Journal of Machine Learning Research 23 (2022) 1-36 Submitted 9/21; Revised 3/22; Published 5/22

Testing Whether a Learning Procedure is Calibrated

Jon Cockayne JON.COCKAYNE@SOTON.AC.UK
Mathematical Sciences
University of Southampton
Highfield
Southampton, SO17 1BJ, UK

Matthew M. Graham M.GRAHAM@UCL.AC.UK
Centre for Advanced Research Computing
University College London
Gower Street
London, WC1E 6BT, UK

Chris J. Oates CHRIS.OATES@NCL.AC.UK
School of Mathematics, Statistics & Physics
Newcastle University
Newcastle Upon Tyne, NE1 7RU, UK

T. J. Sullivan T.J.SULLIVAN@WARWICK.AC.UK
Mathematics Institute and School of Engineering
University of Warwick
Coventry, CV4 7AL, UK

Onur Teymur O@TEYMUR.UK
School of Mathematics, Statistics & Actuarial Science
University of Kent
Cantebury, CT2 7NZ, UK

Editor: Anthony Lee

Abstract

A learning procedure takes as input a dataset and performs inference for the parameters θ of a model that is assumed to have given rise to the dataset. Here we consider learning procedures whose output is a probability distribution, representing uncertainty about θ after seeing the dataset. Bayesian inference is a prime example of such a procedure, but one can also construct other learning procedures that return distributional output. This paper studies conditions for a learning procedure to be considered *calibrated*, in the sense that the true data-generating parameters are plausible as samples from its distributional output. A learning procedure whose inferences and predictions are systematically over- or under-confident will fail to be calibrated. On the other hand, a learning procedure that *is* calibrated need not be statistically efficient. A hypothesis-testing framework is developed in order to assess, using simulation, whether a learning procedure is calibrated. Several vignettes are presented to illustrate different aspects of the framework.

Keywords: calibratedness, credible sets, uncertainty quantification

1. Introduction

Given a parametric model and a dataset purported to be generated from the model, the modern workflow for parameter inference first identifies a statistical paradigm (e.g. Bayesian inference), performs any required numerical computation using an appropriate numerical method, then inspects the results and refines the approach until some *desiderata* (e.g. posterior predictive checks, or a convergence diagnostic for a Markov chain Monte Carlo method) are satisfied. This paper takes a holistic perspective and refers to the overall workflow as a *learning procedure*. Our focus is on learning procedures that produce distributional output, examples of which include workflows based on Bayesian and generalised Bayesian inference (Bissiri et al., 2016), fractional posteriors (Bhattacharya et al., 2019), empirical Bayes (Casella, 1985), variational Bayes (Blei et al., 2017), approximate Bayesian computation (Beaumont et al., 2002), Bayesian synthetic likelihood (Price et al., 2018), and also approaches that have a non-Bayesian motivation, such as the maximum entropy approach (Jaynes, 1982).

It is natural to hope that a learning procedure is *calibrated*, in the sense that the true data-generating parameters are plausible as samples from the distributional output. Indeed, a learning procedure that is *not* calibrated can produce inferences and predictions that are either biased or over/under-confident, and lead users to draw spurious conclusions in model selection problems. The consequences of over-confidence, in particular, could be dire when those inferences are used in safety-critical applications. This point has been discussed at length in the literature, such as in investigating frequentist coverage of credible sets in Bayesian inference and in calibrating probabilistic forecasts. However, the literature appears to lack a definition of "calibration" that is sufficiently general to be applied to an arbitrary learning procedure that produces distributional output. The aim of this paper is to introduce a general definition of "calibration" and accompany this with a methodology for testing whether a learning procedure is calibrated.

The term *calibration* is unfortunately overloaded in the statistical literature. It is also used to refer to the parameter inference task in applications that involve a computer model. For example, Kennedy and O'Hagan (2001) write that "*the process of fitting the model to the observed data by adjusting the parameters is known as calibration*". For avoidance of doubt, we use the standard terminology of *parameter inference* to refer to the task of estimating parameters of a model. The term 'calibration' is also used in the literature on forecast assessment. There the useage is close to the notions proposed in this paper, though in that literature the focus is on testing calibration at the level of the *data* rather than at the level of the parameters. This is discussed further in Sections 2.3.2 and 2.4.2. We reserve the term *calibration* for the specific notions proposed in this paper.

The outline of the paper is as follows: Section 2 presents our proposed definitions, where we identify both *strong* and *weak* senses in which a learning procedure can be said to be calibrated. To ensure our definitions are precise in a mathematical sense, we conceptualise a learning procedure as a mathematical object in Section 2.1 and impose mild regularity assumptions on this object in Section 2.2. In Section 2.3 our notion of strong calibration is presented, illustrated by examples in Section 2.3.1, and compared to existing definitions in the literature in Section 2.3.2. Likewise, in Section 2.4 our notion of weak calibration is presented, illustrated by examples in Section 2.4.1, and compared to existing definitions in

the literature in Section 2.4.2. Several vignettes are provided in Section 3, showing through simulations that our proposed definitions of calibration both accord with intuition and can be tested for. A brief discussion concludes the paper in Section 4.

1.1 Notation

For a measurable space S, $\mathcal{P}(S)$ will denote the set of probability measures on S. For $s \in S$ let $\delta(S) \in \mathcal{P}(s)$ denote the Dirac distribution on s. For a measurable function $f\colon S \to \mathbb{R}$, a measurable set $A \subseteq \mathbb{R}$, and a probability measure $\nu \in \mathcal{P}(S)$, let $f^{-1}(A) \coloneqq \{x \in S \mid f(x) \in A\}$ denote the *preimage* of A and recall that the *pushforward* measure $f_{\#}\nu \in \mathcal{P}(\mathbb{R})$ is defined as $(f_{\#}\nu)(A) \coloneqq \nu(f^{-1}(A))$.

2. What it Means for a Learning Procedure to be Calibrated

This section sets out our proposed definitions of strong and weak calibration, provides examples of learning procedures that are strongly and weakly calibrated, and relates our definitions to existing work.

2.1 Set-Up

Let Θ be a measurable space, which will play the role of the *parameter space* in this work. It is assumed that there is a unique "true" parameter $\theta \in \Theta$ and we consider the *parameter inference* task of estimating θ based on a dataset. Let Y be a measurable space in which datasets are realised.

Definition 1 (Learning Procedure)**.** *A* learning procedure *is a function*

$$\begin{aligned} \mu\colon \mathcal{P}(\Theta) \times Y &\to \mathcal{P}(\Theta) \\ (\mu_0, y) &\mapsto \mu(\mu_0, y). \end{aligned}$$

Here μ_0 is interpreted as an initial belief distribution*, quantifying uncertainty about the parameter θ before any data have been observed, and y denotes a dataset. The distributional output $\mu(\mu_0, y)$ is interpreted as a quantification of the uncertainty associated with the parameter θ, after the data y have been observed.*

The standard example of a learning procedure is Bayesian inference, wherein μ_0 is the prior distribution and $\mu(\mu_0, y)$ is the posterior distribution, this being determined by the prior, the observed data y, and a likelihood function that must be specified. However, Definition 1 is general enough to accommodate any workflow that produces distributional output. In particular, Definition 1 does not pre-suppose that a data-generating model exists or is known to the user, so that the definition of a learning procedure may be applied even in the *M-open* setting (Bernardo and Smith, 1994, §6.1.2). Further, one may consider that computational procedures such as variational inference or Monte Carlo form part of the learning procedure, and in this sense a myriad of different learning procedures can be considered.

Note that we call μ_0 a *belief distribution* following Bissiri et al. (2016) and reserve the term *prior* for use only in the Bayesian context. We also emphasise that a learning

procedure need not depend upon the initial belief distribution μ_0; for example, in the maximum entropy approach (Jaynes, 1982) a distributional output is produced that does not explicitly depend on any initial belief, so that effectively $\mu(\mu_0, y) \equiv \mu(y)$.

In the next section we will introduce the mathematical facts required for our notions of strong and weak calibration in Sections 2.3 and 2.4.

2.2 A Mathematical Characterisation

The definitions that we will present rely on *cumulative distribution functions* (CDFs) and their inverses, and we therefore impose regularity conditions to ensure that such inverse CDFs are well-defined. That is, we impose sufficient regularity to restrict our attention in the sequel to inverse CDFs that are well-defined functions, as opposed to dealing with generalised functions that are set-valued.

Definition 2 (Regular Distribution)**.** *Let Θ be a measurable space equipped with a reference measure λ. A distribution $\nu \in \mathcal{P}(\Theta)$ is* regular *(with respect to λ) if it admits a probability density function (*PDF*) $p_\nu := \mathrm{d}\nu/\mathrm{d}\lambda$ such that $p_\nu > 0$ on Θ (i.e. the measures ν and λ are* equivalent*). The set of all regular distributions will be denoted $\mathcal{P}_r(\Theta)$.*

When Θ is a Borel- or Lebesgue-measurable subset of Euclidean space, the reference measure λ will be assumed to be Lebesgue measure. For $-\infty \leq a < b \leq \infty$ and a univariate distribution $\gamma \in \mathcal{P}((a,b))$, we let $F_\gamma \colon (a,b) \to [0,1]$ denote the associated CDF $F_\gamma(x) := \gamma((a,x]) = \int_a^x \mathrm{d}\gamma$. Our first result, Lemma 3, is classical (e.g. Rosenblatt, 1952) and underpins methods for simulation of univariate random variables using inverse CDFs. This result establishes that the level of regularity in Definition 2 is sufficient for the inverse CDF approach to simulation of such distributions to be applied. It also ensures that our subsequent constructions that depend on Definition 2 are well-defined.

Lemma 3. *For $-\infty \leq a < b \leq \infty$ and $\nu \in \mathcal{P}_r((a,b))$ we have that $F_\nu(X) \sim \mathcal{U}(0,1)$ whenever $X \sim \nu$.*

Proof Since ν admits a PDF p_ν on (a,b), the fundamental theorem of calculus implies that F_ν is differentiable with $DF_\nu(\theta) = p_\nu(\theta)$. In particular, since $p_\nu > 0$ we have that F_ν is continuous and strictly increasing and therefore the sets $F_\nu^{-1}(z)$ are singletons for all $z \in [0,1]$. Let $X \sim \nu$ and $Z := F_\nu(X)$. Then, from the change of variables formula, Z admits a PDF $q(z)$ on $[0,1]$ with

$$q(z) = \sum_{\theta : F_\nu(\theta) = z} p_\nu(\theta) |DF_\nu(\theta)|^{-1} = p_\nu(\theta) \frac{1}{p_\nu(\theta)} = 1,$$

which is indeed the PDF of $\mathcal{U}(0,1)$. ■

The random variable $F_\nu(X)$ is sometimes called the *probability integral transform*; see e.g. Dawid (1984); Diebold et al. (1997). When $\Theta \not\subset \mathbb{R}$, the CDF of a distribution $\nu \in \mathcal{P}_r(\Theta)$ is not in general well-defined. To characterise such distributions analogously to the above, consider a set of test functions of the form $f \colon \Theta \to (a,b)$, with the property that each univariate marginal $f_\# \nu$ *does* admit an invertible CDF. We next establish that regular distributions are characterised by a certain (large) set of such statistics.

Definition 4 (Test Functions $\mathcal{F}_\Theta$)**.** *Consider measurable functions of the form $f\colon \Theta \to (a,b)$ for some $-\infty \le a < b \le \infty$. Then the* test functions $\mathcal{F}_\Theta$ *are the set of all such f for which $f_\# \nu \in \mathcal{P}_r((a,b))$ whenever $\nu \in \mathcal{P}_r(\Theta)$.*

Intuitively, $\mathcal{F}_\Theta$ rules out functions f that take a constant value on a non-null set, in order to avoid the situation where $f_\#\nu$ contains an atom and the CDF $F_{f_\#\nu}\colon (a,b) \to [0,1]$ is not invertible. In the univariate case $\Theta = \mathbb{R}$, the set $\mathcal{F}_\Theta$ contains functions f for which the gradient exists and is nonzero almost everywhere and, moreover, the preimages $f^{-1}(z)$ have cardinality n such that $0 < n < \infty$ for each $z \in (a,b)$. Indeed, in this case $f_\#\nu$ admits an everywhere positive (Lebesgue) PDF on (a,b) of the form

$$p_{f_\#\nu}(z) = \sum_{\theta \in f^{-1}(z)} p_\nu(\theta)\, |Df(\theta)|^{-1}\,. \tag{1}$$

Since by assumption $f_\#\nu$ is regular on (a,b), from Lemma 3 we have that $F_{f_\#\nu}(f(\theta)) \sim \mathcal{U}(0,1)$ whenever $X \sim \nu$. For the multivariate case $\Theta = \mathbb{R}^d$, by the co-area formula the (Lebesgue) PDF of $f_\#\nu$ is

$$p_{f_\#\nu}(z) = \int_{f^{-1}(z)} p_\nu(\theta) |\det(Df(\theta)Df(\theta)^\mathsf{T})|^{-\frac{1}{2}}\, \mathcal{H}^{d-1}(\mathrm{d}\theta), \tag{2}$$

where $\mathcal{H}^{d-1}$ indicates the $(d-1)$ dimensional Hausdorff measure on Θ (Diaconis et al., 2013, Proposition 2). In this case, the requirement on the Jacobian determinant is that $\det(DfDf^\mathsf{T}) \neq 0$ almost everywhere. As $\mathcal{H}^0$ is equivalent to the counting measure, (2) collapses back to (1) when $d = 1$.

The restriction of attention to $\mathcal{F}_\Theta$ is essentially without loss of generality, as evidenced by the following result, whose proof is contained in Section A.1:

Lemma 5 (Regular Distributions are Characterised by $\mathcal{F}_\Theta$)**.** *Let $\Theta = \mathbb{R}^d$ for some $d \in \mathbb{N}$. Suppose that $\mu, \nu \in \mathcal{P}_r(\Theta)$ and $\int f \,\mathrm{d}\mu = \int f \,\mathrm{d}\nu$ for all $f \in \mathcal{F}_\Theta$. Then $\mu = \nu$.*

Now we have the mathematical tools to define what it means for a learning procedure to be calibrated. In Section 2.3 we introduce a strong notion of calibration, which clarifies the sense in which the true parameter can be considered plausible as a sample from the distributional output. Then, in Section 2.4, we consider a strictly weaker notion of calibration that is more easily tested.

2.3 Strongly Calibrated Learning Procedures

To assess whether a learning procedure is calibrated we must specify what it is calibrated *against*, and this requires a data-generating model. Thus, the assessment framework we present exists in the *M-complete* setting (Bernardo and Smith, 1994, §6.1.2).

Definition 6 (Data-Generating Model)**.** *A* data-generating model *is a function*

$$\begin{aligned} P\colon \Theta &\to \mathcal{P}(Y) \\ \theta &\mapsto P_\theta, \end{aligned}$$

where P_θ carries the interpretation of a statistical model from which data are generated.

