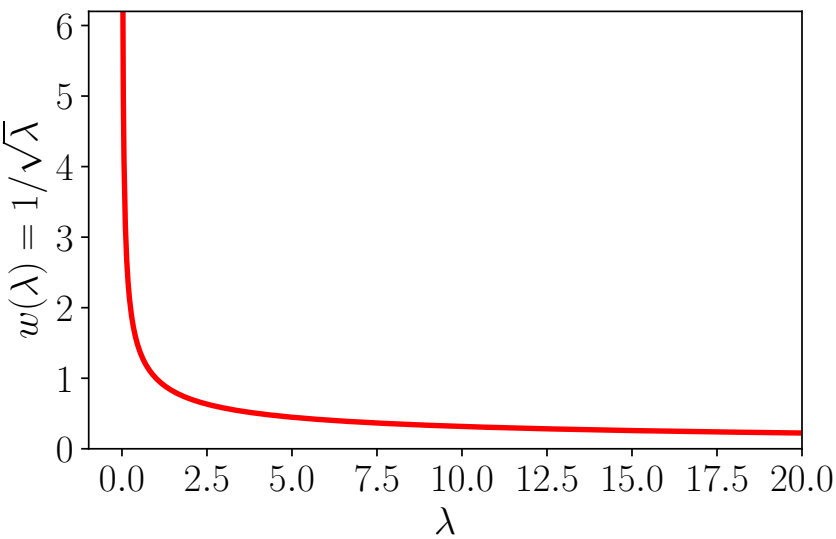

FIG. S1. Jeffreys prior for the Poisson distribution with rate λ, diverges at the boundary of parameter space for $\lambda \to 0$.

3. Gauss distribution

The KL divergence for continuous distributions is defined as

$$D_{\mathrm{KL}}(P||Q) = \int \mathrm{d}x\, P(x) \log \frac{P(x)}{Q(x)}\,. \tag{A17}$$

The KL divergence between two Gaussian distributions with mean μ and variances σ^2 and σ'^2

$$P_{\mu,\,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,,\; P_{\mu,\,\sigma'^2}(x) = \frac{1}{\sqrt{2\pi\sigma'^2}} e^{-\frac{(x-\mu)^2}{2\sigma'^2}}\,, \tag{A18}$$

is given by

$$D_{\mathrm{KL}}(P_{\mu,\,\sigma^2}||P_{\mu,\,\sigma'^2}) = -\frac{1}{2}\left[\log\sigma^2 - \log\sigma'^2 + 1 - \frac{\sigma^2}{\sigma'^2}\right]\,. \tag{A19}$$

The next step is to Taylor expand the KL divergence around σ, by setting $\sigma' = \sigma + \delta\sigma$. For the Taylor expansion we find

$$D_{\mathrm{KL}}(P_{\mu,\,\sigma^2}||P_{\mu,\,\sigma'^2}) = \frac{1}{2}\left[\partial^2_{\sigma'} D_{\mathrm{KL}}(P_{\mu,\,\sigma^2}||P_{\mu,\,\sigma'^2})\right]_{\delta\sigma=0}\delta\sigma^2 + O(\delta\sigma^3) = -\frac{1}{4}\left[\frac{2}{\sigma'^2} - 6\frac{\sigma^2}{\sigma'^4}\right]_{\delta\sigma=0}\delta\sigma^2 + O(\delta\sigma^3) \tag{A20}$$

From this we find the FIM:

$$\mathcal{I}(\sigma) = \frac{1}{\sigma^2} \tag{A21}$$

and Jeffreys prior

$$w(\sigma) = \frac{1}{\sigma}\,. \tag{A22}$$

Jeffreys prior for the Gaussian is shown in Figure S2. It diverges on the boundary of parameter space as $\sigma \to 0$ and approaches zero for $\sigma \to \infty$.

Appendix B: Standard error and diverging KL divergence

Consider the KL divergence for the Bernoulli model

$$D_{\mathrm{KL}}(P_{\theta^*}||P_\theta) = H_{\theta^*}(\theta) - H_{\theta^*}(\theta^*)\,, \tag{B1}$$

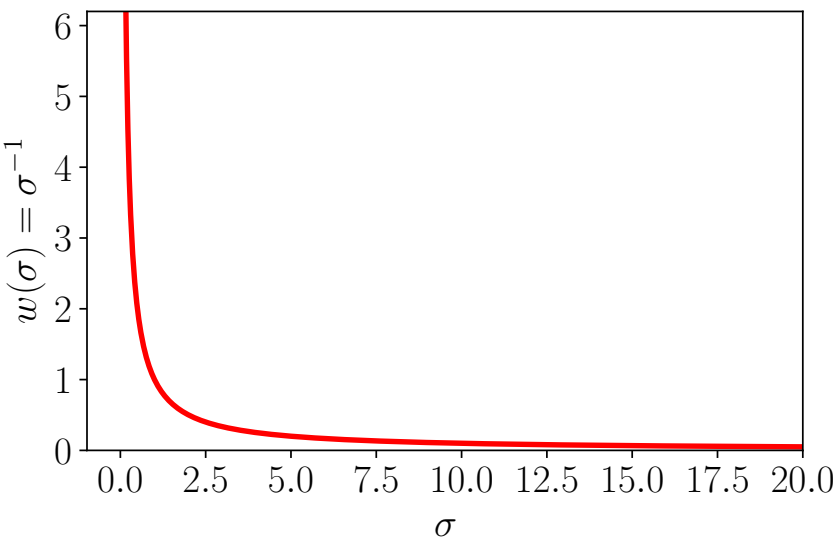

FIG. S2. Jeffreys prior for the Gaussian model with variance σ^2, diverges at the boundary of parameter space for $\sigma \to 0$.

where $H_x(y)$ is the cross-entropy function

$$H_x(y) = -x \log y - (1-x)\log(1-y)\,. \tag{B2}$$

For $\theta \to 0, 1$, the KL divergence diverges to plus infinity. What is the behaviour of the KL divergence when we try to estimate the value of θ using data from many Bernoulli trials generated by the true model θ^*? In particular, how does the KL divergence behave when θ^* lies very close to a boundary of parameter space?

