Fast and Scalable Score-Based Calibration Tests

Abstract

We propose the kernel calibration-conditional Stein discrepancy (KCCSD), a new method to estimate and test calibration of probabilistic models. The method scales linearly with the number of data points, and its test statistics are differentiable and hence can be used as regularizer to obtain better calibrated models. Moreover, we propose the measure-transport based kernel, a family of score-based kernels on probability distributions. We employ it in the formulation of the KCCSD to obtain efficient calibration tests for models for which (only) the score-function of the predictions is readily available, including energy-based models and many unnormalized density models in Bayesian inference. Unlike prior works, our method does not require potentially expensive sampling for calibration evaluation. We apply our method to train and test the calibration of simulation-based inference models, where calibration is crucial for accurate and robust scientific discoveries.

1 INTRODUCTION

Calibration is a statistical property of predictive probabilistic models that ensures that a model's prediction matches the conditional distribution of the predicted variable given the prediction. A calibrated model expresses the uncertainty about its predictions reliably by being neither over- nor underconfident, and hence can be useful even if its accuracy is suboptimal. In safety-critical applications such as autonomous driving it is crucial that predictive models involved in automatic decision making are neither too under- nor, often more importantly, overconfident. Unfortunately, empirical studies revealed that popular machine learning models such as deep neural networks tend to trade off calibration for accuracy **?**. This has lead to an increased interest in the study of calibrated models in recent years.

Calibration has been studied in the metereological and statistical literature for many decades (e.g., **??**). For a long time research on calibration has been focused on different notions of calibration for probabilistic classifiers (e.g., **????????????**) and on calibration of quantiles and confidence intervals for real-valued regression problems (e.g., **?????**).

In recent years, other communities have praised the importance of properties bearing similarities with calibration. One instance of this is in simulation-based inference (SBI), in which probabilistic models are trained to estimate the posterior of scientific parameters of interest given some observed data. These posteriors are continuously-valued and hence many existing notions and evaluations of calibration are not applicable. Therefore the SBI community instead studied the conservativeness of posterior models. However, the current methods available for assessing conservativeness are very expensive to run, as discussed by **?**. There is thus a clear need for fast and scalable methods that can be used to assess the calibration of continuously-valued predictive models.

In parallel to advances in SBI, multiple works (see, e.g., **??**) generalized the notion of calibration introduced for probabilistic classifier to continuous-valued predictive models. In addition, **?** introduced a kernel-based hypothesis test that outputs, given a predictive model and a validation dataset, whether the model is calibrated. However, evaluating their test statistic involves computing expectations with respect to the predictions of the model of interest. These computations can be performed exactly for classification models but in general analytical expressions are available only for specific choices of kernels and models. At the expense of increasing the variance of the test statistic, one can use unbiased Monte Carlo estimates of the expectations if one can sample from the probabilistic predictions. Sampling is possible, for instance, if the predictions are modelled by a normalizing flow or variational autoencoder. However, we argue that in many interesting settings the probabilistic model is intractable and cannot be sampled from in a straightforward way. For instance, in Bayesian inference, and in particular in SBI models like SNRE **?**, SNLE **?** and many of their variants, probabilistic predictions are posterior models over a typically multidimensional target variable (e.g., the parameter) which are often unnormalized and require Markov-chain Monte Carlo (MCMC) methods for approximating expectations. Hence applying the method of **?** to test for calibration would require running an MCMC algorithm for every data sample used for testing, which is prohibitively expensive.

Contributions In this paper, we introduce the kernel calibration-conditional Stein discrepancy (KCCSD), a new score-based test for calibration which addresses computational limitations of existing methods.

- We first identify that calibration can be understood as a conditional goodness-of-fit problem. This allows us to depart from the formulation of **?** and instead leverage statistical tools from the conditional goodness of fit literature investigated by **?** to formulate a new test for calibration based on the KCCSD, which comes with an incomplete U-statistic variant that scales linearly with the number of samples.
- Second, we focus on the design of kernels to compare unnormalized probability distributions, objects which appear in the formulation of our test. Such kernels were studied in many works including **?**, but all require access to empirical estimates of the expectations which we assume are expensive to obtain in our setting. We propose the measure-transport based kernel, a family of kernels between probability distributions which can be computed using the score-function of the (unnormalized) densities only, without having access to empirical probabilities. By considering appropriate special cases, we show this family recovers more traditional kernels as limiting regimes.
- Combining our calibration test with our new kernel, we obtain a fast scalable calibration test that can be used for a wide class of probabilistic models. This test can be used to assess the calibration of a model after it was trained, but also as a differentiable regularizer to train a calibrated model. We demonstrate the properties of our test on synthetic examples, and apply it to problems in simulation-based inference, which has become an important consumer of calibration methods in recent years (**??**).