In this section we present a strong notion of what it means for a learning procedure to be calibrated to a data-generating model. It simplifies matters to restrict to learning procedures that produce regular distributional output:

Definition 7 (Regular Learning Procedure)**.** *A learning procedure* $\mu\colon \mathcal{P}(\Theta) \times Y \to \mathcal{P}(\Theta)$ *is* regular *if* $\mu(\mu_0, y) \in \mathcal{P}_r(\Theta)$ *for all* $\mu_0 \in \mathcal{P}_r(\Theta)$ *and all* $y \in Y$*.*

Definition 8 (Strongly Calibrated)**.** *Let* $B \subseteq \mathcal{P}_r(\Theta)$ *denote a set of belief distributions and* P *a data-generating model. A regular learning procedure* μ *is said to be* strongly calibrated *to* (B, P) *if*

$$\left.\begin{array}{rl} \theta & \sim \mu_0 \\ y \mid \theta & \sim P_\theta \end{array}\right\} \implies F_{f_{\#}\mu(\mu_0, y)}(f(\theta)) \sim \mathcal{U}(0,1) \tag{3}$$

for all $f \in \mathcal{F}_\Theta$ *and for all* $\mu_0 \in B$*. If the set* B *contains a single element,* μ_0*, then we say simply that* μ *is strongly calibrated to* (μ_0, P)*.*

The assumption that both the belief distribution and learning procedure are regular excludes some important learning procedures. For example, in Bayesian inference one sometimes uses an improper, "uninformative" prior such as $p(\theta) \propto 1$, which would not be regular unless Θ_0 is bounded. To study such a learning procedure in the framework of Definition 8 one could consider constructing an "artificial" learning procedure that took a regular distribution μ_0 as input, but ignored this for the purposes of inference and instead used an improper prior—though, one would still need to ensure that the learning procedure itself returned a regular output, which is not guaranteed for an improper prior. In addition to this, any application of Bayesian inference for which the support of the posterior is a strict subset of Θ (e.g. procedures with truncated likelihoods) will fail to be regular. The distributional output of *approximate Bayesian computation* (ABC) may not be regular for similar reasons. This motivates the introduction of *weakly calibrated* learning procedures in Section 2.4, for which the regularity assumption can be relaxed.

To gain intuition for Definition 8, notice that the unknown data-generating parameter θ is statistically identical to a sample from the distributional output $\mu(\mu_0, y)$ when the learning procedure is strongly calibrated. This intuition is clarified in the following remark:

Remark 9 (Correct Coverage for Credible Sets)**.** *Suppose that the learning procedure* μ *is strongly calibrated to* (μ_0, P)*. If the distribution* $\mu(\mu_0, y)$ *is used to construct a* $1-\alpha$ *probability credible set for* θ*, then this interval will indeed contain* θ *with probability* $1-\alpha$ *under the hierarchical data-generating model* $\theta \sim \mu_0$*,* $y \mid \theta \sim P_\theta$*.*

Thus, the distributional output from a strongly calibrated learning procedure can be meaningfully related to the parameter inference task. Note, however, that even a small degree of misspecification can lead to failure of calibration. Thus strong calibration captures the absence of systematic errors, similar to the notion of an unbiased estimator.

Next we present an actionable test for the hypothesis that a learning procedure is strongly calibrated. We emphasise that this test can in theory be applied to *any* learning procedure (i.e. any workflow used for parameter inference that returns distributional output), providing that the regularity requirements are satisfied and that one is able to simulate from the data-generating model.

Remark 10 (Testing whether a Learning Procedure is Strongly Calibrated)**.** *Fix* $\mu_0 \in \mathcal{P}_r(\Theta)$ *and let*

$$\theta_i \overset{\text{IID}}{\sim} \mu_0$$
$$y_i \mid \theta_i \overset{\text{IID}}{\sim} P_{\theta_i}$$

Then we can test whether a (regular) learning procedure μ *is strongly calibrated to* (μ_0, P) *by picking a test function* $f \in \mathcal{F}_\Theta$ *and using any goodness-of-fit test for the hypothesis*

$$F_{f_{\#}\mu(\mu_0, y_i)}(f(\theta_i)) \overset{\text{IID}}{\sim} \mathcal{U}(0,1).$$

Such a test will not have power against all alternatives unless, for example, $d = 1$ *and* $f(\theta) = \theta$*. To increase the power of the test in higher dimensions, multiple* f *should be simultaneously considered. Methodology for selecting a suitable test function is proposed in Section 3.4.*

Remark 11. *For simplicity we have assumed that each* θ_i *is associated with exactly one* y_i*. In practice this need not be the case; each parameter could be associated with many pieces of data. For example in some applications a sample from* μ_0 *may be more difficult to obtain than repeated measurements* $y_1, \ldots, y_n \sim P_{\theta_i}$*. However we note that this will violate the independence assumption in Remark 10, and would require a more complicated test to be used.*

Remark 12 (Quantification of Strong Calibration)**.** *The departure from uniformity of the law of* $F_{f_{\#}\mu(\mu_0, y)}(f(\theta))$ *under* $\theta \sim \mu_0$*,* $y \mid \theta \sim P_\theta$ *can be used to assess the nature and extent to which the learning procedure fails to be strongly calibrated. Histograms can provide an intuitive visualisation; see Section 3.3.*

In the next section we illustrate Definition 8 with some examples for which strong calibration can be verified. Then, in Section 2.3.2 we discuss the relationship between Definition 8 and earlier work.

2.3.1 Examples of Strongly Calibrated Learning Procedures

Our first example confirms the intuition that the Bayesian framework is strongly calibrated to the prior and the data-generating model.

Example 1 (Bayes is Strongly Calibrated)**.** *If* $\theta \sim \mu_0$ *and* $y \mid \theta \sim P_\theta$ *then* (θ, y) *can be considered to be a sample from the joint distribution of the parameters and dataset. In the Bayesian framework (with the data-generating model* P *correctly specified),* $\mu(\mu_0, y)$ *is defined as the conditional distribution of the parameters given the data, and thus* $\theta \mid y \sim \mu(\mu_0, y)$*. Thus if* μ_0 *and* μ *are regular, it follows from Lemma 3 that* $F_{f_{\#}\mu(\mu_0, y)}(f(\theta)) \sim \mathcal{U}(0,1)$ *for all* $f \in \mathcal{F}_\Theta$*. Thus Bayesian inference is strongly calibrated to* $(\mathcal{P}_r(\Theta), P)$*.*

The following example[1] shows that strongly calibrated learning procedures do not necessarily yield accurate estimators:

1. This example is similar in spirit to the *climatological forecaster* in Example 2 of Gneiting et al. (2007), who uses only historical frequencies to predict tomorrow's weather, agnostic of any recent data that may have been obtained.

Example 2 (Data-Agnostic Learning Procedure is Strongly Calibrated)**.** *The trivial learning procedure that takes $\mu(\mu_0, y) \coloneqq \mu_0$ is strongly calibrated to $(\mathcal{P}_r(\Theta), P)$, since for $\theta \sim \mu_0$ and $\mu_0 \in \mathcal{P}_r(\Theta)$,*

$$F_{f_{\#}\mu(\mu_0,y)}(f(\theta)) = F_{f_{\#}\mu_0}(f(\theta)) \sim \mathcal{U}(0,1)$$

for all $f \in \mathcal{F}_\Theta$.

The implication of Example 2 is that strong calibration alone is not sufficient to justify the practical application of a learning procedure, and additional *desiderata*, such as *statistical efficiency*, will typically also need to be taken into account.[2] This paper focusses on calibration and does not attempt to discuss other *desiderata* and how they should be balanced in the applied context.

One can consider situations between the two extremes of Example 1 and Example 2:

Example 3 (Partial Posteriors are Strongly Calibrated)**.** *A* partial posterior *corresponds to performing full Bayesian inference using only summary statistics $s \colon Y \to S$ of the dataset. These have recently been proposed as a tool for compensating for model misspecification (Lewis et al., 2021). For the partial posterior learning procedure, $\mu(\mu_0, y)$ is the conditional distribution of the parameters given the summarised data $s(y)$ and $\theta \mid s(y) \sim \mu(\mu_0, y)$. When both the prior and the partial posterior learning procedure are regular, it follows from Lemma 3 that $F_{f_{\#}\mu(\mu_0,y)}(f(\theta)) \sim \mathcal{U}(0,1)$ for all $f \in \mathcal{F}_\Theta$. Thus partial posteriors are strongly calibrated to $(\mathcal{P}_r(\Theta), P)$.*

Next we present an example that is a clear departure from the Bayesian framework, in that it clearly does not return a posterior distribution and yet is provably strongly calibrated:

Example 4 (Probabilistic Stationary Iterative Methods are Strongly Calibrated)**.** *Let $\Theta = \mathbb{R}^d$ and consider the data-generating model $P_\theta = \delta(A\theta)$ that returns a Dirac distribution on $y = A\theta$, where A is a non-singular matrix. An ideal learning procedure would return $\mu(\mu_0, y) = \delta(A^{-1}y) = \delta(\theta)$, but in many practical scenarios the exact action of A^{-1} on y cannot be computed, either due to poor conditioning of the matrix A or due to the $O(d^3)$ computational cost associated with inverting A. This motivates the use of an alternative procedure, called a* probabilistic iterative method*, recently proposed in Cockayne et al. (2021) and based on classical iterative methods for solving linear systems (see e.g. Saad, 2003). To describe the procedure, let $R^y : \Theta \to \Theta$ be an map, constructed using y, such that θ is a solution of the fixed point equation $\theta = R^y(\theta)$. For example, the choice $R^y(\theta) = (I - \epsilon A)\theta + \epsilon y$, $\epsilon > 0$, corresponds to a classical iterative method called Richardson's method. Consider then the learning procedure $\mu(\mu_0, y) \coloneqq R^y_{\#}\mu_0$, whose output is conjugate under a Gaussian input μ_0, being an affine transform, and can be exactly computed at cost $O(d^2)$. Cockayne et al. (2021) proved that, under mild conditions, the iterative application of R^y produces a sequence of distributions on Θ that contract to $\delta(\theta)$, and that this procedure is strongly calibrated to $(G(\mathbb{R}^d), P)$, where G is the set of all Gaussian distributions*

2. For example, "*maximizing the sharpness of the predictive distributions subject to calibration*" was proposed in Gneiting et al. (2007), although their use of the term "calibration" is distinct from the present paper, being focussed on forecast assessment. See Section 2.3.2 for further discussion of the literature on forecast assessment.

supported on $\mathbb{R}^d$. This example speaks to one potential use of Definition 8, in providing theoretical justification for non-traditional learning procedures which nevertheless produce meaningful distributional output.

Next our attention turns to the relationship between Definition 8 and existing concepts in the literature.

2.3.2 Relation to Existing Concepts

Here we compare and contrast our notion of strong calibration with concepts appearing in earlier work and in related fields.

Frequentist Coverage: There is a rich literature that aims to assess learning procedures according to frequentist *desiderata*. In particular, one can ask whether credible sets have *correct frequentist coverage*, which is analogous to fixing $\theta = \theta_0$ and asking if $y \sim P_{\theta_0}$ implies $F_{f_{\#}\mu(\mu_0,y)}(f(\theta_0)) \sim \mathcal{U}(0,1)$; i.e. the only randomness is introduced during generation of the dataset. This differs to our notion of strong calibration in that we sample θ from μ_0 while, in the frequentist assessment, θ is fixed. In particular, it is possible to prove certain learning procedures are strongly calibrated, but no learning procedure can be expected to attain correct frequentist coverage in general. The literature on frequentist assessment therefore focuses on weaker notions of coverage, such as *asymptotically correct frequentist coverage*, where the data are of the form $y = (y_1, \ldots, y_n)$ and credible sets are required to have correct frequentist coverage in the $n \to \infty$ limit. In finite-dimensional Bayesian analyses where a Bernstein–von–Mises theorem holds, asymptotically correct frequentist coverage is guaranteed (Freedman, 1999). Results on frequentist coverage have also been established in finite dimensions for variational Bayes (Wang and Blei, 2019). In infinite-dimensional settings, a Bayesian learning procedure can fail to have even asymptotically correct frequentist coverage (Cox, 1993; Freedman, 1999). An active area of research is to establish sufficient conditions for asymptotically correct frequentist coverage, and recent results have been established that hold uniformly over a set of values for θ_0; for results in this direction see Szabó et al. (2015) and references therein.

Forecast Assessment: Dawid (1984) refers to the question of whether a probabilistic forecasting system is in some sense "good" as "*the fundamental question of prequential statistics*". Our notion of strong calibration is closely related to a concept developed in that literature to answer this question, for which the term *probabilistic calibration* is used (Dawid, 1982; Diebold et al., 1997; Gneiting et al., 2007; Gneiting and Ranjan, 2013). An important distinction between forecast assessment and the present paper is the sense in which probabilistic calibration is applied; here we estimate a "true" parameter θ, which is not a random variable, whereas in forecast assessment there remains inherent randomness in the quantities being predicted.

In the econometrics literature, Diebold et al. (1997) considered a sequence of forecasts $(Q_i)_{i=1}^n \subset \mathcal{P}_r(\mathbb{R})$, representing predictions for corresponding quantities $(q_i)_{i=1}^n \subset \mathbb{R}$. The authors advocated a visual diagnostic, called a *correlogram*, to assess whether $\{F_{Q_i}(q_i)\}_{i=1}^n$ are plausible as an independent random sample from $\mathcal{U}(0,1)$; see also Christoffersen (1998); Berkowitz (2001). In the statistics community, Gneiting and Ranjan (2013) proposed to compare the variance of the $\{F_{Q_i}(q_i)\}_{i=1}^n$ to $1/12$, the variance of a $\mathcal{U}(0,1)$ random variable,

with the sequence of forecasts being called *overdispersed* if this variance is smaller than $1/12$, and *underdispersed* if it is larger; see the review of Gneiting and Katzfuss (2014). This literature contains elements that are similar in spirit to our notion of strong calibration, except that a parametric statistical model is not explicitly involved; an important distinction that we require when assessing whether a learning procedure is calibrated.

In the meteorology literature, the calibration of probabilistic forecasts is routinely assessed using *rank histograms* (Anderson, 1996; Talagrand et al., 1997; Hamill and Colucci, 1997; Hamill, 2001). For computational reasons, a forecast is typically represented by a discrete distribution $\mu(\mu_0, y) \approx \frac{1}{M}\sum_{m=1}^{M}\delta(\theta^m)$, produced based on initial belief μ_0 and after observing data y, assumed to have arisen from a data-generating model P. To assess the forecast, an ensemble of synthetic datasets $\{y^m\}_{m=1}^M$ is simulated as $y^m \sim P_{\theta^m}$. For a test function $f \in \mathcal{F}_Y$, the rank statistic

$$r(\{f(y^m)\}_{m=1}^M, f(y)) \coloneqq \sum_{m=1}^{M} \mathbb{I}[f(y^m) < f(y)]$$

will be uniformly distributed on $\{0, 1, \ldots, M\}$ if the forecast is calibrated. This is assessed empirically by producing a histogram of rank statistics for a collection of T ensembles of synthetic datasets $\{\{y_t^m\}_{m=1}^M\}_{t=1}^T$ and corresponding real datasets $\{y_t\}_{t=1}^T$, where $t = 1, \ldots, T$ may index distinct times, spatial locations, or both. Denoting the empirical measure associated with an ensemble of synthetic datasets as $\hat{\nu}_t = \frac{1}{M}\sum_{m=1}^{M}\delta(y_t^m)$, the rank statistic $r(\{f(y_t^m)\}_{m=1}^M, f(y_t))$ is related to the CDF of $\hat{\nu}_t$ by

$$F_{f_{\#}\hat{\nu}_t}(f(y_t)) = \frac{1}{M} r(\{f(y_t^m)\}_{m=1}^M, f(y_t)).$$

Checking for rank histogram uniformity is therefore similar in spirit to the test for strong calibration in Remark 10, with relaxations to allow for the fact that the learning procedure produces an empirical distribution output and that the true parameters $\{\theta_t\}_{t=1}^T$ that gave rise to the real datasets $\{y_t\}_{t=1}^T$ are unknown, so that testing occurs in the data domain Y rather than in the parameter domain Θ.