To answer this question we consider the empirically estimated parameter $\hat{\theta}$. This parameter is subject to statistical fluctuations due to finite data size. In general, the standard error for a parameter estimated from N independent trials is given by

$$\sigma_N = \frac{\sigma}{\sqrt{N}}\,, \tag{B3}$$

where σ is the standard deviation of the distribution we are sampling from. Within the standard error the estimated parameter lies within the bounds

$$\hat{\theta}_\pm = \theta^* \pm \sigma_N\,. \tag{B4}$$

For the Bernoulli model, the empirical estimate of the probability parameter is obtained as the mean number of events observed in N independent Bernoulli trials and σ is the standard deviation of the Bernoulli distribution

$$\sigma = \sqrt{\theta^*(1-\theta^*)} = \theta^*\sqrt{1/\theta^* - 1}\,. \tag{B5}$$

The statistically expected bounds are

$$\hat{\theta}_\pm = \theta^*\left(1 - \sqrt{\frac{1/\theta^* - 1}{N}}\right)\,. \tag{B6}$$

Let us now focus on the boundary at $\theta = 0$ and recall that the KL divergence diverges at this boundary. Therefore, one way to check whether we can expect the KL divergence to diverge for a given model θ^* and number of observations N, is to check whether the boundary $\theta = 0$ lies within the standard error bound $\hat{\theta}_-$. From the last equation we see that the boundary is included if the condition

$$\sqrt{\frac{1/\theta^* - 1}{N}} > 1 \tag{B7}$$

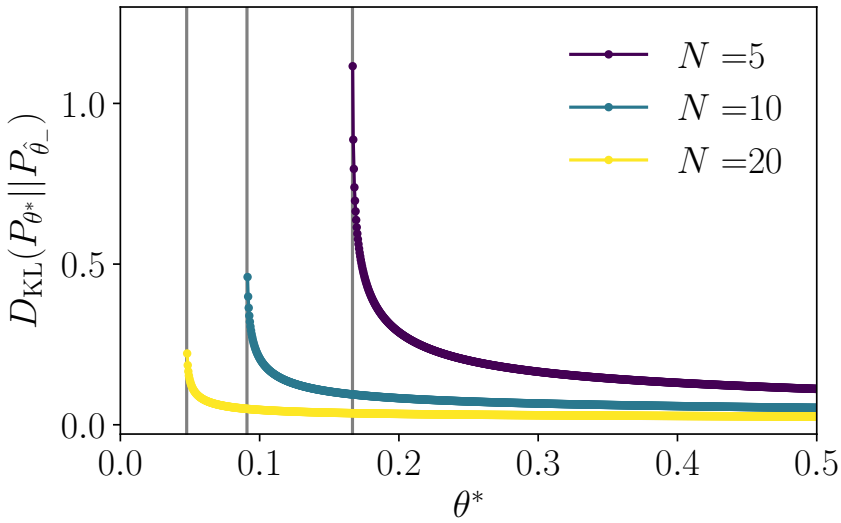

FIG. S3. Diverging behaviour of $D_{\mathrm{KL}}(P_{\theta^*}||P_{\hat{\theta}_{-}=\theta^*-\sigma_N})$ for three values of N. Vertical gray lines show the position of diverging behaviour at $1/(N+1)$.

is satisfied. Rearranging this inequality gives

$$\theta^* < \frac{1}{N+1}\,. \tag{B8}$$

Thus true models which are within $1/(N+1)$ of the boundary, the boundary where the KL divergence diverges lies within the standard error and statistically we can expect the KL divergence to diverge. See Figure S3 for examples plots.

For the Poisson model, the empirical estimate of the parameter lambda is obtained as the mean number of events observed in N independent Poisson trials and σ is the standard deviation of the Poisson distribution

$$\sigma_N = \sqrt{\frac{\lambda^*}{N}} = \lambda^* \frac{1}{\sqrt{\lambda^* N}}\,. \tag{B9}$$

The statistically expected bounds are

$$\hat{\lambda}_{\pm} = \lambda^* \pm \sigma_N = \lambda^* \left(1 \pm \frac{1}{\sqrt{\lambda^* N}}\right)\,. \tag{B10}$$

The boundary of parameter space lies at $\lambda = 0$. Therefore we consider $\hat{\lambda}_-$ as the important bound and the condition for the boundary of parameter space to be included is

$$\frac{1}{\sqrt{\lambda^* N}} > 1\,, \tag{B11}$$

or

$$\lambda^* < \frac{1}{N}\,. \tag{B12}$$

Therefore, for any given N we can find a models λ^* which lie close enough to the boundary according to the above condition, and for which we should expect the KL divergence to diverge.

Appendix C: Optimality condition

1. The optimality condition for exponential families in mean parameterisation

This derivation makes use of the fact that the Kullback-Leibler divergence between two distributions from the same exponential family in canonical form, can be expressed as the Bregman divergence (defined with respect to the dual

of the log partition function, F^*) between the mean parameters $\boldsymbol{\theta}$:

$$D_{\mathrm{KL}}(Q_{\boldsymbol{\eta}^*}||Q_{\boldsymbol{\eta}}) = B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta})\,. \tag{C1}$$

Writing out the definition of the Bregman divergence, the right-hand side is of the form

$$B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}) = F^*(\boldsymbol{\theta}^*) - F^*(\boldsymbol{\theta}) - [\boldsymbol{\theta}^* - \boldsymbol{\theta}]^T \cdot \nabla_{\boldsymbol{\theta}} F^*(\boldsymbol{\theta})\,, \tag{C2}$$

where

$$\boldsymbol{\theta} = \kappa \boldsymbol{\theta}_{\mathrm{bias}} + (1-\kappa)\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\,. \tag{C3}$$