2 BACKGROUND

Notation We consider probabilistic systems characterized by a joint distribution $\mathbb{P}(X, Y)$ of random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$, and study *probabilistic models* $P_{|\cdot} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ to approximate the unknown conditional probability of Y given $X = x$: $P_{|x}(\cdot) \simeq \mathbb{P}(Y \in \cdot \mid X = x)$. Target variable Y is typically a parameter of a probabilistic system of interest—like synapses in biological neural networks—while input variable X is observed data—like neuron voltage traces measured using electrophysiology.

2.1 CALIBRATION OF PREDICTIVE MODELS

Calibration: General Definition A probabilistic model $P_{|\cdot}$ is called calibrated or reliable (**???**) if it satisfies

$$P_{|X} = \mathbb{P}\left(Y \in \cdot \mid P_{|X}\right) \qquad \mathbb{P}(X)\text{-a.s.}. \tag{1}$$

Note that this definition applies to general predictive probabilistic models, also beyond classification, and only assumes that the conditional distributions on the left-hand side exist.

Hypothesis Testing: Kernel Calibration Error There are multiple ways to study the calibration of predictive models. In this section, we introduce the approach of **?** and its later generalization (**?**), on top of which CCKSD is built and on which we focus. These works introduce kernel-based tests to assess whether a predictive model is calibrated. These tests turn the equality between conditional-distribution present in **??** into a more classical equality between joint distributions. The transformation is achieved by noting that

$$\begin{aligned} P_{|X} = \mathbb{P}\left(Y \in \cdot \mid P_{|X}\right) \quad \mathbb{P}(X)\text{-a.s.} \\ \iff (P_{|X}, Y) \overset{d}{=} (P_{|X}, Z) \end{aligned} \tag{2}$$

where Z is an "auxiliary" variable such that $Z \mid P_{|X} \sim P_{|X}(\cdot)$. This identity between probability distribution was used by **?** to construct an MMD-type calibration test based on the statistic

$$\sup_{h \in \mathcal{B}(0_{\mathcal{H}}, 1)} \mathbb{E}_{(x,y,z) \sim \mathbb{P}(X,Y,Z)} \left[h(P_{|x}, y) - h(P_{|x}, z)\right], \tag{3}$$

called the (squared) kernel calibration error (SKCE). Here, $\mathcal{B}(0_{\mathcal{H}}, 1)$ is the unit ball of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions with positive definite kernel $k_{\mathcal{H}} : (P_{|\mathcal{X}} \times \mathcal{Y})^2 \to \mathbb{R}$. As noted by **?**, this formulation generalizes the notion of (squared) kernel classification calibration error (SKCCE) defined for the special case of discrete output spaces $\mathcal{Y} = \{1, \ldots, d\}$ only (**?**). **?** then constructed a test that estimates the SKCE based on n pairs of samples $\{(P_{|x^i}, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$:

$$\begin{aligned} \frac{1}{n(n-1)} \sum_{i \neq j} \Big\{ & k((P_{|x^i}, y^i), (P_{|x^j}, y^j)) \\ & - \mathbb{E}_{z^i \sim P_{|x^i}} k((P_{|x^i}, z^i), (P_{|x^j}, y^j))) \\ & - \mathbb{E}_{z^j \sim P_{|x^j}} k((P_{|x^i}, y^i), (P_{|x^j}, z^j))) \\ & + \mathbb{E}_{z^i \sim P_{|x^i}} \mathbb{E}_{z^j \sim P_{|x^j}} k((P_{|x^i}, z^i), (P_{|x^j}, z^j)) \Big\}. \end{aligned} \tag{4}$$

As discussed above and by **?**, **??** alludes at two important sources of possible intractability:

First Problem The last three terms in the sum are expectations under predictions of the probabilistic model of interest. However, closed-form expressions for these expectations are only available in restricted cases, such as for classification and for Gaussian models coupled with Gaussian kernels. When these expectations are not available, they must be approximated numerically. If the distributions $P_{|X}$ are given in the form of unnormalized models, this approximation requires running expensive approximation methods that often take the form of an MCMC algorithm and must be performed for every sample of $P_{|X}$ used to estimate the test statistic.