Signal Processing: An important goal in signal processing is to estimate a time-dependent latent state $\{\theta_t\}_{t=1}^T$, $\theta_t \in \mathbb{R}^d$, based on time-series data $\{y_t\}_{t=1}^T$. For Gaussian filtering algorithms, such as the extended Kalman filter (see Law et al., 2015, p84), the output of the learning procedure is a sequence of Gaussian distributions $\mathcal{N}(m_t, \Sigma_t)$. These serve to quantify uncertainty as to the unknown value of the parameter θ_t, $t = 1, \ldots, T$. Such a filtering algorithm is considered to be calibrated if the *Z-score* $\Sigma_t^{-1/2}(\theta_t - m_t)$ is plausible as a sample from $\mathcal{N}(0, 1)$. The *average normalised estimation error squared* (ANEES) (Bar-Shalom and Birmiwal, 1983; Drummond et al., 1998)

$$\frac{1}{T}\sum_{t=1}^{T}(\theta_t - m_t)^\top \Sigma_t^{-1}(\theta_t - m_t)$$

attempts to quantify this property, with values of ANEES close to 1 when the learning procedure is calibrated. Li et al. (2002) argued against the use of ANEES on the grounds

that it "*penalises optimism much more severely than pessimism*".[3] These authors then proposed the *non-credibility index* (NCI)

$$\frac{10}{T}\sum_{t=1}^{T}\log_{10}\left(\frac{(\theta_t - m_t)^\top \Sigma_t^{-1}(\theta_t - m_t)}{(\theta_t - m_t)^\top \bar{\Sigma}_t^{-1}(\theta_t - m_t)}\right)$$

where $\bar{\Sigma}_t$ is the covariance matrix of the random vector $\theta_t - m_t$, where the randomness here refers to the generation of the dataset. The NCI, which is also called the *inclusion indicator* in Li and Zhao (2006), takes values close to 0 if the filtering algorithm is calibrated and is quite widely used (e.g. Prüher et al., 2020). Further discussion can be found in Li et al. (2011). The ANEES is similar in spirit to our Definition 8, but it is adapted to learning procedures that produce Gaussian output and to a temporal data-generating model.

Validation of Algorithms for Bayesian Computation: Cook et al. (2006) observed that Bayesian inference is strongly calibrated to the prior and the data-generating model[4] and presented the argument used in Example 1. Their interest was in validating software for Bayesian inference, and general learning procedures were not considered. They proposed a goodness-of-fit test for the case $\Theta = \mathbb{R}^d$ that corresponds to Remark 10, using a test statistic of the form

$$T \coloneqq \sum_{i=1}^{n}(F^{-1}_{\mathcal{N}(0,1)}(F_{f_{\#}\mu(\mu_0,y_i)}(f(\theta_i))))^2 \tag{4}$$

for some $f \in \mathcal{F}_\Theta$. If the null hypothesis holds and the learning procedure is strongly calibrated, then $T \sim \chi^2_n$. Cook et al. (2006) focused on software that uses *Markov chain Monte Carlo* (MCMC), meaning that CDFs are not exactly computed, and advocated an empirical approximation to the CDF based on approximate samples $\{\theta_i^m\}_{m=1}^M$ from $\mu(\mu_0, y_i)$ generated using MCMC.

A similar approach was used to analyse ABC in Wegmann et al. (2009), who performed a *Kolmogorov–Smirnov* (KS) test for uniformity, and in Prangle et al. (2014) who used the name *coverage property* and advocated a visual diagnostic plot. In more recent work, Lee et al. (2019); Xing et al. (2019) proposed the use of credible sets to circumvent access to CDFs; this is similar in spirit to taking f to be an indicator function in Definition 8. In Talts et al. (2018) the authors modified the approach of Cook et al. (2006) to address issues surrounding empirical approximation of the CDF, such as discretisation artefacts when displayed as a histogram if an appropriate continuity correction or binning scheme is not used. Talts et al. (2018) showed that, for IID samples $\{\theta_i^m\}_{m=1}^M$ from the posterior given y_i, rank statistics $r(\{f(\theta_i^m)\}_{m=1}^M, f(\theta_i))$ for a test function $f \in \mathcal{F}_\Theta$ will follow a discrete uniform distribution on $\{0, 1, \dots M\}$, and proposed to use this to test calibration rather than checking the (continuous) uniformity of estimated quantiles. Further, Talts et al. (2018) proposed to alleviate departures from uniformity in the rank statistics arising from the use of dependent MCMC rather than IID samples by thinning the MCMC samples using a heuristic based on the estimated chain autocorrelations.

3. It is unclear to us whether this is a problem, since in most statistical applications estimates that are conservative are generally preferred to estimates that are over-confident.
4. Though, the result was not described in such terms in that work.

Validation of Bayesian Workflows: The aforementioned authors including Cook et al. (2006) focussed on the correctness of algorithms for Bayesian computation, but one can take a broader view in which a *Bayesian workflow* (e.g. including prior elicitation, selection of a likelihood, and so forth; see Gelman et al. (2020)), also form part of the learning procedure to be assessed. The earliest related work in this direction of which we are aware is Monahan and Boos (1992), who stated a definition similar to our strong calibration (albeit in terms of credible sets). These authors considered generalised Bayesian inference and provided the argument used in Example 3. A KS test for uniformity of $F_{\mu(\mu_0,y_i)}(\theta_i)$ was proposed in the case where Θ is one-dimensional.

Harrison et al. (2015) proposed a notion of calibration that is similar in spirit to our Definition 8, motivated by the often challenging computational workflows encountered in applications to astronomy. First, the authors take a collection of candidate values θ_i for the parameter and generate associated datasets $y_i \mid \theta_i \overset{\text{IID}}{\sim} P_{\theta_i}$. The values θ_i "*may be the same for each simulation generated or differ between them, depending on the nature of the inference problem*". Then, recasting into our notation, these authors proposed to "*test the null hypothesis that each set of assumed parameter values θ_i is drawn from the corresponding derived posterior $\mu(\mu_0, y)$*". This procedure coincides with our notion of strong calibration only if $\theta_i \overset{\text{IID}}{\sim} \mu_0$. The authors considered Bayesian workflows ("*our validation procedure [...] allows for the verification of the implementation and any simplifying assumptions of the data model*") and proposed a "*multiple simultaneous version of [a novel, multi-dimensional] Kolmogorov–Smirnov test*" for the calibrated null hypothesis. This *multi-dimensional Kolmogorov–Smirnov* (MKS) test provides an ingenious way to circumvent the selection of a test function f in Remark 10, being based on highest probability density regions instead of CDFs. However, the MKS test does not have power against all alternatives to the calibrated null hypothesis, even in dimension $d = 1$, and the description of the test as a multi-dimensional KS test is misleading, as when $d = 1$ the test does not correspond to a standard KS test.

Summary: In summary, the content of Sections 2.1 to 2.3 and 2.3.1 departs from existing work on this topic in that:

1. where similar hypothesis tests have been performed in Monahan and Boos (1992); Cook et al. (2006); Harrison et al. (2015), they were used only to verify the correctness of algorithms and/or workflows for some form of Bayesian computation, while we proposed a notion of strong calibration that is ambivalent to any particular statistical framework;

2. Definition 8 is sufficiently precise to allow for logical deduction, such as proving the strong calibration property holds for a non-traditional learning procedure such as that in Example 4.

The main drawback with Definition 8 appears to be practical, since testing for strong calibration in principle requires access to the CDF of $f_{\#}\mu(\mu_0, y)$ for at least one test function $f \in \mathcal{F}_\Theta$. In some cases the CDF will be explicitly available or easily approximated, but in other cases it will not. Therefore, in the next section we propose a second, strictly weaker notion of calibration which can be tested without access to the CDF.

2.4 Weakly Calibrated Learning Procedures

Testing whether a learning procedure is strongly calibrated may be challenging in practice. Furthermore, as discussed in Section 2.3, the requirement that both μ_0 and the learning procedure are regular in the sense of Definitions 2 and 7 will often be too strong, given the diverse algorithms for uncertainty quantification that have been proposed in literature. We therefore propose a second, weaker definition that requires neither additional structure to define a CDF nor regularity of the distributions involved:

Definition 13 (Weakly Calibrated)**.** *Let $B \subseteq \mathcal{P}(\Theta)$ denote a set of belief distributions and P a data-generating model. A learning procedure μ is said to be* weakly calibrated *to (B, P) if either of the following equivalent properties hold:*

(i) $\displaystyle\iint \mu(\mu_0, y)\,\mathrm{d}P_\theta(y)\,\mathrm{d}\mu_0(\theta) = \mu_0.$

(ii) $\theta \mapsto \displaystyle\int \mu(\mu_0, y)\,\mathrm{d}P_\theta(y)$ *is a μ_0-invariant Markov kernel on Θ.*

for all $\mu_0 \in B$. If the set B contains a single element, μ_0, we say simply that μ is weakly calibrated to (μ_0, P).

To give some intuition, the definition (i) above states that if one randomises the true parameter according to $\theta \sim \mu_0$, generates synthetic data according to $y \sim P_\theta$, and then samples $\vartheta \sim \mu(\mu_0, y)$ from the distributional output, this should be identical in distribution to sampling ϑ from μ_0 directly. Similarly to Remark 12, one could consider quantifying departures from weak calibration in terms of a statistical divergence between the two measures appearing in (i), but here we focus on testing for equality and quantitative descriptions will not be pursued. Focussing on (ii), note that a sufficient condition is provided by the *detailed balance* condition (Eq. 20.5 in Meyn and Tweedie, 2009)

$$\mu_0(\mathrm{d}\theta) \int \mu(\mu_0, y)(\mathrm{d}\vartheta)\,\mathrm{d}P_\theta(y) = \mu_0(\mathrm{d}\vartheta) \int \mu(\mu_0, y)(\mathrm{d}\theta)\,\mathrm{d}P_\vartheta(y), \qquad \forall \theta, \vartheta \in \Theta. \tag{5}$$

On the other hand, the existence of non-reversible Markov kernels that are μ_0 invariant (e.g. Bierkens, 2016) demonstrates that (5) is not a necessary condition for (ii) to hold.

The main practical advantage of Definition 13 is that we may test whether a learning procedure is weakly calibrated without access to CDFs of any univariate summary $f_\# \mu(\mu_0, y)$, $f \in \mathcal{F}_\Theta$:

Remark 14 (Testing whether a Learning Procedure is Weakly Calibrated)**.** *Let $\mu_0 \in \mathcal{P}(\Theta)$ and let*

$$\begin{aligned} \theta_i &\overset{\text{IID}}{\sim} \mu_0 \\ y_i \mid \theta_i &\overset{\text{IID}}{\sim} P_{\theta_i} \\ \vartheta_i \mid \theta_i, y_i &\overset{\text{IID}}{\sim} \mu(\mu_0, y_i). \end{aligned}$$

Then weak calibration of a learning procedure μ to (μ_0, P) can be tested using any goodness-of-fit test for the null hypothesis that $\vartheta_i \overset{\text{IID}}{\sim} \mu_0$. Alternatively if μ_0 and μ are each regular,

one could instead test for weak calibration by picking one or more functions $f \in \mathcal{F}_\Theta$ and using any goodness-of-fit test for the null hypothesis

$$F_{f_{\#}\mu_0}(f(\vartheta_i)) \overset{\text{IID}}{\sim} \mathcal{U}(0,1).$$

This is of course equivalent to the procedure described in Remark 14 provided a sufficiently large set of $f \in \mathcal{F}_r(\Theta)$ are used, but we write it in this way to draw a comparison with Remark 10.

2.4.1 Examples of Weakly Calibrated Learning Procedures

A natural question is whether a learning procedure that is strongly calibrated to (B, P) is also weakly calibrated to (B, P), as the nomenclature suggests. This is indeed the case, as stated below and proven in Section A.2.

Lemma 15 (Strongly Calibrated $\implies$ Weakly Calibrated)**.** *Let $\Theta = \mathbb{R}^d$ for some $d \in \mathbb{N}$. Suppose that μ is a regular learning procedure that is strongly calibrated to (B, P), where $B \subseteq \mathcal{P}_r(\Theta)$ and P is a data-generating model. Then the learning procedure μ is also weakly calibrated to (B, P).*

By virtue of Lemma 15, the learning procedures that were shown to be strongly calibrated in Section 2.3.1 are also weakly calibrated. However, the converse is not true in general, and the following example provides a cautionary tale:

Example 5 (Weakly Calibrated $\not\implies$ Strongly Calibrated)**.** *A learning procedure μ may produce quite unreasonable distributional output $\mu(\mu_0, y)$ and yet be weakly calibrated. As a concrete example, consider $\Theta = \mathbb{R}$, an initial belief distribution $\mu_0 = \mathcal{N}(0, 1)$, and a data-generating model $P_y(\theta)$ distributed according to $y = \theta + \epsilon$, with independent noise $\epsilon \sim \mathcal{N}(0, 1)$. The Bayesian learning procedure produces $\mu(\mu_0, y) = \mathcal{N}(y/2, 1/2)$ and is both weakly and strongly calibrated to (μ_0, P_y) (see the left hand panel in Figure 1). The "mirror Bayes" learning procedure, which flips the sign of the datum y before the Bayesian learning procedure is applied, produces $\mu(\mu_0, y) = \mathcal{N}(-y/2, 1/2)$, which is not strongly calibrated to (μ_0, P_y) but is nevertheless weakly calibrated to (μ_0, P_y) (see the right hand panel in Figure 1).*

Thus there is a trade-off between strong and weak calibration, where the more straightforward approach to testing afforded by weak calibration occurs at the expense of failing to rule out pathologically bad learning procedures, such as Example 5.