To compute the condition for the optimal model, we take the derivative of the Bregman divergence with respect to the bias:

$$\frac{1}{\kappa}\nabla_{\theta_{\mathrm{bias}\,i}} B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}) = \nabla_{\theta_i} B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}) = -\nabla_{\theta_i} F^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^{*T} \cdot \nabla_{\theta_i}\nabla_{\boldsymbol{\theta}} F^*(\boldsymbol{\theta}) + \nabla_{\theta_i}\{\boldsymbol{\theta}^T \cdot \nabla_{\boldsymbol{\theta}} F^*(\boldsymbol{\theta})\} \tag{C4}$$

$$= -\nabla_{\theta_i} F^*(\boldsymbol{\theta}) - \sum_{j=1}^{d} \theta_j^* \nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta}) + \nabla_{\theta_i} F^*(\boldsymbol{\theta}) + \sum_{j=1}^{d} \theta_j \nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta}) \tag{C5}$$

$$= -\sum_{j=1}^{d} [\theta_j^* - \theta_j] \nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta})\,. \tag{C6}$$

Next, we use the relation

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta})\,, \tag{C7}$$

relating the Fisher information metric in mean parameterization to the dual function. We derive this relation starting with Eq. (8)

$$D_{\mathrm{KL}}(Q_{\boldsymbol{\eta}^*(\boldsymbol{\theta}^*)}||Q_{\boldsymbol{\eta}(\boldsymbol{\theta})}) = B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta})\,, \tag{C8}$$

replacing $\boldsymbol{\theta} = \boldsymbol{\theta}^* - \delta\boldsymbol{\theta}$, and Taylor expanding in $\delta\boldsymbol{\theta}$. Using the definition of the Bregman divergence Eq. (C2) and noting that $\nabla_{\delta\theta_i} = -\nabla_{\theta_i}$, the coefficients of the first three orders in $\delta\boldsymbol{\theta}$ are given by

$$B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}^* - \delta\boldsymbol{\theta})\big|_{\delta\boldsymbol{\theta}=0} = 0\,, \tag{C9}$$

$$\nabla_{\delta\theta_i} B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}^* - \delta\boldsymbol{\theta})\big|_{\delta\boldsymbol{\theta}=0} = \left[\nabla_{\theta_i} F^*(\boldsymbol{\theta}) - \nabla_{\theta_i} F^*(\boldsymbol{\theta}) + \sum_j \delta\theta_j \nabla_{\theta_i}\nabla_{\theta)j} F^*(\boldsymbol{\theta})\right]_{\delta\boldsymbol{\theta}=0} = 0\,, \tag{C10}$$

$$\nabla_{\delta\theta_k}\nabla_{\delta\theta_i} B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}^* - \delta\boldsymbol{\theta})\big|_{\delta\boldsymbol{\theta}=0} = \Big[-\nabla_{\delta\theta_k}\nabla_{\theta_i} F^*(\boldsymbol{\theta}) + \nabla_{\delta\theta_k}\nabla_{\theta_i} F^*(\boldsymbol{\theta}) + \delta_{kj}\nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta}) \tag{C11}$$

$$- \sum_j \delta\theta_j \nabla_{\delta\theta_k}\nabla_{\theta_i}\nabla_{\theta_j} F^*(\boldsymbol{\theta})\Big]_{\delta\boldsymbol{\theta}=0} = \nabla_{\theta_i}\nabla_{\theta_k} F^*(\boldsymbol{\theta}^*)\,. \tag{C12}$$

Using this relation, we rewrite the above condition as

$$\frac{1}{\kappa}\nabla_{\boldsymbol{\theta}_{\mathrm{bias}}} B_{F^*}(\boldsymbol{\theta}^* : \boldsymbol{\theta}) = -\mathcal{I}(\boldsymbol{\theta})\,[\boldsymbol{\theta}^* - \boldsymbol{\theta}] = \mathcal{I}(\boldsymbol{\theta})\delta\boldsymbol{\theta}\,, \tag{C13}$$

where $-\delta\boldsymbol{\theta} \equiv \boldsymbol{\theta}^* - \boldsymbol{\theta}$. Thus, in mean parameterisation, the optimality constraint in Eq. (6) of the main text, takes the simple form

$$\frac{1}{\kappa}\nabla_{\boldsymbol{\theta}_{\mathrm{bias}}} \big\langle \mathcal{L}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\big\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \big\langle \mathcal{I}(\boldsymbol{\theta})\delta\boldsymbol{\theta}\big\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0\,. \tag{C14}$$

2. The optimality condition for the Bernoulli model in mean parameterization derived from Taylor expansion

As an example, we evaluate the optimality condition

$$\nabla_{\boldsymbol{\theta}_{\mathrm{bias}}} \big\langle D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\big\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0 \tag{C15}$$

for the Bernoulli model written in mean parameters. In this derivation we will use Taylor expansion of the KL divergence and find that all but a single term of the infinite series cancel to yield a simplified form of the condition in Eq. (9) of main text.

We consider KL divergence between two Bernoulli distributions, denoted by $P_{\boldsymbol{\theta}^*}$ and $P_{\boldsymbol{\theta}}$, respectively:

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = D_{\mathrm{KL}}(P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}||P_{\boldsymbol{\theta}}) = \sum_x P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}} \log\left(\frac{P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}}{P_{\boldsymbol{\theta}}}\right), \tag{C16}$$

where $\boldsymbol{\theta}^* \equiv \boldsymbol{\theta} + \delta\boldsymbol{\theta}$ and expand the right-hand side in $\delta\boldsymbol{\theta}$. For the first few orders we find

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = 0\,, \tag{C17}$$

$$\partial D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = 1\,, \tag{C18}$$

$$\partial^2 D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = 1/\boldsymbol{\theta}\,, \tag{C19}$$

$$\partial^3 D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = -1/\boldsymbol{\theta}^2\,, \tag{C20}$$

$$\partial^4 D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = 2/\boldsymbol{\theta}^3\,, \tag{C21}$$

$$\partial^5 D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\Big|_{\delta\boldsymbol{\theta}=0} = -6/\boldsymbol{\theta}^4\,, \tag{C22}$$

$$\vdots$$

where $\partial \equiv \partial_{\delta\boldsymbol{\theta}}$ and powers on the right-hand side act element-wise. The Taylor expansion is thus