Second Problem The second source is the choice and evaluation of the kernel function k. We restrict our attention to the conventional form of tensor-product type kernels $k((p, y), (p', y')) = k_P(p, p')k_Y(y, y')$ chosen in this setting. While typically many tractable choices for the kernel k_Y exist (taking as input discrete or Euclidean values), the choices for k_P, defined for two probability distribution p and p', are more limited and require expensive approximations methods when working with unnormalized models.

A popular approach to design kernels on distributions (**??**) is to first embed the probability distributions in an Hilbert space $\mathcal{H}$ using a map ϕ, and then compose it with a kernel $k_{\mathcal{H}}$ on $\mathcal{H}$:

$$k_P(p, p') = k_{\mathcal{H}}(\phi(p), \phi(p')).$$

Any valid kernel on $\mathcal{H}$, like the linear kernel $k_{\mathcal{H}}(z, z') = \langle z, z'\rangle_{\mathcal{H}}$, the Gaussian kernel $k_{\mathcal{H}}(z, z') = e^{-\|z-z'\|^2_{\mathcal{H}}}$, or the multiquadric kernel $k_{\mathcal{H}}(z, z') = (1+\|z - z'\|^2_{\mathcal{H}})^{-1}$ can be used. In practice, the map ϕ can be set to be the *mean embedding* map to an RKHS $\mathcal{H}$, e.g., $\phi(\mu) = \int k_{\mathcal{H}}(z, \cdot)\, \mu(\mathrm{d}z)$. Kernels $k_{\mathcal{H}}$ that are functions of $\|\phi(\mu) - \phi(\nu)\|^2_{\mathcal{H}} := \mathrm{MMD}^2(\mu, \nu)$, are often referred to as MMD-type kernels **?**. Other distances, like the Wasserstein distance in 1 dimension or the sliced Wasserstein distance (**?**) in multiple dimensions, also take this form for some choice of ϕ and $\mathcal{H}$, and can thus be used to construct kernels on distributions **?**. In general, however, estimating $k_P(p, p')$ becomes intractable apart from special cases such as when p and p' are Gaussian distributions. **While there exist finite-samples estimators for such kernels, a fast calibration estimation method based on ?? would require an estimator that does not require samples from p and p'.**

2.2 CALIBRATION BEYOND CLASSIFICATION: APPLICATION IN BAYESIAN INFERENCE

Bayesian Inference and SBI One main motivation for studying calibration of generic probabilistic models is Bayesian inference, which seeks to compute the (approximate) posterior distribution of a parameter y of interest given some observed variable x. These posterior distributions are special instances of probabilistic models $P_{|x}$, depending on the observed data x, and thus calibration tests can be used to assess the calibration of these Bayesian models. In the case of SBI, a family of Bayesian inference approaches particularly popular in scientific domains such as neuroscience or physics, it is particularly important that the posterior estimate exhibits reliability properties as non-reliable models can cause incorrect scientific conclusions. As shown by **?** such reliability properties are not always verified by the approximate posteriors returned by SBI methods. Unfortunately, the reliability evaluation process is complicated by the fact current reliability estimation methods are very costly to run **?**. Thus it is crucial to provide the Bayesian inference community with practical and fast reliability estimation methods.

Reliability in Bayesian Inference versus Calibration It is important to note that reliability metrics traditionally used in Bayesian inference such as posterior coverage **?** differ from the notion of calibration in **??**. However, in **??** we show that a probabilistic model that is calibrated according to **??** is also reliable in the sense of **?**, grounding the use of our tests in Bayesian inference.