An important class of learning procedures that are widely used and yet are *not* weakly calibrated are the generalised Bayesian learning procedures (Bissiri et al., 2016). These are typically not weakly calibrated to the data-generating model and the prior, since these learning procedures are motivated by the *M-open* setting (Bernardo and Smith, 1994, §6.1.2) where the data-generating model may be misspecified. A canonical example of a generalised Bayesian procedure is presented next:

Example 6 (Fractional Posteriors are not Weakly Calibrated)**.** *To avoid technical obfuscation, in this example we abuse notation and assume that μ_0 and $\mu(\mu_0, y)$ can be identified with densities with respect to the reference measure λ on Θ, i.e. $\mu_0(A) = \int_A \mu_0(\theta) \, \mathrm{d}\lambda(\theta)$*

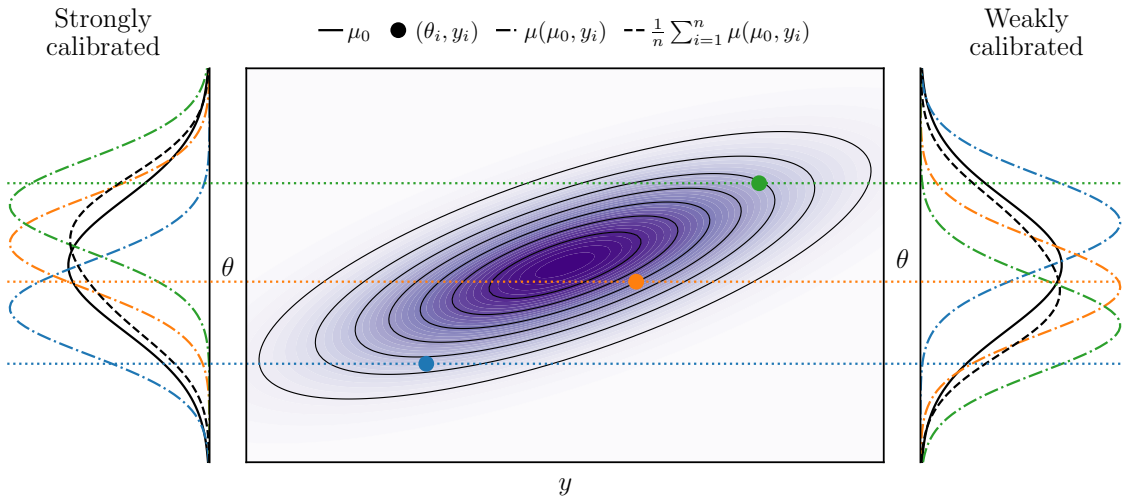

Figure 1: Strong versus weak calibration. Centre: (θ_i, y_i) pairs (blue, orange, green circles) were generated from the joint distribution described by a reference distribution μ_0 and data generating model P (purple heatmap and contours). Left: Distributional output $\mu(\mu_0, y_i)$ (blue, orange, and green dash-dotted lines) from a learning procedure that is strongly calibrated. Note how the true parameters θ_i (blue, orange, and green dotted lines) are plausible as samples from their associated distributions $\mu(\mu_0, y_i)$. Right: Distributional output from a learning procedure that is not strongly calibrated but nevertheless weakly calibrated. Note that the average of the distributional outputs $\mu(\mu_0, y_i)$ (black dashed line) is close to μ_0 (solid black line), even though the individual θ_i, in some cases, lie far out in the tails of the associated distributions $\mu(\mu_0, y_i)$, and thus are not plausible as samples from said distributions.

for each μ_0-measurable set A (and analogously for $\mu(\mu_0, y)$). Similarly, we assume that P_θ admits a density $p(\cdot \mid \theta)$ with respect to a suitable reference measure $\mathrm{d}y$ on Y.[5]

Here we consider fractional posteriors (Bhattacharya et al., 2019), a prototypical instance of a generalised Bayesian learning procedure. As with partial posteriors in Example 3, fractional posteriors have been proposed as a remedy for model misspecification (e.g. in SafeBayes, Grünwald and van Ommen (2017)). The distributional output of a fractional posterior is defined as $\mu(\mu_0, y)(\theta) := p(y|\theta)^t \mu_0(\theta)/p_t(y)$, $\theta \in \Theta$, where $t \in [0, 1]$ and we have defined $p_t(y) := \int p(y \mid \vartheta)^t \mu_0(\vartheta)\,\mathrm{d}\vartheta$, assuming that $p_t(y) > 0$. As an example, consider $\mu_0 = \mathcal{N}(0, 1)$, $p(y \mid \theta) = \mathcal{N}(y; \theta, \sigma^2)$, $\sigma > 0$. Our aim is to verify condition (i) in Definition 13, which requires the distribution

$$\iint \mu(\mu_0, y)\,\mathrm{d}P_\theta(y)\,\mathrm{d}\mu_0(\theta) = \mathcal{N}\left(\vartheta; 0, \frac{t^2(\sigma^2+1) + \sigma^2(t+\sigma^2)}{(t+\sigma^2)^2}\right)$$

5. Note that this is not the same as assuming μ and μ_0 are regular, since their PDFs are not required to be positive on Θ.

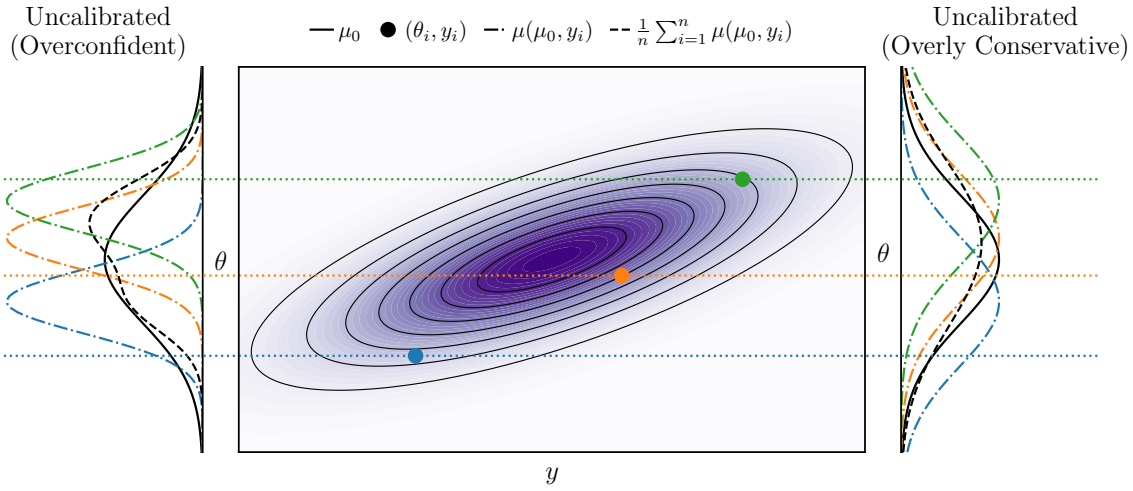

Figure 2: Potential consequences of uncalibrated methods. Centre: as in Fig. 1. Left: Distributional output $\mu(\mu_0, y_i)$ (blue, orange, and green dash-dotted lines) from a learning procedure that is *overconfident*. The true parameters θ_i will tend to be further in the tails of $\mu(\mu_0, y_i)$. Right: Distributional output from a learning procedure that is *overly conservative*. True parameters θ_i tend to be close to the mode of $\mu(\mu_0, y_i)$. The width of the distribution $\mu(\mu_0, y_i)$ suggests that, when the mode is used as an estimator, this estimator has higher uncertainty than the typical error. In both cases the average of the distributional outputs (black dashed line) differs from μ_0 (solid black line).

to be equal to $\mathcal{N}(\vartheta; 0, 1)$*, i.e.*

$$\frac{t^2(\sigma^2+1)+\sigma^2(t+\sigma^2)}{(t+\sigma^2)^2} = 1 \implies t^2(\sigma^2+1)+\sigma^2(t+\sigma^2) = (t+\sigma^2)^2 \implies \sigma^2 t(t-1) = 0.$$

Thus, fractional posteriors are weakly calibrated if and only if either $t = 1$*, which reduces to standard Bayesian inference (Example 1), or* $t = 0$*, which is data-agnostic (Example 2).*

Finally we present two examples of learning procedures that are neither strongly nor weakly calibrated, to demonstrate the potential consequences of methods not being calibrated.

Example 7 (Consequences of Uncalibrated Methods)**.** *We return to the setting of Example 5. Recall that we have an initial belief distribution* $\mu_0 = \mathcal{N}(0, 1)$ *and a data-generating model* $P_y(\theta)$ *such that* $y = \theta + \epsilon$ *with independent noise* $\epsilon \sim \mathcal{N}(0, 1)$*. The Bayesian learning procedure* $\mu(\mu_0, y) = \mathcal{N}(y/2, 1/2)$ *is both weakly and strongly calibrated to* (μ_0, P_y)*.*

Consider the setting of learning procedures that return distributional output $\tilde{\mu}(\mu_0, y) = \mathcal{N}(y/2, \eta/2)$ *for some* $\eta > 0$*, that is, the procedures have the same mean as the Bayesian learning procedure but a different variance for* $\eta \neq 1$*. We illustrate the output in Fig. 2. When* $\eta = 0.5 < 1$ *(left panel), the learning procedures are* overconfident*. The output* $\tilde{\mu}(\mu_0, y)$ *is narrower and more peaked than the correctly specified Bayesian procedure*

$\mu(\mu_0, y)$, with the consequence that the true parameter θ_i typically lies further in the tails of the distribution than for the correctly specified procedure. Thus, the misspecified procedure will often suggest a high degree of confidence in the wrong value of the parameter θ_i.

Conversely, when $\sigma = 2 > 1$ (right panel), the learning procedures are overly conservative. *The procedure $\tilde{\mu}(\mu_0, y)$ produces a distributional output that is wider and flatter than the correctly specified Bayesian procedure $\mu(\mu_0, y)$. Thus the true value of the parameter θ_i will typically be closer to the mean than the posterior variance would suggest, with the consequence that a user will often associate an accurate estimator of θ with a high degree of uncertainty. In both cases $\sigma = 0.5$ and $\sigma = 2$ note that the average of $\mu(\mu_0, y_i)$ differs from μ_0.*

2.4.2 Relation to Existing Concepts

Here we compare and contrast our notion of weak calibration with concepts appearing in earlier work and in related fields.

Forecast Assessment: Our notion of weak calibration is closely related to a concept developed in the literature on forecast assessment, for which the term *marginal calibration* is used (Gneiting et al., 2007). As previously mentioned in Section 2.3.2, an important distinction between forecast assessment and the present paper is the sense in which notions such as probabilistic calibration and marginal calibration are applied. This leads to major differences between forecast assessment and the present work. For example, probabilistic calibration does not imply marginal calibration in the context of forecast assessment,[6] while our notion of strong calibration *does* imply weak calibration in the context of testing whether learning procedures are calibrated, as established in Lemma 15.

Validation of Algorithms for Bayesian Computation: The invariance property that underpins our notion of weak calibration has previously been noted in the Bayesian context. Talts et al. (2018) call this "*self-consistency of the data-averaged posterior*". It appears to have been first used in Geweke (2004), who proposed to use it to check the correctness of MCMC algorithms and their code. Therein, the author proposed to alternatively sample from $y \mid \theta$ and $\theta \mid y$, the latter using MCMC. For a correctly implemented MCMC method, θ will be marginally distributed according to the prior μ_0 after an initial burn-in period has passed. Geweke (2004) performed a collection of univariate hypothesis tests for this weak calibration null hypothesis, followed by a Bonferroni correction to adjust for multiple testing. Our Definition 13 is similar in spirit, but is precise enough to permit logical deduction, such as Lemma 15, and yet general enough to cover learning procedures which need not exist within a Bayesian context. Additionally, we do not assume the structure of MCMC that is required to render this Gibbs-like approach practical.

This completes our formal discussion of what it means for a learning procedure to be called "calibrated". The next section presents several vignettes designed to illustrate our the general framework.

6. A simple example of a forecaster who is marginally calibrated but not probabilistically calibrated is provided by the *unfocussed forecaster* of Gneiting et al. (2007); see also Hamill (2001). These examples have no analogue in our context, due to the fact that there is no inherent randomness in the "true" parameter θ, while the quantity being predicted is inherently random in the setting of forecast assessment.

3. Vignettes

In this section we exploit our framework to test whether or not several popular learning procedures are calibrated, with five separate vignettes presented. The first two vignettes, Sections 3.1 and 3.2, consider learning procedures that are motivated as being approximations to Bayesian inference and are widely used: Gaussian approximations to non-Gaussian posteriors and *approximate Bayesian computation*, respectively. In challenging applications, the output produced using these approximations can fail to resemble the usual Bayesian posterior; we therefore view these approximations as learning procedures in their own right and we ask whether their distributional output is calibrated. Section 3.3 presents a topical application to recently developed probabilistic *ordinary differential equation* (ODE) solvers. Section 3.4 concerns the challenge of performing a goodness-of-fit test for strong calibration in multiple dimensions, where a suitable test function f must first be identified. The final vignette, Section 3.5 examines how our notions of calibration can be extended to the setting where the data-generating model is misspecified.

3.1 Gaussian Approximations

A common approach in statistics is to output a Gaussian distribution which approximates, in some sense, the distributional output of an idealised learning procedure. The targeted learning procedure will often be Bayesian inference, however Gaussian approximations can also be used within different inferential paradigms. As an example of such an approach, Gaussian approximations are often the output of variational inference methods, wherein the learning procedure outputs the member of a family of distributions (in this case Gaussian) which minimises a divergence from the target distribution (Blei et al., 2017). A distinct but related approach is that of fitting a Gaussian approximation based on only local information. The Laplace approximation, which outputs a Gaussian distribution centred at a maximum of the log density of the target distribution and with covariance equal to the inverse of the Hessian of the log density at this point, is a canonical example of such a method.

As a first simulation study we test the calibration of Laplace approximations to the Bayesian posterior in a model with a location parameter θ. We assign a prior $\mu_0 = \mathcal{N}(0, 1)$, and a Student's t data-generating model P_θ such that y consists of n independent draws from a $\mathcal{T}(\theta, 1, \nu)$ distribution. To be specific, $y = (y^{(n)})_{n=1}^N$, with $y^{(n)} \overset{\text{IID}}{\sim} \mathcal{T}(\theta, 1, \nu)$ for $n \in \{1, \ldots, N\}$.

The true posterior in this case is non-Gaussian and so our expectation is that a Laplace approximation will be neither strongly nor weakly calibrated. However, for $\nu \to \infty$ or $N \to \infty$ (and $\nu > 2$) the posterior will become increasingly close to Gaussian, in the former case due to the Student's t distribution becoming increasingly close to Gaussian as $\nu \to \infty$, and in the latter due to the asymptotic normality of the posterior as $N \to \infty$ by the Bernstein–von Mises theorem for $\nu > 2$. We therefore would expect it to be increasingly challenging for the tests in Remark 10 and Remark 14 to reject respectively strong and weak calibration as $\nu \to \infty$ or $N \to \infty$.