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = \sum_i \left\{\delta\theta_i + \frac{1}{2!}\frac{\delta\theta_i^2}{\theta_i} - \frac{1}{3!}\frac{\delta\theta_i^3}{\theta_i^2} + \frac{1}{4!}\frac{2\delta\theta_i^4}{\theta_i^3} - \frac{1}{5!}\frac{6\delta\theta_i^5}{\theta_i^4} + O(\delta_i^6)\right\} \tag{C23}$$

$$= 1 + \sum_i \left\{\frac{1}{2!}\frac{\delta\theta_i^2}{\theta_i} - \frac{1}{3!}\frac{\delta\theta_i^3}{\theta_i^2} + \frac{1}{4!}\frac{2\delta\theta_i^4}{\theta_i^3} - \frac{1}{5!}\frac{6\delta\theta_i^5}{\theta_i^4} + O(\delta_i^6)\right\} \tag{C24}$$

Next, we act with the derivative with respect to the bias:

$$\partial_{\theta_{\mathrm{bias}\,i}} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = (\partial\theta_i/\partial\theta_{\mathrm{bias}\,i})\partial_{\theta_i} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = \kappa\partial_{\theta_i} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\,. \tag{C25}$$

Remembering that $\delta\boldsymbol{\theta}$ is a function of $\boldsymbol{\theta}$, $\partial\delta\theta/\partial\theta = -1$, we obtain:

$$\partial_{\theta_i} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = -\frac{1}{\theta_i}\delta\theta_i + \left[-\frac{1}{2!}\frac{\delta\theta_i^2}{\theta_i^2} + \frac{3}{3!}\frac{\delta\theta_i^2}{\theta_i^2}\right] + \left[\frac{2}{3!}\frac{\delta\theta_i^3}{\theta_i^3} - \frac{4}{4!}\frac{2\delta\theta_i^3}{\theta_i^3}\right] + ... \tag{C26}$$

$$= -\frac{1}{\theta_i}\delta\theta_i\,, \tag{C27}$$

where on in the first line on the right-hand side we have grouped terms such that expressions in the brackets vanish to show how terms of successive order in $\delta\theta$ cancel. Using $\theta_2 = 1 - \theta_1$, we finally see that in matrix notation the expression takes the form

$$\partial_{\boldsymbol{\theta}} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}}) = -\begin{pmatrix} 1/\theta_1 & 0 \\ 0 & 1/(1-\theta_1)\end{pmatrix}\delta\boldsymbol{\theta} = -\mathcal{I}(\boldsymbol{\theta})\delta\boldsymbol{\theta}\,, \tag{C28}$$

where we recognised the 2×2 matrix to be the Fisher information matrix. Thus, we obtain the optimality condition

$$-\left\langle \partial_{\boldsymbol{\theta}} D_{\mathrm{KL}}(P_{\boldsymbol{\theta}^*}||P_{\boldsymbol{\theta}})\right\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = \left\langle \mathcal{I}(\boldsymbol{\theta})\delta\boldsymbol{\theta}\right\rangle_{\hat{\boldsymbol{\theta}}_{\mathrm{ML}}} = 0\,. \tag{C29}$$

Appendix D: Optimal parameter calculation for Poisson distribution with unknown rate parameter

We consider two Poisson distributions with parameters λ^* and λ:

$$P_{\lambda^*}(n) = \frac{\lambda^{*\,n} e^{-\lambda^*}}{n!}\,,\; P_\lambda(n) = \frac{\lambda^n e^{-\lambda}}{n!}\,. \tag{D1}$$

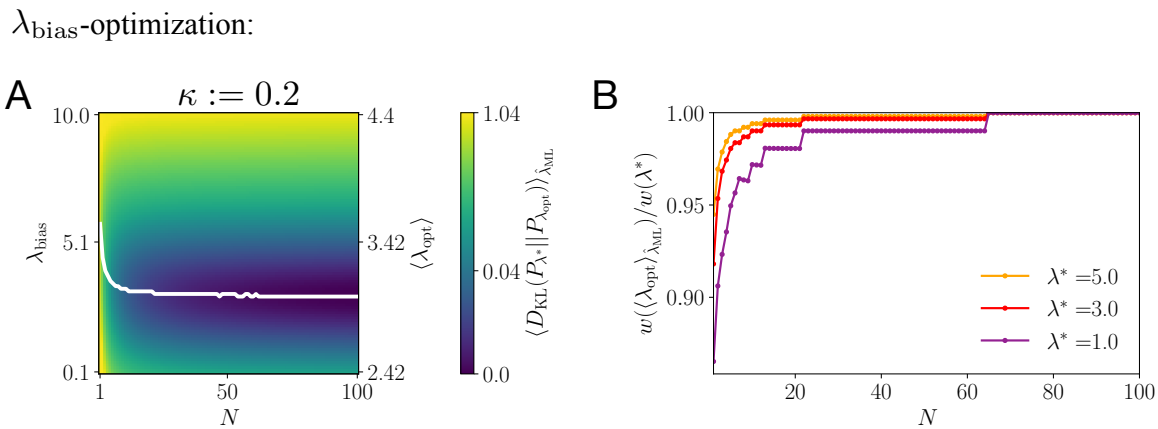

FIG. S4. Optimization of the expected loss for the Poisson distribution with respect to the prior shows a bias to low complexity models. The true model lies at $\lambda^* = 3$. (A) The bias λ_{bias} is optimised while $\kappa = 0.2$ ($q = 0.25$) is kept fixed. The plot shows the landscape of the expected loss as a function of the bias and the number of observations. We are using a non-linear colormap based on the cumulative distribution of the loss. (B) The plot shows for different true models λ^* the ratio of the local complexity measure by Jeffreys prior of the expected optimal to the true model.