2.3 KERNEL CONDITIONAL GOODNESS-OF-FIT TEST

Conditional goodness-of-fit (or CGOF) testing tests whether

$$H_0 \colon P_{|X} = \mathbb{P}(Y \in \cdot \mid X) \qquad \mathbb{P}(X)\text{-a.s.} \tag{5}$$

given a model $P_{|x}$ for the conditional distribution $\mathbb{P}(Y \in \cdot \mid X = x)$ and samples $\{(x^i, y^i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathbb{P}(X, Y)$. This problem was studied by **?** for the case $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ and models $P_{|x}$ with a differentiable, strictly positive density $f_{P_{|x}}$. They proposed a kernel CGOF test for **??** based on the (squared) kernel conditional Stein discrepancy (KCSD)

$$D_{P_{|\cdot}}(\mathbb{P}) := \left\|\mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)}\left[K_x \xi_{P_{|x}}(y, \cdot)\right]\right\|^2_{\mathcal{F}_K} \tag{6}$$

where

- $\mathcal{F}_l$ is an RKHS on $\mathcal{Y}$ with kernel $l \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and $\mathcal{F}_l^{d_y} = \otimes_{d_y} \mathcal{F}_l$,
- $\xi_{P_{|x}}$ is the "kernelized score"

  $$\xi_{P_{|x}}(y, \cdot) = l(y, \cdot)\nabla_y \log f_{P_{|x}}(y) + \nabla_y l(y, \cdot) \in \mathcal{F}_l^{d_y},$$

- $\mathcal{F}_K$ is an $F_l^{d_y}$-vector-valued RKHS with kernel $K \colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F}_l^{d_y}, \mathcal{F}_l^{d_y})$, and
- K_x is its associated linear operator on $\mathcal{Y}$:

  $$\begin{aligned} K_x \colon \mathcal{Y} &\longrightarrow \mathcal{L}(\mathcal{X}, \mathcal{Y}), \\ y &\longmapsto K_x y = K(x, \cdot)y. \end{aligned}$$

Under certain assumptions (**?**, Theorem 1), such as kernels K and l being C_0-universal, the null hypothesis in **??** is true if and only if $D_{P_{|\cdot}}(\mathbb{P}) = 0$. An example of an $\mathcal{F}_l^{d_y}$-reproducing C_0-universal kernel K is

$$K(x, x') = k(x, x') I_{\mathcal{F}_l^{d_y}} \tag{7}$$

where $I_{\mathcal{F}_l^{d_y}} \in \mathcal{L}(\mathcal{F}_l^{d_y}, \mathcal{F}_l^{d_y})$ is the identity operator and k is a real-valued C_0-universal kernel (**?**). **?** showed that the CGOF statistic $D_{P_{|\cdot}}(\mathbb{P})$ admits an unbiased consistent estimator and used it to construct hypothesis tests of **??** with operator-valued kernels of the form in **??**.

3 KERNEL CALIBRATION-CONDITIONAL STEIN DISCREPANCY

Calibration testing in the sense of **??** is an instance of *conditional goodness-of-fit* testing of **??** with input $P_{|X}$, target Y, and models $P_{|x} \mapsto P_{|x}$. Assuming that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ and that distributions $P_{|x}$ have a differentiable, strictly positive density $f_{P_{|x}}$, thus the (squared) kernel conditional Stein discrepancy in **??** becomes

$$C_{P_{|\cdot}}(\mathbb{P}) := \left\| \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \left[K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right] \right\|_{\mathcal{F}_K}^2, \tag{8}$$

where now K is a kernel on $P_{|\mathcal{X}}$. To emphasize the calibration setting, we call $C_{P_{|\cdot}}$ the kernel calibration-conditional Stein discrepancy (KCCSD). Similar to the KCSD, given samples $\{P_{|x^i}, y^i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$ and assuming a kernel K of the form in **??**, statistic $C_{P_{|\cdot}}(\mathbb{P})$ has an unbiased consistent estimator

$$\widehat{C_{P_{|\cdot}}} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H((P_{|x^i}, y^i), (P_{|x^j}, y^j))$$

where

$$H((p, y), (p', y')) := k(p, p') h((p, y), (p', y')) \tag{9}$$

with

$$\begin{aligned} h((p, y), (p', y')) &:= l(y, y') s_p(y)^\top s_{p'}(y') \\ &+ \sum_{i=1}^{d_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(y, y') + s_p(y)^\top \nabla_{y'} l(y, y') \\ &+ s_{p'}(y')^\top \nabla_y l(y, y'), \end{aligned} \tag{10}$$

where $s_p(y) := \nabla_y \log f_p(y)$ (resp. $s_{p'}(y)$) is the *score* of p (resp. p'). In Section A in the supplement we discuss how the formula of $\widehat{C}_{P_{|\cdot}}$ generalizes to operator-valued kernels that are not of the form in **??**.