In univariate cases such as this, we may employ the identity test function $f(\theta) = \theta$ and a one-sample KS test to check for uniformity in the tests in Remarks 10 and 14. Laplace approximations were computed for 10^6 realisations from the hierarchical model $\theta_i \overset{\text{IID}}{\sim} \mu_0$, $y_i \mid \theta_i \overset{\text{IID}}{\sim} P_{\theta_i}$, for each of $\nu \in \{1, 2, \ldots 20\}$ with $N = 5$ and for each of $N \in \{1, 2, \ldots 20\}$

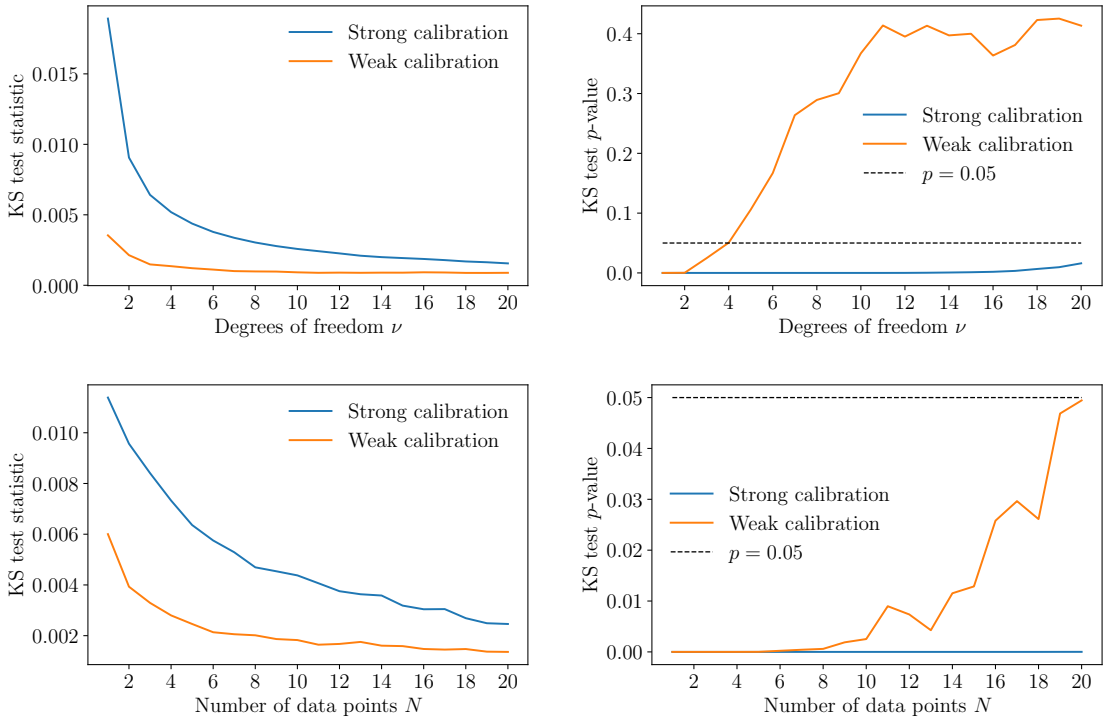

Figure 3: Gaussian approximations: KS test statistics (left) and p-values (right) for strong and weak calibration of Laplace approximations in the t-distribution example for varying degrees of freedom ν and $N = 5$ (top) and varying number of data N and $\nu = 3$ (bottom).

with $\nu = 3$. The strong and weak calibration test results are summarised in Figure 3. As expected, we see that the power of both the strong and weak calibration tests decrease as ν and N increase, with the KS test statistics (defined in (7)) showing decreasing departures from uniformity. While the strong calibration test rejects the null hypothesis at a 0.05 significance criterion for all values of ν and N tested, the weak calibration test fails to reject at a 0.05 level for most of the ν range. However, for the results with varying N, we see that weak calibration test correctly rejects the null hypothesis at a 0.05 significance level up to $N = 20$.

A test of strong calibration is clearly preferable to a test of weak calibration in situations where it is possible to be performed. However, these results indicate that the weaker test in Remark 14 is still able to provide a useful check of calibration in some situations, with the benefit of being simpler to compute and more widely applicable than the test in Remark 10.

3.2 Approximate Bayesian Computation

Performing Bayesian inference in settings for which the data-generating model P_θ does not have a tractable PDF is challenging, with ABC methods (Beaumont et al., 2002) often used

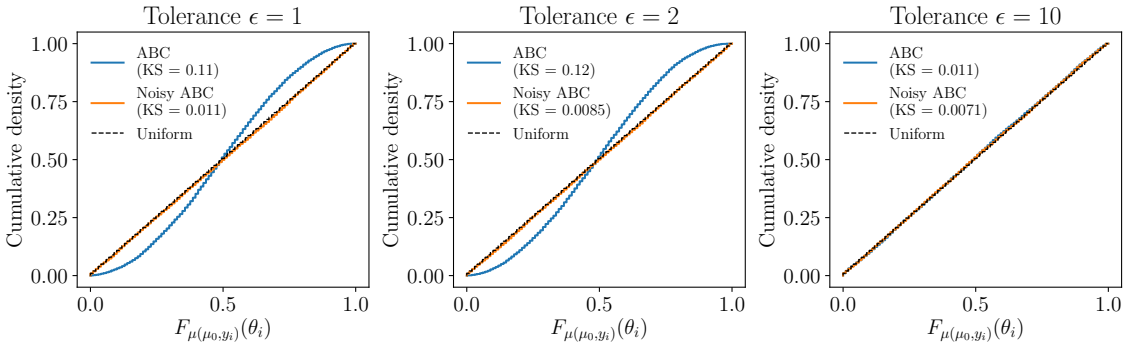

Figure 4: Approximate Bayesian computation: Empirical CDFs for both ABC and noisy ABC in the g-and-k quantile distribution example, for tolerances $\epsilon = 1$ (left), $\epsilon = 2$ (middle) and $\epsilon = 10$ (right). The values of the KS test statistics are shown in the legend.

as an alternative in such situations. The key idea in ABC is that, in contrast to the standard Bayesian procedure of conditioning on the observed dataset $y = y_{\text{obs}}$, one instead conditions on the event that $d(y, y_{\text{obs}}) < \epsilon$, for some distance $d\colon Y \times Y \to [0, \infty)$ and some *tolerance* $\epsilon > 0$. Typically the distance is specified by embedding the data into a finite-dimensional normed vector space S via a *summary statistic* function $s\colon Y \to S$ and specifying the distance as $d(y, y_{\text{obs}}) = \|s(y) - s(y_{\text{obs}})\|$.

As a consequence of Example 3, the learning procedure that exactly conditions on $s(y) = s(y_{\text{obs}})$, i.e. ABC with tolerance $\epsilon = 0$, is guaranteed to be strongly calibrated. Likewise in the limit of $\epsilon \to \infty$ the ABC posterior will be strongly calibrated, as the posterior will collapse to the prior (see Example 2). For $\epsilon \in (0, \infty)$ the ABC posterior will in general however be neither strongly nor weakly calibrated. To resolve this lack of calibration of ABC methods, Fearnhead and Prangle (2012) proposed the *noisy* ABC algorithm, which is calibrated for any tolerance $\epsilon \geq 0$. Rather than conditioning on the event $\|s(y) - s(y_{\text{obs}})\| < \epsilon$, noisy ABC replaces $s(y_{\text{obs}})$ with noisy summary statistics $\tilde{s}_{\text{obs}}$ generated according to $\tilde{s}_{\text{obs}} = s(y_{\text{obs}}) + \epsilon x$, with x uniformly distributed on the unit ball in S. The distributional output of noisy ABC is the partial posterior based on $\tilde{s}_{\text{obs}}$, which takes into account the additional noise in the data-generating model, and is therefore strongly calibrated by an extension of the argument in Example 3.

Here we consider the parameter inference task for a g-and-k distribution. The g-and-k distribution is defined through the inverse of its CDF (quantile function) and it does not have a closed-form PDF (though the PDF can be evaluated numerically; Rayner and MacGillivray, 2002). Here we aim to infer the location parameter θ, which is assigned a prior $\mu_0 = \mathcal{N}(0, 1)$, given a dataset $y \in \mathbb{R}^N$, $N = 20$, generated according to the data-generating model

$$P_\theta \;:\; y^{(n)} = \theta + b\left(1 + 0.8\,\frac{1 - \exp(-g u_n)}{1 + \exp(-g u_n)}\right) u_n (1 + u_n^2)^k,$$

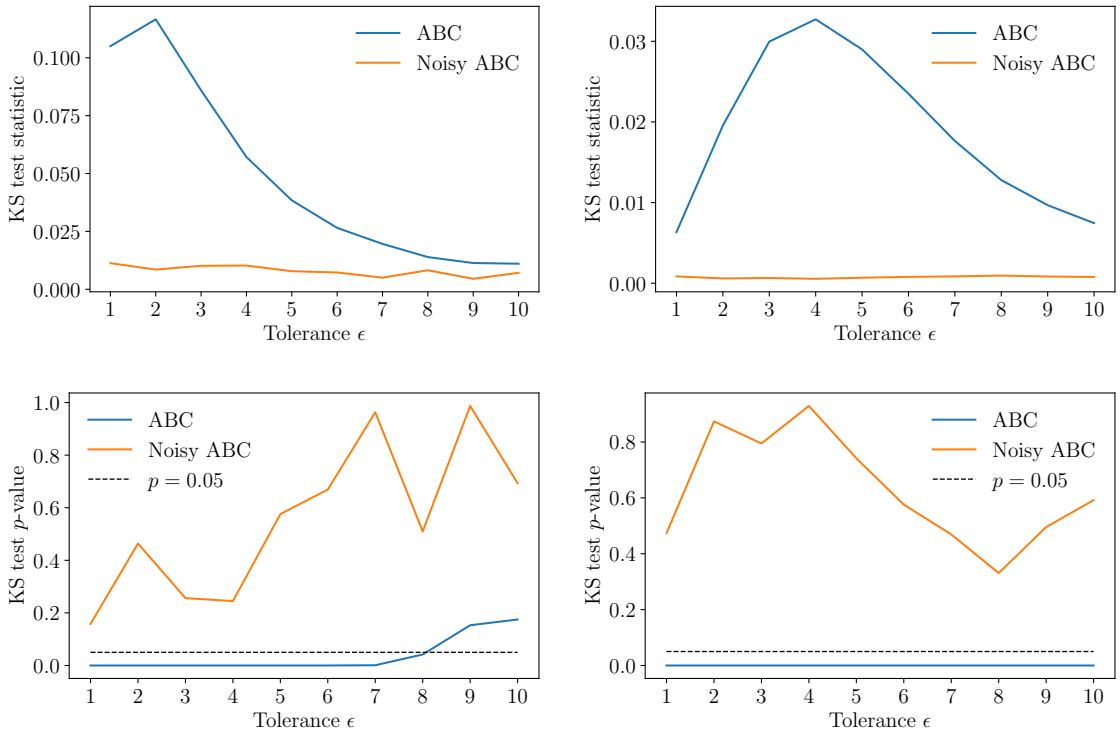

Figure 5: Approximate Bayesian computation: KS test statistics (top) and p-values (bottom) for strong (left) and weak (right) calibration of ABC and noisy ABC in the g-and-k quantile distribution example, for various tolerances $\epsilon > 0$.

with $u_n \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$, $n \in \{1,2,\dots N\}$, $b = 1$, $g = 2$ and $k = 0.5$. For the tests that follow we computed independent realisations from the hierarchical model $\theta_i \overset{\text{IID}}{\sim} \mu_0$, $y_i \mid \theta_i \overset{\text{IID}}{\sim} P_{\theta_i}$. In each case, data were summarised as a vector $s\colon \mathbb{R}^N \to \mathbb{R}^5$ consisting of the five quartiles of the dataset, and rejection sampling was used to generate M samples $\{\theta_i^m\}_{m=1}^M$ from the distributional output of both ABC and noisy ABC, for tolerances $\epsilon \in \{1,2,\dots 10\}$. Single samples ($M = 1$) can be directly used to test for weak calibration, as per Remark 14. However, the intractability of the distributional output for ABC and noisy ABC precludes a straightforward test for strong calibration. Instead, we consider a variant of the test for strong calibration in Remark 10, which in a similar spirit to Talts et al. (2018), wherein we test whether the rank statistics $r(\{\theta_i^m\}_{m=1}^M, \theta_i)$ are IID samples from the discrete uniform distribution on $\{0,1,\dots M\}$. For testing strong calibration, a total of 10^4 realisations of the hierarchical model were considered with $M = 100$, while for the less computationally demanding test for weak calibration a total of 10^6 realisations were considered with $M = 1$.

Figure 4 presents empirical CDFs for both ABC and noisy ABC, on which our test for strong calibration is based. Figure 5 presents the KS test statistics and corresponding p-values for both strong and weak calibration, for different values of the tolerance $\epsilon > 0$. In each case noisy ABC is, as expected, seen to be better calibrated than ABC. Both the strong

and weak calibration tests correctly fail to reject the null hypothesis at a 0.05 significance level for noisy ABC, which is strongly (and weakly) calibrated, for all values of the tolerance ϵ. The strong calibration test fails to reject the null hypothesis that ABC is strongly calibrated for the highest two tolerances $\epsilon \geq 9$. The weak calibration test on the other hand correctly rejects at a 0.05 level the null hypothesis that ABC is weakly calibrated for all ϵ. The apparent greater power of the weak calibration test here likely arises from the much larger number of model realisations used — 10^6 compared to 10^4 for the strong test — for a given computational expenditure due to the need to generate only $M = 1$ ABC sample per realisation rather than $M = 100$. A final interesting point of note is that both weak and strong calibration show a "dip" in the KS test statistic at $\epsilon = 1$, reflecting that as $\epsilon \to 0$ classical ABC tends towards a Bayesian procedure, which is guaranteed to be calibrated.

3.3 Calibration of Probabilistic ODE Solvers

A traditional (adaptive) numerical method for the approximate solution of an ODE accepts, as its input, an error tolerance $\tau > 0$ and returns, as its output, an approximation to the solution of the ODE. In general it is not guaranteed that the resulting approximation has error less than τ, but empirical analysis over a range of typical ODEs can provide reassurance that the error will be below τ for many problems practically encountered. In contrast to the traditional approach, there has been a concerted research effort in recent years to develop *probabilistic numerical method*s (PNMs) for ODEs. A PNM returns a probability distribution over the solution space of the ODE, representing epistemic uncertainty associated with the unknown true solution of the ODE. The scale of this distributional output can be used as the basis for selecting a suitable time step size in order to drive the uncertainty below a user-specified tolerance τ, if desired. Compared to traditional numerical methods, which have benefited from over a century of development, important questions regarding their behaviour of PNMs remain unanswered, including whether such methods are calibrated. Most PNMs exploit *Gaussian process* (GP) models for the solution of the ODE, motivated by mathematical convenience rather than detailed knowledge of the ODE to be solved. These models typically include hyperparameters for the GP, which are jointly estimated along with the solution of the ODE. Given that PNM act on the basis of a default GP model, essentially independent of initial belief μ_0 regarding the ODE at hand, it is unclear whether hyperparameter estimation is sufficient to ensure PNM are calibrated.