The KL divergence between the two distributions is given by

$$D_{\text{KL}}(P_{\lambda^*}||Q_\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^{*\,n} e^{-\lambda^*}}{n!} \log\left(\frac{\lambda^{*\,n} e^{-\lambda^*}}{n!} \frac{n!}{\lambda^n e^{-\lambda}}\right) = \lambda^* \log\frac{\lambda^*}{\lambda} + (\lambda - \lambda^*)\,. \tag{D2}$$

We want to minimise the posterior mean loss

$$\left\langle \mathcal{L}_{\lambda^*} \right\rangle_{\text{post}}(\lambda) = \int \mathrm{d}\tilde{\lambda} P(\tilde{\lambda}|\mathbf{n}) D_{\text{KL}}(P_{\tilde{\lambda}}||Q_\lambda) \tag{D3}$$

with respect to λ. Taking the derivative with respect to λ we obtain the condition

$$\partial_\lambda \left\langle \mathcal{L}_{\lambda^*} \right\rangle_{\text{post}} = \partial_\lambda \int \mathrm{d}\tilde{\lambda} P(\tilde{\lambda}|\mathbf{n}) D_{\text{KL}}(P_{\tilde{\lambda}}||Q_\lambda) = \int \mathrm{d}\tilde{\lambda} P(\tilde{\lambda}|\mathbf{n}) \left(-\tilde{\lambda}/\lambda + 1\right) = 0\,. \tag{D4}$$

The condition is satisfied by

$$\lambda = \int \mathrm{d}\tilde{\lambda} P(\tilde{\lambda}|\mathbf{n}) \tilde{\lambda}\,. \tag{D5}$$

The conjugate prior for the mean number of counts λ is the Gamma distribution and the posterior takes the form

$$P(\lambda|\mathbf{n}) = \text{Gamma}(\alpha', \beta') = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \lambda^{\alpha'-1} e^{-\beta'\lambda} \tag{D6}$$

with

$$\alpha' = \alpha + N\bar{X} = \alpha + N\frac{1}{N}\sum_{i=1}^{N} x_i\,, \tag{D7}$$

$$\beta' = \beta + N\,, \tag{D8}$$

and the hyperparameters $\alpha, \beta > 0$. $\bar{X}$ is an empirical estimate of the rate λ from N observations. The posterior mean which gives the optimal parameter choice is given by

$$\lambda(\hat{\lambda}_{\text{ML}}; \alpha, \beta) = \frac{\alpha'}{\beta'} = \frac{\alpha + N\hat{\lambda}_{\text{ML}}}{\beta + N} = \frac{\frac{\beta}{N}\frac{\alpha}{\beta} + \hat{\lambda}_{\text{ML}}}{\frac{\beta}{N} + 1} = \frac{q\lambda_{\text{bias}} + \hat{\lambda}_{\text{ML}}}{q + 1} \tag{D9}$$

and we define as before

$$N_0 \equiv \beta\,, \quad \lambda_{\rm bias} \equiv \frac{\alpha}{\beta}\,, \quad \text{and} \quad q \equiv \frac{\beta}{N}\,. \tag{D10}$$

The hyperparameter β can thus be interpreted as virtual observations and α as the sum of counts across β observations. If we do not trust in the empirical estimate, then we should take α large compared to β. In this limit, the optimal λ values becomes large and lies further away from the boundary of parameter space situated at $\lambda = 0$.

Up to a scaling by $1/N$, the maximum-likelihood value of the average count $\hat{\lambda}_{\rm ML}$ computed from N observations, is distributed according to the true Poisson distribution with the rate rescaled by N. This can be seen from the characteristic function of the Poisson distribution given by $\exp(\lambda(e^{it}-1))$ and the fact that the characteristic function of the sum of N i.i.d. random variables is given by the N'th power of the characteristic function of a single random variable. The empirically averaged expected loss is then given by

$$\left\langle \mathcal{L}_{\lambda^*}(\lambda) \right\rangle_{\hat{\lambda}_{\rm ML}} = \sum_{n=0}^{\infty} {\rm Poi}(n; N\lambda^*) D_{\rm KL}(P_{\lambda^*} || Q_{\lambda(n/N;\alpha,\beta)})\,. \tag{D11}$$

We minimise this loss with respect to the hyperparameters, for instance by keeping β fixed an minimising with respect to α. In Figure S4 we show the optimal choice of the hyperparameter α as a function of the data size, confirming that for small N, α is large, while for increasing data size $\alpha \to 0$.

Appendix E: Optimal parameter calculation for Gaussian with known mean and unknown variance

Consider two Gaussian distributions with the same known mean μ and unknown variances σ^{*2} and σ^2

$$P_{\sigma^{*2}}(x) = \frac{1}{\sqrt{2\pi\sigma^{*2}}} e^{-\frac{(x-\mu)^2}{2\sigma^{*2}}}\,,\; P_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,. \tag{E1}$$

The KL divergence two between these distributions is given by

$$D_{\rm KL}(P_{\sigma^{*2}} || P_{\sigma^2}) = -\frac{1}{2}\left[\log\sigma^{*2} - \log\sigma^2 + 1 - \frac{\sigma^{*2}}{\sigma^2}\right]\,. \tag{E2}$$

We want to minimise the posterior mean loss

$$\left\langle \mathcal{L}_{\sigma^*} \right\rangle_{\rm post}(\sigma) = \int {\rm d}\tilde{\sigma}^2 P(\tilde{\sigma}^2|\mathbf{n}) D_{\rm KL}(P_{\tilde{\sigma}^2} || P_{\sigma^2}) \tag{E3}$$

with respect to σ^2, where $P(\tilde{\sigma}^2|\mathbf{n})$ is the posterior. Taking the derivative with respect to σ we obtain the condition

$$\partial_\sigma \int {\rm d}\tilde{\sigma}^2 P(\tilde{\sigma}^2|\mathbf{n}) D_{\rm KL}(P_{\tilde{\sigma}^2} || P_{\sigma^2}) = \int {\rm d}\tilde{\sigma}^2 P(\tilde{\sigma}^2|\mathbf{n}) \left(\frac{1}{\sigma} - \frac{\tilde{\sigma}^2}{\sigma^3}\right) = 0 \tag{E4}$$

satisfied by

$$\sigma^2 = \int {\rm d}\tilde{\sigma}^2 \tilde{\sigma}^2 P(\tilde{\sigma}^2|\mathbf{n})\,. \tag{E5}$$