The above framing of the calibration problem conveniently avoids the first source of possible intractability present in the SKCE. For instance, for Gaussian models the test statistic can be evaluated exactly for arbitrary kernels l on $\mathcal{Y}$ whereas a closed-form expression of the SKCE is known only in the special case where l is a Gaussian kernel.

?? shows that the KCCSD can be viewed as a special case of the SKCE. More generally, as shown in Section B, the KCSD is a special form of the MMD.

Proposition 3.1 (Special case of Lemma B.1)**.** *Under weak assumptions (see Lemma B.1), the KCCSD with respect to kernels* $l \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *and* $k \colon P_{|\mathcal{X}} \times P_{|\mathcal{X}} \to \mathbb{R}$ *is equivalent to the SKCE with kernel* $H \colon (P_{|\mathcal{X}} \times \mathcal{Y}) \times (P_{|\mathcal{X}} \times \mathcal{Y}) \to \mathbb{R}$ *defined in* **??**.

Linear Time Variants The minimum-variance estimator of this U-statistic requires $n(n-1)$ evaluations and thus scales quadratically with the sample size. Therefore in our tests we employ linear-time tests variants, based on incomplete U-statistics estimation **?** or on B-tests **?**. Indeed, sample sizes in typical Bayesian application are such that operations of with a complexity scaling quadratically with the numbers of samples are intractable on commodity hardware. These linearization methods are conceptually similar to their quadratic counterparts, and do not require solving an additional optimization problem, as the test points methods of **??**.

I think we need to do the theory if we actually use incomplete U-statistics linearization?

The full testing procedure is outlined in **??**. The computations can be performed with kernels K of the form in **??** or more general operator-valued kernels, but crucially the method requires that K is tractable. Thus for general models of probability distributions, such as energy-based models and other unnormalized density models, it remains to address the second source of intractability, namely to construct a kernel K that can be evaluated efficiently.

4 TRACTABLE KERNELS FOR GENERAL UNNORMALIZED DENSITIES

In this section, we introduce two kernels between (density-based) probability distributions that admit unbiased estimates that neither require samples from the said distributions nor require access their normalizing constant. Crucially, the properties of these new kernels allow to extend the scope of calibration tests to a larger setting, including Bayesian inference.

General Recipe As in prior work on kernels for distributions (**??**), our proposed kernels take the form of exponentiated Hilbertian metrics

$$k(p, q) = e^{-\|\phi(p) - \phi(q)\|_H^2 / (2\sigma^2)}$$

Algorithm 1: CGOF Calibration Test (Tractable Kernel)

```
Data: Pairs $\{(P_{|x^i}, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$
Result: Whether to reject $H_0$: "model is calibrated"
Parameters: Number of data samples $n$, kernel
        $l\colon \mathcal{Y}^2 \to \mathbb{R}$, kernel $k\colon (P_{|\mathcal{X}})^2 \to \mathbb{R}$, set
        of indices pairs $R \subset \{1, \dots, n\}^2$,
        significance level $\alpha$
/* Estimate KCCSD using ?? or (A.1)                          */
$\widehat{C} \leftarrow \frac{1}{|R|} \sum_{(i,j) \in R} H((P_{|x^i}, y^i), (P_{|x^j}, y^j))$
/* Use e.g. bootstrap (?)                                    */
$\widehat{C}_\alpha \leftarrow$ approximate $(1-\alpha)$-quantile of $\widehat{C}$
if $\widehat{C} < \widehat{C}_\alpha$ then
    return Fail to reject $H_0$
else
    return Reject $H_0$
end
```

where H is some Hilbert space, $\phi\colon p \mapsto \phi(p) \in H$ is a feature map, and σ is a bandwidth parameter. Our contributions in this section consist in pairs of carefully designed ϕ and H that will allow approximating k easily.