The principal application of PNMs for ODEs is to *inverse problems*, where an ODE's parameters are to be estimated based on a dataset. This usually requires the numerical solution of many ODEs, each corresponding to different values of the parameters, to see which parameter values are compatible with the dataset. The motivation for PNMs in this setting is that the solution of the ODEs can be viewed as an unknown latent quantity and integrated out, potentially using a fast-but-crude PNM in place of an adaptive ODE solver and adjusting credible sets for ODE parameters in a way commensurate with the accuracy of the PNM used. However, the success of this approach hinges on whether the underlying PNM is calibrated, as otherwise under- or over-confident parameter inferences could be produced. To shed light on this question, we considered the probabilistic numerical solution of the

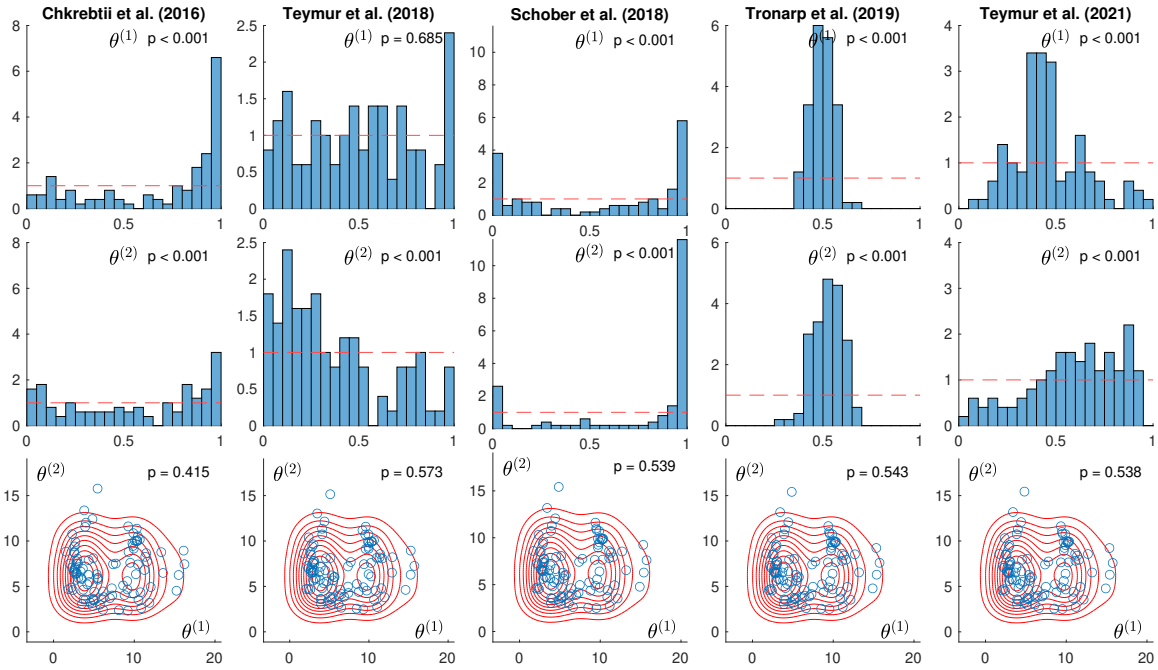

Figure 6: Calibration of probabilistic ODE solvers: Samples (blue) from the strong calibration test statistic $F_{f_{\#}\mu(\mu_0,y_i)}(f(\theta_i))$ (c.f. Remark 10), with $f(\theta) = \theta^{(1)}(10)$ (top row), $\theta^{(2)}(10)$ (middle row), and from the weak calibration test statistic $\vartheta_i \sim \mu(\mu_0, y_i)$ (bottom row; c.f. Remark 14). The reference distribution, corresponding to the null hypothesis that the learning procedures are calibrated, is in each case shown in red. The p-values for the associated hypothesis tests (see main text for details) are shown in the top right-hand corner of each panel.

following Lotka–Volterra ODE

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \begin{bmatrix} \alpha\theta^{(1)} + \beta\theta^{(1)}\theta^{(2)} \\ \gamma\theta^{(2)} + \delta\theta^{(1)}\theta^{(2)} \end{bmatrix}, \quad \theta(0) = \begin{bmatrix} 10 \\ 10 \end{bmatrix}, \quad t \in [0, 10], \tag{6}$$

with an initial belief distribution μ_0 induced over the solution space of differentiable functions $\theta(t)$ on $[0, 10]$ by sampling parameters $(\alpha, \beta, \gamma, \delta)$ from a probability distribution π on $[0, \infty)^4$. For this experiment we took the distribution π to be

$$\alpha, \gamma \overset{\text{iid}}{\sim} \text{logNormal}(0, 0.25), \qquad \beta, \delta \overset{\text{iid}}{\sim} \text{logNormal}(-2, 0.1),$$

which produces a variety of periodic trajectories typically associated with this type of predator-prey model. The following PNMs were considered: Chkrebtii et al. (2016), which employs a particle-based approach requiring parallel simulations to produce empirical credible sets; Teymur et al. (2018), which is based on stochastic perturbation of traditional numerical methods, continuing a line of work that originated in Conrad et al. (2017); Schober et al. (2019) and Tronarp et al. (2019), which are both based on Gaussian filtering but with

different approaches to the (local) linearisation of (6); and Teymur et al. (2021), which is based on a probabilistic version of Richardson extrapolation. Each method has user-defined settings that can in principle affect the selection of its hyperparameters, and thus, its calibration in the senses used in this paper; for this experiment we considered one setting only for each PNM, with full details contained in Appendix B. In particular, default settings were used for some PNMs, whilst the settings of other PNMs were manually selected. Thus we do not claim to draw general conclusions about the specific PNMs involved; our aim is only to show how diverse algorithms can be analysed using the notions of calibration we have introduced.

Tests of strong and weak calibration were performed, in each case using the test functions $f_j(\theta) = \theta^{(j)}(10)$, $j \in \{1, 2\}$, i.e. the value of the solution at the final time point. Results are displayed in Figure 6. The top two rows show histograms of $F_{f_{\#}\mu(\mu_0, y_i)}(\theta_i^{(j)}(10))$ for $j \in \{1, 2\}$, using 100 samples θ_i drawn from μ_0. A Kolmogorov–Smirnov test of uniformity was then used to test whether the PNMs are strongly calibrated (c.f. Remark 10). The bottom panels show scatter plots of samples $\vartheta_i(10)$ where $\vartheta_i \sim \mu(\mu_0, y_i)$, overlaid on contours of μ_0 (empirically obtained). A kernel two-sample test (Gretton et al., 2012) was performed based on samples from the intractable distribution μ_0 to assess whether the PNMs are weakly calibrated (c.f. Remark 14). The results of these simulations show that strong calibration is not a property enjoyed by most PNM at present. The only instance where strong calibration was not emphatically rejected is Teymur et al. (2018), for inference of the first component $\theta^{(1)}(10)$. It is interesting to note that Teymur et al. (2018) performs an exhaustive grid search for GP hyperparameter estimation, which can require more computation compared to the other PNM considered, and this may explain its relative success in this calibration assessment. The remaining PNM perform poorly in different ways, including being over-confident (e.g. Schober et al., 2019) and under-confident (e.g. Tronarp et al., 2019). However, we reiterate that these conclusions will depend on additional user-specified settings, specific to how each PNM is implemented. On the other hand, weak calibration was never rejected, and indeed this was also the case over a much wider variety of algorithm settings (not presented). This suggests that weak calibration of PNMs, in as far as this testing framework is concerned, is indeed a weak requirement.

3.4 Data-Driven Goodness-of-Fit Testing for Strong Calibration

For multivariate parameter inference tasks, where e.g. $\Theta = \mathbb{R}^d$, $d > 1$, it will not be possible in general to identify a single test function $f \in \mathcal{F}_\Theta$ that has power against all alternatives to the strong calibration null. Indeed, even a simultaneous test using all coordinate functions $f_i(\theta) = \theta^{(i)}$, $i = 1, \ldots, d$, does not have power against all alternatives, since a multivariate distribution is not uniquely determined by its univariate marginals. Nevertheless, the richness of the set $\mathcal{F}_\Theta$ is such that we expect *some* $f \in \mathcal{F}_\Theta$ to yield a test with the power to reject the null hypothesis, due to Lemma 5. A strategy to select a suitable test function f is therefore required.

Following a generic approach to goodness-of-fit testing, one way to proceed is to consider splitting the collection of simulated parameter-dataset pairs into two disjoint sets: $\mathcal{S}_1 := \{(\theta_i, y_i)\}_{i=1}^{s}$, $\mathcal{S}_2 := \{(\theta_i, y_i)\}_{i=s+1}^{S}$. The first subset $\mathcal{S}_1$ can be used to identify a suitable test function f, after which a goodness-of-fit test can be conducted using f and $\mathcal{S}_2$. The

independence of $\mathcal{S}_1$ and $\mathcal{S}_2$ ensures that a test conducted in this way is valid. To select a suitable test function, one first identifies a sufficiently small subset $\mathcal{F}_s \subset \mathcal{F}_\Theta$ of test functions and, for each $f \in \mathcal{F}_s$, a univariate goodness-of-fit test is performed using $\mathcal{S}_1$. The element of $\mathcal{F}_s$ that gives rise to the strongest evidence against the null hypothesis, based on $\mathcal{S}_1$, is selected. The main advantage of a data-splitting approach is that the selection of f is data-driven, as opposed to f being user-specified. The role of data to inform the selection of f is anticipated to be increasingly important in higher dimensional settings, $d \gg 1$. To explore this, we consider now a setting that is, at least notionally, infinite dimensional.

Let $\theta : [0,1] \to \mathbb{R}$ be a continuous function-valued parameter, so that $\Theta = C(0,1)$ is the set of continuous functions on $[0,1]$. For μ_0 we consider a hierarchical, non-stationary GP of the form $\theta(x) := \sigma(x)g(x)$, $g \sim \mathcal{GP}(0,k)$ with $k(x,x') := \exp(-(x-x')^2/\ell^2)$, $\sigma \sim \nu$ for some distribution ν to be specified, and for simplicity $\ell = 0.1$ is fixed. Consider the data-generating model that returns $y = (y^{(1)}, \ldots, y^{(10)})$, where $y^{(n)} = \theta(x_n)$ and $x_n \sim \mathcal{U}(0,1)$ are independently sampled. A popular, pragmatic workflow acknowledges the non-stationarity encoded in μ_0 but, for computational convenience, fits instead a stationary, non-hierarchical GP of the form $\theta(x) = \sigma_0 g(x)$, where the scalar σ_0 is estimated using maximum likelihood. Estimating σ_0 from data enables the scale of the distributional output to roughly adapt to the scale of the dataset, but this is insufficient to ensure the learning procedure is strongly calibrated (Karvonen et al., 2020). Our interest here is in whether we can detect failure of strong calibration, and for this purpose we consider a simple form of ν that sets $\sigma(x) = 1 + x$ with probability one. It can be expected that simplified GP regression produces a "compromise" value of σ_0, which leads to under-confident inferences for $\theta(x)$ when x is close to 0 and over-confident inferences when x is close to 1.

For the set of candidate test functions $\mathcal{F}_s$, we consider the evaluation functions $f_x(\theta) := \theta(x)$, indexed by $x \in [0,1]$. A number, S, of parameter-dataset pairs were generated, of which $s = \frac{S}{2}$ were assigned to $\mathcal{S}_1$ and used to identify a promising location $x_* \in [0,1]$ at which to perform a hypothesis test of strong calibration using the held-out $\mathcal{S}_2$. Since the marginals $(f_x)_\# \mu(\mu_0, y)$ are Gaussian, it is natural to use a χ^2_s test, as per (4). Thus we select x_* to minimise the p-value of a two-sided χ^2_s test, based on f_x and computed using $\mathcal{S}_1$, over $x \in [0,1]$. The total number of simulated parameter-dataset pairs S was varied from 10 to 150 and, through repeated simulation, the p-values of a two-sided χ^2_s test of strong calibration, based on the estimated x_* and $\mathcal{S}_1$, were computed. As a baseline, we also computed p-values for a user-specified test function centred at $x_b := 0.5$. In Figure 7 (left) we plot log p-values as a function of x, for $s = 10$ (top) and $s = 150$ (bottom), for one typical realisation of $\mathcal{S}_1$. These results indicate that values of x close to 0 are likely to provide the most power for our hypothesis test. Here x_* is indicated as a vertical red line and x_b indicated as a vertical blue line; the identification of a suitable x_* is seen to be easier when the number, s, of simulations available in $\mathcal{S}_1$ is increased. Finally, in Figure 7 (right) we plot the p-values obtained when the x_*-based and x_b-based tests are applied to $\mathcal{S}_2$. To avoid reporting an artefact of the random seed, average log p-values are reported, along with standard errors, based on 100 independent realisations of $\mathcal{S}_1$ and $\mathcal{S}_2$. It is seen that the data-driven goodness-of-fit test (based on x_*) is more powerful than the user-specified test (based on x_b).

This illustration makes clear that, for a data-splitting approach to work well, the size of the set $\mathcal{F}_s$ of candidate test functions should be carefully controlled, relative to the number

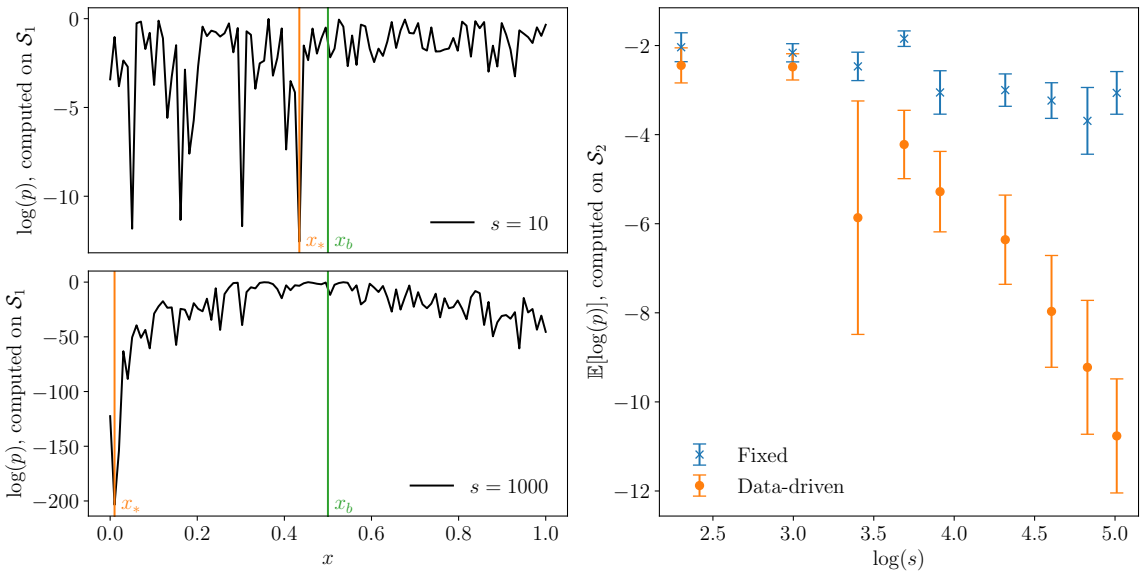

Figure 7: Data-driven goodness-of-fit testing for strong calibration: Here we consider an infinite-dimensional parameter $\theta \in C(0,1)$. On the left, we display typical p-values obtained using a test based on the evaluation functional $f_x(\theta) = \theta(x)$ and a number s of independent simulations of the parameter and dataset (top: $s = 10$, bottom: $s = 150$). The right hand panel displays average p-values obtained when the procedure is applied to 100 independent realisations of $\mathcal{S}_1$ and $\mathcal{S}_2$, each of size s, using either the data-driven choice $x = x_*$ or the fixed choice $x = x_b$ of test.

s of samples in $\mathcal{S}_1$. For example, if we simply took $\mathcal{F}_s = \mathcal{F}_\Theta$, then for any $\alpha \in (0,1)$ there would be infinitely many elements of $\mathcal{F}_s$ for which the null hypothesis is rejected at level α by virtue only of the fact that $\mathcal{S}_1$ is a finite set. Consideration of multiple data splits can also be exploited to increase the power of such a test (Romano and DiCiccio, 2019).