The conjugate prior for the variance is the scaled inverse-chi-squared distribution χ^{-2}. In [56], the posterior distribution for the variance, is given as

$$P(\sigma^2|\nu_N, \sigma_N^2) = \chi^{-2}(\sigma^2|\nu_N, \sigma_N^2) = \frac{1}{\Gamma(\nu_N/2)} \left(\frac{\nu_N \sigma_N^2}{2}\right)^{\nu_N/2} x^{-\nu_N/2-1} e^{-\frac{\nu_N \sigma_N^2}{2x}}\,, \tag{E6}$$

where

$$\nu_N = \nu + N\,, \quad \text{with} \quad \nu > 2\,, \tag{E7}$$

$$\sigma_N^2 = \frac{1}{\nu_N}\left(\nu\sigma_0^2 + \sum_{i=1}^{N}(x_i - \mu)^2\right)\,. \tag{E8}$$

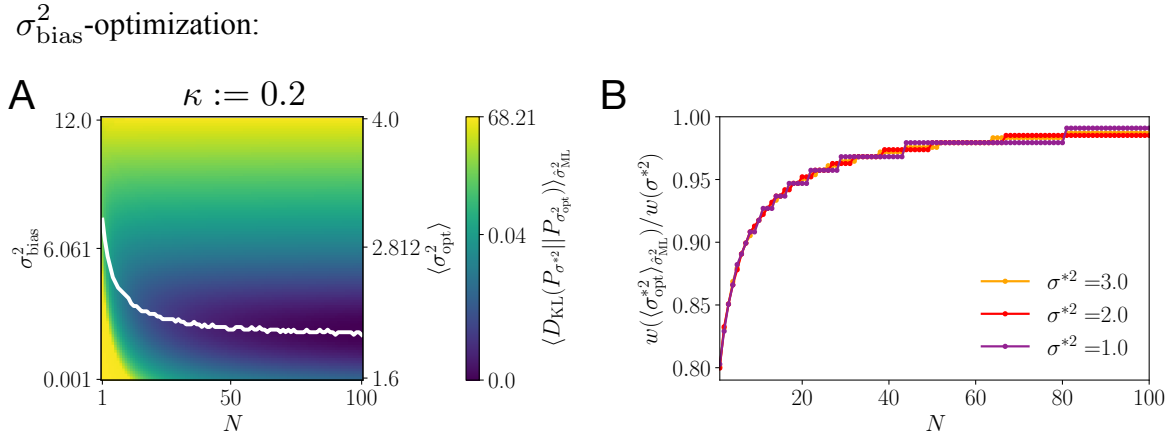

FIG. S5. Optimization of the expected loss for the Gaussian distribution with respect to the prior shows a bias to low complexity models. The true model lies at $\sigma^{*2} = 2$ and we set $\mu = 0$ for the known mean value. (A) The bias σ^2_{bias} is optimised while $\kappa = 0.2$ ($q = 0.25$) is kept fixed. The plot shows the landscape of the expected loss as a function of the bias and the number of observations. We are using a non-linear colormap based on the cumulative distribution of the loss. (B) The plot shows for different true models σ^{*2} the ratio of the local complexity measure by Jeffreys prior of the expected optimal to the true model.

The loss-minimizing value for σ^2 is given by the Bayesian mean which comes out as

$$\sigma^2(\hat{\sigma}^2_{\text{ML}};\nu,\sigma_0^2) = \int \mathrm{d}\tilde{\sigma}^2\tilde{\sigma}^2\chi^{-2}(\tilde{\sigma}^2|\nu_N,\sigma_N^2) = \frac{\nu_N}{\nu_N-2}\sigma_N^2 = \frac{\nu\sigma_0^2+\sum_i(x_i-\mu)^2}{\nu-2+N} = \frac{\frac{\nu-2}{N}\frac{\nu\sigma_0^2}{\nu-2}+\frac{1}{N}\sum_i(x_i-\mu)^2}{\frac{\nu-2}{N}+1} \tag{E9}$$

and we define

$$N_0 \equiv \nu - 2\,, \tag{E10}$$

$$\sigma^2_{\text{bias}} \equiv \frac{\nu\sigma_0^2}{\nu-2}\,, \tag{E11}$$

$$\hat{\sigma}^2_{\text{ML}} \equiv \frac{1}{N}\sum_i (x_i-\mu)^2\,. \tag{E12}$$

Up to a scaling of σ^{*2}/N, the empirical variance is distributed according to the chi-squared distribution of order N (recall, the variance is given by the sum of N i.i.d. squared Gaussian random variables divided by N) given by

$$\chi^2_N(x) = \frac{1}{2^{N/2}\Gamma(N/2)}\,x^{N/2-1}e^{-x/2}\,. \tag{E13}$$

Finally, we optimise the expected loss

$$\langle\mathcal{L}_{\sigma^*}(\sigma)\rangle_{\hat{\sigma}_{\text{ML}}} = \int \mathrm{d}x\,\chi^2_N(x) D_{\text{KL}}(P_{\sigma^{*2}}||P_{\sigma^2(\sigma^{*2}x/N;\nu,\sigma_0^2)}) \tag{E14}$$

with respect to the bias term, while κ is fixed. In this equation we are using the continuous version of the Kullback-Leibler divergence. By Eq. (18) of the main text, κ fixes q which in turn provides a value for N_0 given N:

$$q = \frac{N_0}{N} = \frac{\kappa}{1-\kappa}\,, \quad N_0 = Nq = \nu - 2\,, \tag{E15}$$

where we have also used the definition of N_0 in terms of the hyperparameter ν. For the hyperparameters ν and the bias we can thus write:

$$\nu = N_0 + 2\,, \tag{E16}$$

$$\sigma^2_{\text{bias}} = \frac{N_0+2}{N_0}\sigma_0^2\,. \tag{E17}$$

Minimising the expected loss given N, yields the optimal choice for σ^2_{bias}. In Figure S5 we show the optimal bias value as a function of the data size.

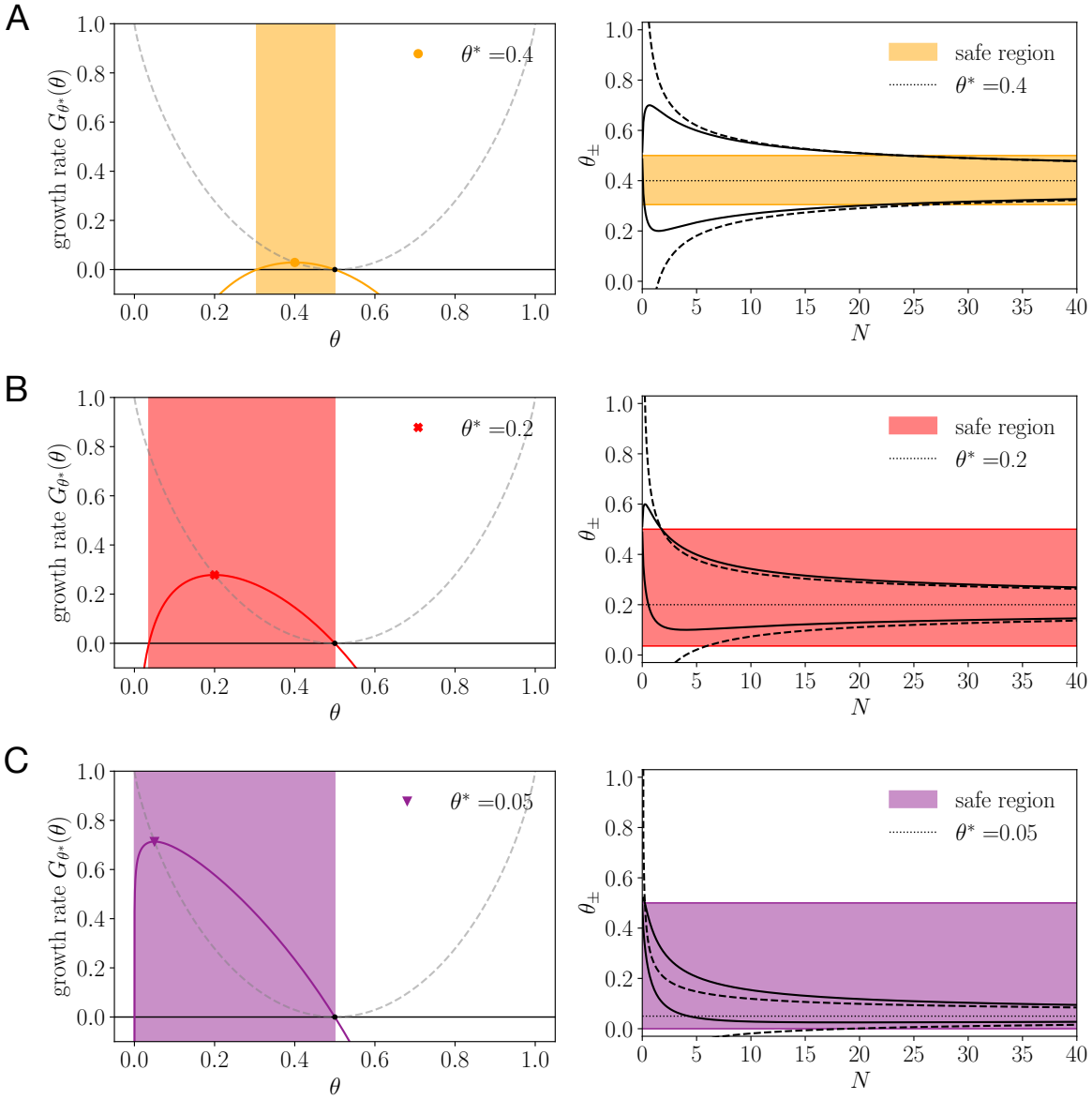

FIG. S6. The prior choice in Bayesian inference controls the safety of the learning strategy. Left column: long-term growth rate as a function of the adopted model ϕ for three different values of the true model $\theta^* = 0.4$, 0.2 and 0.05. Shaded boxes indicate the regions of non-negative growth rate. Right column: Magnitude of fluctuations of the conditional optimal model as a function of the number of observations N for two different learning strategies: maximum-likelihood estimation ($\alpha = 0$; dashed line) and Jeffreys prior ($\alpha = 1/2$; solid line).

Appendix F: Statistical fluctuations of different learning strategies

In Figure S6, we show the magnitude of statistical fluctuations of the inferred model in relation to the region of non-negative long-term growth rates. Subfigures A, B and C show different true models $\theta^* = 0.4$, 0.2 and 0.05, respectively. In the left column of subfigures we show the growth rates $G_{\theta^*}(\theta)$ as a function of the model choice θ. Regions where the growth rate is non-negative are indicated by color shaded boxes. In the right column of subfigures we show the statistical fluctuations of the inferred model as function of N for two different learning strategies. The first learning strategy is maximum-likelihood estimation $\alpha = 0$ and the second strategy uses $\alpha = 1/2$. The later strategy ($\alpha = 1/2$) is safer, because the magnitude of fluctuations is shrunk and more contained in the interval of non-negative growth rates in comparison to the maximum-likelihood strategy.

[1] F. M. Burnet, A modification of Jerne's theory of antibody production using the concept of clonal selection, CA: A Cancer Journal for Clinicians **26**, 119 (1976).

[2] E. Kussell and S. Leibler, Phenotypic diversity, population growth, and information in fluctuating environments, Science **309**, 2075 (2005).

[3] B. Xue and S. Leibler, Benefits of phenotypic plasticity for population growth in varying environments, Proceedings of the National Academy of Sciences **115**, 12745 (2018).

[4] J. J. Hopfield, Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity, Proceedings of the National Academy of Sciences **71**, 4135 (1974).

[5] I. Martincorena, A. S. Seshasayee, and N. M. Luscombe, Evidence of non-random mutation rates suggests an evolutionary risk management strategy, Nature **485**, 95 (2012).

[6] L. D. Harder and L. A. Real, Why are bumble bees risk averse?, Ecology **68**, 1104 (1987).

[7] U. Bhat, C. P. Kempes, and J. D. Yeakel, Scaling the risk landscape drives optimal life-history strategies and the evolution of grazing, Proceedings of the National Academy of Sciences **117**, 1580 (2020).

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).

[9] M. Vergassola, E. Villermaux, and B. I. Shraiman, 'infotaxis' as a strategy for searching without gradients, Nature **445**, 406 (2007).

[10] D. Cohen, Optimizing reproduction in a randomly varying environment, Journal of Theoretical Biology **12**, 119 (1966).