3.5 Robust Calibration

Our proposed notions of strong and weak calibration can be extended to the *M-open* setting (Bernardo and Smith, 1994, §6.1.2) where the data-generating model may be misspecified. This permits us to define notions of "robust calibration", which are analogous (and orthogonal) to the notions of "robust estimation" that are already widely studied (Berger, 1994; Huber and Ronchetti, 2009). For example, suppose that a learning procedure μ is strongly calibrated to (μ_0, P). Then, for any $f \in \mathcal{F}_\Theta$, the distribution $\mathcal{U}_{f,P}$ of the random variable $F_{f_\#\mu(\mu_0,y)}(f(\theta))$, where $\theta \sim \mu_0$, $y \mid \theta \sim P_\theta$, is by definition $\mathcal{U}(0,1)$. Thus, when the data-generating model P is misspecified, we may quantify the loss of strong calibration in terms of a statistical divergence between $\mathcal{U}_{f,Q}$ and $\mathcal{U}(0,1)$.

Here we adopt a more practical perspective, using the framework of Section 2.3 to test the strong calibration null hypothesis in settings where the data-generating model is misspecified. For example, consider a Bayesian learning procedure μ for a location parameter

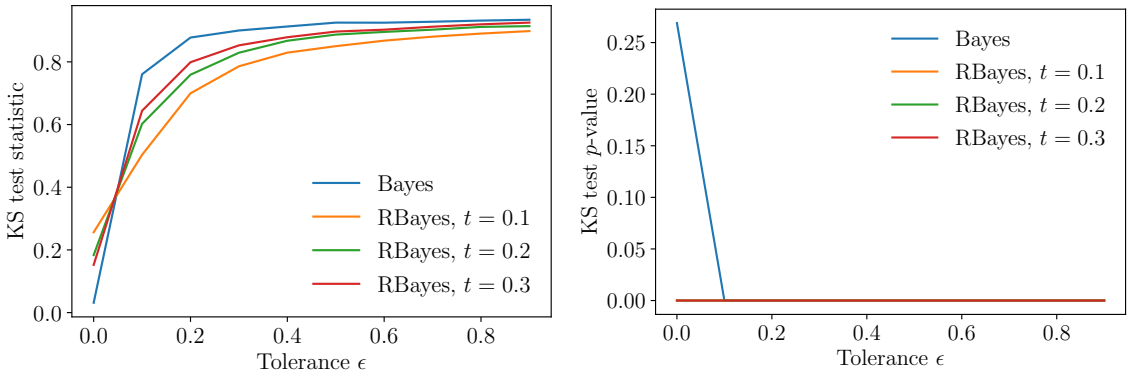

Figure 8: Robust calibration: Results of KS tests for strong calibration, comparing standard Bayesian inference ("Bayes") to the method of fractional posteriors (also called *robust Bayes*; "RBayes") with exponents $t \in \{0.1, 0.2, 0.3\}$, in a setting where the likelihood is misspecified. Note that in the right panel the RBayes lines for $t = 0.1, t = 0.2$ and $t = 0.3$ coincide.

θ, which is assigned a prior $\mu_0 = \mathcal{N}(0, 3)$, based on a likelihood $Y \mid \theta \sim \mathcal{N}(\theta, 1)$. Our assessment will be performed using the data-generating model

$$P_\theta \ : \ Y \mid \theta \sim \begin{cases} \mathcal{N}(\theta, 1) & \text{w.p. } 1 - \epsilon \\ \mathcal{N}(5, 1) & \text{w.p. } \epsilon \end{cases},$$

where $\epsilon \in [0, 1]$ is a probability of obtaining a contaminated observation, so that for $\epsilon > 0$ the likelihood is misspecified and the Bayesian learning procedure is not strongly calibrated to (μ_0, P). Fractional posteriors with exponent $t \in [0, 1]$, as defined in Example 6, have been proposed as learning procedures that can offer robustness to misspecification of the likelihood e.g. in Grünwald and van Ommen (2017). Our aim is to assess this claim within our testing framework.

Results of performing a KS test of the strong calibration null hypothesis, using the identity test function $f(\theta) = \theta$, are displayed for a variety of values of t and ϵ in Figure 8. Clearly the only circumstance in which any of the learning procedures is strongly calibrated is when $\epsilon = 0$ and the Bayesian procedure is used. Otherwise, according to the test statistic in the left panel, fractional posteriors are marginally better calibrated when $\epsilon > 0$ than the Bayesian procedure, though regarding the p-values in the right panel one sees that the values of the statistic in these cases are still sufficiently sufficiently large to emphatically reject the strong calibration null hypothesis.

Finally, we note that other senses of "robust calibration" could be considered, analogous to the various notions of "robust estimation" that have been studied (Berger, 1994; Huber and Ronchetti, 2009). For example, one could consider a setting where true parameters θ are drawn from a distribution other than μ_0 and assess the consequences, in terms of calibration, for a learning procedure that uses μ_0 as the initial belief distribution.

4. Discussion

The desire that a parameter used to generate a dataset should appear plausible as a sample from the distributional output of a learning procedure, such as a Bayesian posterior, is foundational and, at least in an informal sense, widely understood and accepted. Despite this, a precise and widely applicable notion of what it means for a learning procedure to be "calibrated" appears not to have been put forward. Our aim in this paper was to propose such a definition, together with a framework for testing whether a learning procedure is calibrated. In particular, we proposed a property called *strong calibration* (Definition 8), which provides an explicit sense in which output from the learning procedure can be considered to be meaningful. A strictly weaker property, called *weak calibration*, was also proposed (Definition 13), which has the advantage of being more straightforward to test. Several vignettes were provided to illustrate the generality and usefulness of the framework.

Our hope, in writing this manuscript, is to stimulate further critical discussion around calibration as a *desideratum* for a learning procedure, and to bring together some of the disparate strands of literature where related concepts and domain-specific definitions have been developed (cf. Section 2.3.2).

4.1 Further Work

A particularly promising avenue for further research would be to develop *measures* of miscalibration using the ideas proposed in this paper. Generally speaking, when using approximate methods such as Laplace approximation (cf. Section 3.1) or ABC (cf. Section 3.2), or generalised Bayesian methods (cf. Section 3.5), a user has purposefully departed from the Bayesian framework due to challenges such as its lack of computational tractability or the possibility that the model is misspecified. In such settings a measure of miscalibration is likely to be of more use than a test for calibration, since exact calibration cannot be expected to hold. A *measure* of miscalibration might allow a user to select the "most calibrated" method from among multiple alternatives, or perhaps even incorporate calibration into a variational objective in a variational Bayesian framework (Knoblauch et al., 2021).

In the context of Definition 8, such a measure could be constructed by selecting some test function $f^* \in \mathcal{F}_\Theta$ and computing a statistical divergence between $F_{f^*_\# \mu(\mu_0, y)}(f^*(\theta))$ and $\mathcal{U}(0,1)$. The former quantity is unlikely to be available in closed-form but could be estimated using Monte Carlo techniques. Immediate challenges with this would concern selection of a suitable divergence and a suitable f^*. For the latter, one could perhaps instead consider selecting a subset $\mathcal{F}_{\text{test}} \subset \mathcal{F}_\theta$ over which a supremum can be taken tractably. However, we leave this task for future work.

Acknowledgements

JC was supported by Wave 1 of the UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Digital Twins for Complex Engineering Systems" theme within that grant, and the Alan Turing Institute, UK. MMG and CJO were supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. TJS has been supported in part by the German Research Foundation (Deutsche Forschungsgemeinschaft) through project 415980428 and the Excellence Cluster

"MATH+ The Berlin Mathematics Research Centre" (EXC-2046/1, project 390685689). The authors thank Dennis Prangle for feedback on an earlier version of the manuscript.

Appendix A. Proof of Theoretical Results

This appendix contains proofs for all novel results in the main text. For $x \in \mathbb{R}^d$, we let $(-\infty, x] \coloneqq (-\infty, x_1] \times \cdots \times (-\infty, x_d]$ and we write $y \leq x$ whenever $y \in (-\infty, x]$, i.e. when $y_i \leq x_i$ for $i = 1, \ldots, d$.

A.1 Proof of Lemma 5

Our proof of Lemma 5 makes use of the *Kolmogorov distance*

$$d_{\mathrm{K}}(\mu, \nu) \coloneqq \sup_{x \in \mathbb{R}^d} d_{\mathrm{K}}(x; \mu, \nu), \tag{7}$$

$$d_{\mathrm{K}}(x; \mu, \nu) \coloneqq \left| \int 1_{(-\infty, x]} \, \mathrm{d}\mu - \int 1_{(-\infty, x]} \, \mathrm{d}\nu \right|, \tag{8}$$

which is a metric on $\mathcal{P}(\mathbb{R}^d)$ (Shorack, 2000, Theorem 2.4).

Proof [Proof of Lemma 5] Suppose that $\mu \neq \nu$, so that it suffices to exhibit an element $f \in \mathcal{F}_\Theta$ for which $\int f \, \mathrm{d}\mu \neq \int f \, \mathrm{d}\nu$. From the metric property of d_{K}, there must exist $x^* \in \mathbb{R}^d$ such that $\varepsilon \coloneqq d_{\mathrm{K}}(x^*; \mu, \nu) > 0$. Now, for $c > 0$, consider the function

$$f_{x^*}^{(c)} \colon \mathbb{R}^d \to (0, 1), \qquad\qquad f_{x^*}^{(c)}(x) \coloneqq \prod_{i=1}^{d} \frac{1}{1 + e^{2c(x_i - x_i^*)}}, \tag{9}$$

which satisfies $f_{x^*}^{(c)} \in \mathcal{F}_\Theta$. Since $f_{x^*}^{(c)}$ converges pointwise to $f_{x^*}^{(\infty)} \coloneqq 1_{(-\infty, x^*]}$ outside of a null set and $|f_{x^*}^{(c)}| \leq 1$, the dominated convergence theorem implies that $f_{x^*}^{(c)}$ is a consistent approximation of $f_{x^*}^{(\infty)}$ in the $c \to \infty$ limit in both $L^1(\mu)$ and $L^1(\nu)$. Therefore, there exists $c^* > 0$ such that $\|f_{x^*}^{(c^*)} - f_{x^*}^{(\infty)}\|_{L^1(\mu)} < \varepsilon/2$ and $\|f_{x^*}^{(c^*)} - f_{x^*}^{(\infty)}\|_{L^1(\nu)} < \varepsilon/2$. For this $f_{x^*}^{(c^*)} \in \mathcal{F}_\Theta$ we have from the reverse triangle inequality that

$$\begin{aligned}
\left| \int f_{x^*}^{(c^*)} \, \mathrm{d}\mu - \int f_{x^*}^{(c^*)} \, \mathrm{d}\nu \right| &= \left| \left(\int f_{x^*}^{(c^*)} \, \mathrm{d}\mu - \int f_{x^*}^{(\infty)} \, \mathrm{d}\mu \right) + \left(\int f_{x^*}^{(\infty)} \, \mathrm{d}\mu - \int f_{x^*}^{(\infty)} \mathrm{d}\nu \right) \right. \\
&\qquad \left. + \left(\int f_{x^*}^{(\infty)} \, \mathrm{d}\nu - \int f_{x^*}^{(c^*)} \, \mathrm{d}\nu \right) \right| \\
&\geq \Bigg| \underbrace{\left| \left(\int f_{x^*}^{(c^*)} \, \mathrm{d}\mu - \int f_{x^*}^{(\infty)} \, \mathrm{d}\mu \right) + \left(\int f_{x^*}^{(\infty)} \, \mathrm{d}\nu - \int f_{x^*}^{(c^*)} \, \mathrm{d}\nu \right) \right|}_{(*)} \\
&\qquad - \underbrace{\left| \int f_{x^*}^{(\infty)} \, \mathrm{d}\mu - \int f_{x^*}^{(\infty)} \, \mathrm{d}\nu \right|}_{= \varepsilon} \Bigg|.
\end{aligned}$$

The triangle inequality implies that

$$|(*)| \leq \|f_{x^*}^{(c^*)} - f_{x^*}^{(\infty)}\|_{L^1(\mu)} + \|f_{x^*}^{(\infty)} - f_{x^*}^{(c^*)}\|_{L^1(\nu)} < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

and so it follows that $\left|\int f_{x^*}^{(c^*)}\,\mathrm{d}\mu - \int f_{x^*}^{(c^*)}\mathrm{d}\nu\right| \neq 0$. Thus we have exhibited an element $f_{x^*}^{(c^*)} \in \mathcal{F}_\Theta$ for which $\int f_{x^*}^{(c^*)}\,\mathrm{d}\mu \neq \int f_{x^*}^{(c^*)}\,\mathrm{d}\nu$. This completes the proof. ■

A.2 Proof of Lemma 15

First we derive a corollary of Lemma 5 that will be used to prove Lemma 15:

Corollary 16. *Let $\Theta = \mathbb{R}^d$ for some $d \in \mathbb{N}$. Suppose that $\mu, \nu \in \mathcal{P}_r(\Theta)$ and that the independent random variables $\theta \sim \mu$, $\vartheta \sim \nu$ satisfy $\mathbb{P}(f(\theta) \leq f(\vartheta)) = 1/2$ for all $f \in \mathcal{F}_\Theta$. Then $\mu = \nu$.*

Proof If $\mu \neq \nu$ then, as in the proof of Lemma 5, we can identify $x^* \in \mathbb{R}^d$ such that $d_{\mathrm{K}}(x^*; \mu, \nu) > 0$. Since μ and ν are regular, the function $x \mapsto d_{\mathrm{K}}(x; \mu, \nu)$ is continuous on $\mathbb{R}^d$ and there exists an open neighbourhood $N(x^*)$ of x^* such that $d_{\mathrm{K}}(x; \mu, \nu) > 0$ for all $x \in N(x^*)$.

Suppose, to arrive at a contradiction, that $\mathbb{P}(f(\theta) \leq f(\vartheta)) = 1/2$ for all $f \in \mathcal{F}_\Theta$. Then, for all $x \in N(x^*)$, we can construct functions $f_x^{(c)} \in \mathcal{F}_\Theta$ as per (9), for which it holds that

$$\mathbb{P}(\vartheta \leq x) = \mathbb{P}(f_x^{(\infty)}(\theta) \leq f_x^{(\infty)}(\vartheta)) = \lim_{c\to\infty} \mathbb{P}(f_x^{(c)}(\theta) \leq f_x^{(c)}(\vartheta)) = \frac{1}{2}.$$

But ν was assumed to be regular, meaning that ν has a positive Lebesgue PDF, so that $\mathbb{P}(\vartheta \leq x) = 1/2$ cannot simultaneously hold for all $x \in N(x^*)$. Indeed, since $N(x^*)$ is open, there exists $x \in N(x^*)$ such that $x_i^* < x_i$ for all $i = 1, \ldots, d$. Then $\mathbb{P}(\vartheta \leq x) = \mathbb{P}(\vartheta \leq x^*) + \nu(S)$, where $S := (-\infty, x] \setminus (-\infty, x^*]$ is a measurable set with $\nu(S) > 0$. This contradiction completes the proof. ■

Proof [Proof of Lemma 15] Fix $\mu_0 \in B$. Let $\theta \sim \mu_0$, $y|\theta \sim P_\theta$ and $\vartheta|\theta, y \sim \mu(\mu_0, y)$. First we argue that the distribution $\nu := \iint \mu(\mu_0, y)\,\mathrm{d}P_\theta(y)\,\mathrm{d}\mu_0(\theta)$ of the random variable ϑ is regular. Since μ is a regular learning procedure, $\mu(\mu_0, y)$ admits a PDF $p_{\mu(\mu_0,y)}$ for each $y \in \mathbb{R}^d$. Thus, ν admits the PDF

$$p_\nu(x) := \int p_{\mu(\mu_0,y)}(x)\,\mathrm{d}Q(y), \qquad Q := \int P_\theta\,\mathrm{d}\mu(\theta)$$

and our task is to establish that this PDF is positive on $\mathbb{R}^d$. Fix $x \in \mathbb{R}^d$. Now, since $p_{\mu(\mu_0,y)}(x) > 0$ for all $y \in \mathbb{R}^d$, we have

$$\mathbb{R}^d = \bigcup_{n\in\mathbb{N}} S_n, \qquad S_n := \left\{ y \in \mathbb{R}^d \,\middle|\, p_{\mu(\mu_0,y)}(x) > \frac{1}{n} \right\}.$$

Since Q is a probability distribution on $\mathbb{R}^d$, it follows that for some $n \in \mathbb{N}$, $Q(S_n) > 0$. Therefore

$$p_\nu(x) = \int p_{\mu(\mu_0,y)}(x)\,\mathrm{d}Q(y) > \frac{1}{n} Q(S_n) > 0$$

and, since this argument holds for all $x \in \mathbb{R}^d$, p_ν is a positive PDF on $\mathbb{R}^d$ and ν is regular.

Next, since μ_0 and the learning procedure μ are regular, and μ is strongly calibrated, for each $f \in \mathcal{F}_\Theta$,

$$F_{f_\#\mu(\mu_0,y)}(f(\theta)) = \mathbb{P}(f(\vartheta) \leq f(\theta)|\theta, y) \sim \mathcal{U}(0,1) \tag{10}$$

and taking expectations of both sides yields

$$\mathbb{P}(f(\vartheta) \leq f(\theta)) = \frac{1}{2}. \tag{11}$$

Since both μ_0 and ν are regular, it follows from Corollary 16 and (11) that $\mu_0 = \nu$, and so ϑ has the marginal distribution μ_0. Thus we have shown that the learning procedure μ is weakly calibrated to the belief distribution μ_0 and the data-generating model P. ■

Appendix B. Probabilistic Numerical Methods for ODEs

This appendix contains full details of how the PNMs in Section 3.3 were implemented:

- The code for Chkrebtii et al. (2016) was taken from `git.io/J33lL` and the step-size was set at $h = 0.1$. The following settings were used: `nsolves` $= 100$, `N` $= 100$, `nevalpoints` $= 500$, `lambda` $= 0.08$ and `alpha` $= 1$. These values were manually selected, over the default values recommended in the code, since they led to improved calibration of the output. Rigorous optimisation of these settings was not attempted.

- The code for Teymur et al. (2018) was provided to us by the authors and is not yet publicly released. The method used is the 2-step (i.e. order 3) probabilistic Adams–Moulton method with step-size $h = 0.5$ and overall scaling parameter $\alpha = 0.3$. These values were manually selected with the intention of improving calibration of the output, but rigorous optimisation of these settings was not attempted. The stepwise perturbations are scaled using the global calibration procedure described in Conrad et al. (2017).

- The code for both Schober et al. (2019) and Tronarp et al. (2019) derives from the comprehensive open-source Python package `probnum`. On the advice of the authors of this package we implemented the adaptive routine `probnum/diffeq.probsolve_ivp`. In this case the default values of tolerances were used. The only hyperparameter it is required to set is `algo_order`, which we set to `3`. The setting `method` $=$ `EK0` corresponds to Schober et al. (2019), and `method` $=$ `EK1` corresponds to Tronarp et al. (2019).

- The code for Teymur et al. (2021) was provided to us by the authors and expected to be made public on full publication of that paper. This method is based on multi-fidelity simulation, so we take $h \in \{0.1, 0.2, 0.4\}$ and solve the ODE using a 2-step (i.e. order 2) Adams–Bashforth method. All other hyperparameters are optimised automatically as part of the routine.

References

Jeffrey L Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, 9(7):1518–1530, 1996. doi: 10.1175/1520-0442(1996)009⟨1518:AMFPAE⟩2.0.CO;2.

Yaakov Bar-Shalom and K. Birmiwal. Consistency and robustness of PDAF for target tracking in cluttered environments. *Automatica*, 19(4):431–437, 1983. doi: 10.1016/0005-1098(83)90059-6.

Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

James O. Berger. An overview of robust Bayesian analysis: With comments and a rejoinder by the author. *Test*, 3(1):5–124, 1994. doi: 10.1007/BF02562676.

Jeremy Berkowitz. Testing density forecasts, with applications to risk management. *J. Bus. Econom. Statist.*, 19(4):465–474, 2001. doi: 10.1198/07350010152596718.

Jose-M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1994. doi: 10.1002/9780470316870.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *Ann. Statist.*, 47(1):39–66, 2019. doi: 10.1214/18-AOS1712.

Joris Bierkens. Non-reversible Metropolis–Hastings. *Stat. Comput.*, 26(6):1213–1228, 2016. doi: 10.1007/s11222-015-9598-x.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

George Casella. An introduction to empirical Bayes data analysis. *Amer. Statist.*, 39(2):83–87, 1985. doi: 10.2307/2682801.

Oksana A. Chkrebtii, David A. Campbell, Ben Calderhead, and Mark A. Girolami. Bayesian solution uncertainty quantification for differential equations. *Bayesian Anal.*, 11(4):1239–1267, 2016. doi: 10.1214/16-BA1017.

Peter F. Christoffersen. Evaluating interval forecasts. *Int. Econ. Rev.*, 39(4):841–862, 1998. doi: 10.2307/2527341.

Jon Cockayne, Chris J. Oates, Ilse C. F. Ipsen, and Tim Reid. Probabilistic iterative methods for linear systems. *Journal of Machine Learning Research*, 2021. To appear. arXiv:2012.12615.

Patrick R. Conrad, Mark Girolami, Simo Särkkä, Andrew Stuart, and Konstantinos Zygalakis. Statistical analysis of differential equations: Introducing probability measures on numerical solutions. *Stat. Comput.*, 27(4):1065–1082, 2017. doi: 10.1007/s11222-016-9671-0.

Samantha R. Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Statist.*, 15(3):675–692, 2006. doi: 10.1198/106186006X136976.

Dennis D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923, 1993. doi: 10.1214/aos/1176349157.

A. P. Dawid. The well-calibrated Bayesian. *J. Amer. Statist. Assoc.*, 77(379):605–603, 1982. doi: 10.1080/01621459.1982.10477856.

A. P. Dawid. Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A*, 147 (2):278–292, 1984. doi: 10.2307/2981683.

Persi Diaconis, Susan Holmes, and Mehrdad Shahshahani. Sampling from a manifold. In *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, volume 10 of *Inst. Math. Stat. (IMS) Collect.*, pages 102–125. Inst. Math. Statist., Beachwood, OH, 2013.

Francis X. Diebold, Todd A. Gunther, and Anthony S. Tay. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.*, 39(4):863–883, 1997. doi: 10.2307/2527342.

Oliver E. Drummond, X. Rong Li, and Chen He. Comparison of various static multiple-model estimation algorithms. In *Signal and Data Processing of Small Targets 1998*, volume 3373, pages 510–527. International Society for Optics and Photonics, 1998.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(3):419–474, 2012. doi: 10.1111/j.1467-9868.2011.01010.x.

David Freedman. On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1140, 1999. doi: 10.1214/aos/1017938917.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow, 2020. arXiv:2011.01808.

John Geweke. Getting it right: Joint distribution tests of posterior simulators. *J. Amer. Statist. Assoc.*, 99(467):799–804, 2004. doi: 10.1198/016214504000001132.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, 1(1):125–151, 2014. doi: 10.1146/annurev-statistics-062713-085831.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electron. J. Stat.*, 7:1747–1782, 2013. doi: 10.1214/13-EJS823.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(25):723–773, 2012. URL `https://www.jmlr.org/papers/volume13/gretton12a/gretton12a.pdf`.

Peter Grünwald and Thijs van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4), December 2017. doi: 10.1214/17-ba1085. URL `https://doi.org/10.1214/17-ba1085`.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.*, 129(3):550–560, 2001. doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2.

Thomas M. Hamill and Stephen J. Colucci. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.*, 125(6):1312–1327, 1997. doi: 10.1175/1520-0493(1997)125⟨1312:VOERSR⟩2.0.CO;2.

Diana Harrison, David Sutton, Pedro Carvalho, and Michael Hobson. Validation of Bayesian posterior distributions using a multidimensional Kolmogorov–Smirnov test. *Mon. Not. R. Astron. Soc.*, 451(3):2610–2624, 2015. doi: 10.1093/mnras/stv1110.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. doi: 10.1002/9780470434697.

Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proc. IEEE*, 70(9):939–952, 1982. doi: 10.1109/PROC.1982.12425.

Toni Karvonen, George Wynne, Filip Tronarp, Chris Oates, and Simo Särkkä. Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA J. Uncertain. Quantif.*, 8(3):926–958, 2020. doi: 10.1137/20M1315968.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *Journal of Machine Learning Research*, 2021. To appear.

Kody Law, Andrew Stuart, and Konstantinos Zygalakis. *Data Assimilation: A Mathematical Introduction*, volume 62 of *Texts in Applied Mathematics*. Springer, Cham, 2015. doi: 10.1007/978-3-319-20325-6.

Jeong Eun Lee, Geoff K. Nicholls, and Robin J. Ryder. Calibration procedures for approximate Bayesian credible sets. *Bayesian Anal.*, 14(4):1245–1269, 2019. doi: 10.1214/19-BA1175.

John R. Lewis, Steven N. MacEachern, and Yoonkyung Lee. Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression (with discussion). *Bayesian Analysis*, 16(4), December 2021. doi: 10.1214/21-ba1257. URL `https://doi.org/10.1214/21-ba1257`.

X. Rong Li and Zhanlue Zhao. Measuring estimator's credibility: Noncredibility index. In *Proceedings of the 9th International Conference on Information Fusion*, pages 1–8. IEEE, 2006. doi: 10.1109/ICIF.2006.301770.

X. Rong Li, Zhanlue Zhao, and Vesselin P. Jilkov. Estimator's credibility and its measures. In *Proceedings of the IFAC 15th World Congress*, 2002.

X. Rong Li, Zhanlue Zhao, and Xiao-Bai Li. Evaluation of estimation algorithms: Credibility tests. *IEEE T. Syst. Man Cy. A*, 42(1):147–163, 2011. doi: 10.1109/TSMCA.2011.2158095.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, second edition, 2009. doi: 10.1017/CBO9780511626630.

John F. Monahan and Dennis D. Boos. Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278, 1992. doi: 10.1093/biomet/79.2.271.

D. Prangle, M. G. B. Blum, G. Popovic, and S. A. Sisson. Diagnostic tools for approximate Bayesian computation using the coverage property. *Aust. N. Z. J. Stat.*, 56(4):309–329, 2014. doi: 10.1111/anzs.12087.

Leah F. Price, Christopher C. Drovandi, Anthony Lee, and David J. Nott. Bayesian synthetic likelihood. *J. Comput. Graph. Statist.*, 27(1):1–11, 2018. doi: 10.1080/10618600.2017.1302882.

Jakub Prüher, Toni Karvonen, Christopher James Oates, Ondřej Straka, and Simo Särkkä. Improved calibration of numerical integration error in sigma-point filters. *IEEE Trans. Automat. Contr.*, 66(3):1286–1292, 2020. doi: 10.1109/TAC.2020.2991698.

G. D. Rayner and H. L. MacGillivray. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Stat. Comput.*, 12(1):57–75, 2002. doi: 10.1023/A:1013120305780.

Joseph P. Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, Department of Statistics, Stanford University, 2019. URL `https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/2019-03.pdf`. Technical Report No. 2019-03.

Murray Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statistics*, 23: 470–472, 1952. doi: 10.1214/aoms/1177729394.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003. doi: 10.1137/1.9780898718003.

Michael Schober, Simo Särkkä, and Philipp Hennig. A probabilistic model for the numerical solution of initial value problems. *Stat. Comput.*, 29(1):99–122, 2019. doi: 10.1007/s11222-017-9798-7.

Galen R. Shorack. *Probability for Statisticians*. Springer Texts in Statistics. Springer-Verlag, New York, 2000.

Botond Szabó, A. W. van der Vaart, and J. H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428, 2015. doi: 10.1214/14-AOS1270.

O. Talagrand, R. Vautard, and B. Strauss. Evaluation of probabilistic prediction systems. In *Proceedings of the ECMWF Workshop on Predictability*, pages 1–25. ECMWF, 1997. URL `https://www.ecmwf.int/node/12555`.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration, 2018. arXiv:1804.06788.

Onur Teymur, Han Cheng Lie, T. J. Sullivan, and Ben Calderhead. Implicit probabilistic integrators for ODEs. *NeurIPS 31*, pages 7255–7264, 2018.

Onur Teymur, Chris N. Foley, Philip G. Breen, Toni Karvonen, and Chris J. Oates. Black-box probabilistic numerics, 2021. arXiv:2106.13718.

Filip Tronarp, Hans Kersting, Simo Särkkä, and Philipp Hennig. Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: A new perspective. *Stat. Comput.*, 29(6):1297–1315, 2019. doi: 10.1007/s11222-019-09900-1.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.*, 114(527):1147–1161, 2019. doi: 10.1080/01621459.2018.1473776.

Daniel Wegmann, Christoph Leuenberger, and Laurent Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009. doi: 10.1534/genetics.109.102509.

Hanwen Xing, Geoff Nicholls, and Jeong Lee. Calibrated approximate Bayesian inference. In *International Conference on Machine Learning*, pages 6912–6920, 2019.