[11] D. R. Rubenstein, Spatiotemporal environmental variation, risk aversion, and the evolution of cooperative breeding as a bet-hedging strategy, Proceedings of the National Academy of Sciences **108**, 10816 (2011).

[12] R. Levins, Theory of fitness in a heterogeneous environment. i. the fitness set and adaptive function, The American Naturalist **96**, 361 (1962).

[13] I. Nemenman, Information theory and adaptation, Quantitative biology: from molecular to cellular systems **4**, 73 (2012).

[14] T. J. Perkins and P. S. Swain, Strategies for cellular decision-making, Molecular systems biology **5**, 326 (2009).

[15] P. S. Laplace, marquis de, *Essai philosophique sur les probabilités* (Courcier, 1814).

[16] J. M. Keynes, *A Treatise on Probability* (Dover Publications, 1921).

[17] S. Kullback, *Information theory and statistics* (Courier Corporation, 1997).

[18] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst (Cambridge University Press, 2003).

[19] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[20] E. T. Jaynes, Information theory and statistical mechanics. ii, Phys. Rev. **108**, 171 (1957).

[21] J. Rissanen, Modeling by shortest data description, Automatica **14**, 465 (1978).

[22] V. Balasubramanian, Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions, Neural Comput. **9**, 349–368 (1997).

[23] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **186**, 453 (1946).

[24] J. L. Kelly Jr., A new interpretation of information rate, Bell System Technical Journal **35**, 917 (1956).

[25] O. Rivoire and S. Leibler, The value of information for populations in varying environments, Journal of Statistical Physics **142**, 1124 (2011).

[26] R. A. Fisher and E. J. Russell, On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **222**, 309 (1922).

[27] S.-I. Amari, A foundation of information geometry, Electronics and Communications in Japan (Part I: Communications) **66**, 1 (1983).

[28] F. Nielsen and V. Garcia, Statistical exponential families: A digest with flash cards (2011), arXiv:0911.4863 [cs.LG].

[29] M. J. Wainwright, M. I. Jordan, *et al.*, Graphical models, exponential families, and variational inference, Foundations and Trends in Machine Learning **1**, 1 (2008).

[30] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR computational mathematics and mathematical physics **7**, 200 (1967).

[31] F. Nielsen, An elementary introduction to information geometry, Entropy **22**, 1100 (2020).

[32] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty, Clustering with Bregman Divergences., Journal of machine learning research **6** (2005).

[33] P. Diaconis and D. Ylvisaker, Conjugate priors for exponential families, The Annals of Statistics **7**, 269 (1979).

[34] G. J. Lidstone, Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities, Transactions of the Faculty of Actuaries **8**, 13 (1920).

[35] I. Nemenman, F. Shafee, and W. Bialek, Entropy and inference, revisited, in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01 (MIT Press, Cambridge, MA, USA, 2001) p. 471–478.

[36] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, How a well-adapted immune system is organized, Proceedings of the National Academy of Sciences **112**, 5950 (2015).

[37] J. R. Gremer and D. L. Venable, Bet hedging in desert winter annual plants: optimal germination strategies in a variable environment., Ecology letters **17 3**, 380 (2014).

[38] N. Balaban, J. Merrin, R. Chait, L. Kowalik, and S. Leibler, Bacterial persistence as a phenotypic switch, Science **305**, 1622 (2004).

[39] H. S. Moyed and K. P. Bertrand, hipa, a newly recognized gene of escherichia coli k-12 that affects frequency of persistence after inhibition of murein synthesis., Journal of Bacteriology **155**, 768 (1983).

[40] H. Mattingly, K. Kamino, B. Machta, and T. Emonet, E. coli chemotaxis is information-limited, bioRxiv 10.1101/2021.02.22.432091 (2021).

[41] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, Maximizing the information learned from finite data selects a simple model, Proceedings of the National Academy of Sciences **115**, 1760 (2018).

[42] K. N. Quinn, M. C. Abbott, M. K. Transtrum, B. B. Machta, and J. P. Sethna, Information geometry for multiparameter models: New perspectives on the origin of simplicity (2021), arXiv:2111.07176 [cond-mat.stat-mech].

[43] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models, PLoS computational biology **3**, e189 (2007).

[44] O. H. Schnaack, L. Peliti, and A. Nourmohammad, Risk-utility tradeoff shapes memory strategies for evolving patterns, arXiv preprint arXiv:2110.15008 (2021).

[45] E. D. Lee, J. C. Flack, and D. C. Krakauer, Outsourcing memory through niche construction, bioRxiv , 2022 (2022).

[46] J. Feldman, The simplicity principle in perception and cognition, Wiley Interdisciplinary Reviews: Cognitive Science **7**, 330 (2016).

[47] K. Koffka, *Principles of Gestalt psychology*, Vol. 44 (routledge, 2013).

[48] N. Chater and P. Vitányi, Simplicity: a unifying principle in cognitive science?, Trends in cognitive sciences **7**, 19 (2003).

[49] S. J. Gershman and Y. Niv, Perceptual estimation obeys occam's razor, Frontiers in psychology **4**, 623 (2013).

[50] G. Tavoni, T. Doi, C. Pizzica, V. Balasubramanian, and J. I. Gold, Human inference reflects a normative balance of complexity and accuracy, Nature Human Behaviour **6**, 1153 (2022).

[51] E. Piasini, S. Liu, P. Chaudhari, V. Balasubramanian, and J. I. Gold, How occam's razor guides human decision-making, bioRxiv , 2023 (2023).

[52] J. J. Rissanen, Fisher information and stochastic complexity, IEEE transactions on information theory **42**, 40 (1996).

[53] P. D. Grünwald, *The minimum description length principle* (MIT press, 2007).

[54] W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity, and learning, Neural computation **13**, 2409 (2001).

[55] V. Sachdeva, T. Mora, A. M. Walczak, and S. E. Palmer, Optimal prediction with resource constraints using the information bottleneck, PLOS Computational Biology **17**, e1008743 (2021).

[56] K. P. Murphy, Conjugate bayesian analysis of the gaussian distribution, `https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf` (2007), (version: 2023-04-13).