Tractable Hilbertian Distances between unnormalized densities Our starting point is the *Fisher Divergence* (???), also known as the *Relative Fisher Information* (?), between two probability densities p and q, which is given by

$$\mathrm{FD}(p, q) := \int_{\mathcal{X}} \|s_p(x) - s_q(x)\|^2 \, p(x) \, \mathrm{d}x.$$

The Fisher Divergence is a convenient tool to compare unnormalized densities of the form

$$p(x) := \frac{\overbrace{f(x)}^{\text{tractable}}}{\underbrace{Z_f}_{\text{intractable}}} \quad \text{where} \quad Z_f := \int_{\mathcal{X}} f(x) \, \mathrm{d}x$$

as the score of p can be evaluated without knowing Z_f:

$$s_p(x) = \nabla_x (\log f(x)/Z_f) = \nabla_x \log f(x).$$

This property confers the Fisher Divergence with a tractable unbiased estimator given n i.i.d. samples $\{X^i\}_{i=1}^n$ from p, which takes the form:

$$\widehat{\mathrm{FD}(p, q)} = \frac{1}{n} \sum_{i=1}^n \|s_p(X^i) - s_q(X^i)\|^2.$$

While the assumption ensuring access to samples from p is realistic in the unsupervised learning literature (?), or when dealing with special instances of unnormalized densities such as truncated densities $f(x) = p(x)\mathbf{1}_{x \in \mathcal{C}}$, it does not hold in the context of studying the calibration of unnormalized models, where the samples y^i are drawn from the unknown $p^\star(y|x^i)$, which may or may not equal the model $p(y|x^i)$. We overcome this issue by constructing a generalized version of the Fisher Divergence:

Definition 4.1 (Generalized Fisher Divergence)**.** Let p, q be two probability densities on $\mathcal{X}$, and ν a probability measure on $\mathcal{X}$. The *Generalized Fisher Divergence* between p and q is defined as

$$\mathrm{GFD}_\nu(p, q) := \int_{\mathcal{X}} \|s_p(x) - s_q(x)\|^2 \, \nu(\mathrm{d}x),$$

if $\mathbb{E}_\nu \|s_p\|^2, \mathbb{E}_\nu \|s_q\|^2 < +\infty$, and $+\infty$ otherwise.

The Generalized Fisher Divergence differs from the Fisher Divergence in that the integration is performed with respect to some given base measure ν instead of p. If the support of ν covers the support of p and q, then we have that $\mathrm{GFD}_\nu(p, q) = 0$ iff. $p = q$. Moreover, if ν can be sampled from in a tractable manner, then $\mathrm{GFD}_\nu(p, q)$ admits a tractable estimator given samples $\{Z^i\}_{i=1}^n$ from ν of the form

$$\widehat{\mathrm{GFD}}_\nu(p, q) = \frac{1}{n} \sum_{i=1}^n \|s_p(Z^i) - s_q(Z^i)\|^2.$$

In practice, the tractability assumption as well as the support assumption for any p, q are verified by setting ν to be a standard Gaussian distribution.

The Exponentiated-GFD Kernel Importantly, the (square root of the) Generalized Fisher Divergence is a Hilbertian metric on the space of probability densities. Indeed, for p, q such that $\mathbb{E}_\nu \|s_p\|^2, \mathbb{E}_\nu \|s_q\|^2 < +\infty$, we have that

$$\mathrm{GFD}_\nu(p, q) = \|\phi(p) - \phi(q)\|^2_{\mathcal{L}_2(\nu)}$$

where $\phi\colon p \mapsto s_p(\cdot) \in \mathcal{L}_2(\nu)$. The latter fact allows to construct a kernel K_ν on the space of probability densities based on the Generalized Fisher Divergence as follows:

Definition 4.2 (Exponentiated GFD Kernel)**.** Let p, q be two probability densities on $\mathcal{X}$, and ν a probability measure on $\mathcal{X}$. The *exponentiated GFD kernel* between p and q is defined as

$$K_\nu(p, q) := e^{-\mathrm{GFD}_\nu(p,q)/(2\sigma^2)}$$

Since the GFD is a Hilbertian metric, K_ν is positive definite (?), and can be estimated given samples of ν by replacing GFD_ν with its empirical counterpart. We summarize the computation method for K_ν in **??**.

Note the difference in the estimation error of such kernels compared to traditional MMD-type kernels of the form

$$K_{\mathrm{MMD}}(p, q) = e^{-\mathrm{MMD}^2(p,q)/(2\sigma^2)}.$$

Algorithm 2: Exponentiated GFD Kernel

```
Data: Probability densities $p, q$ on $\mathcal{X}$
Result: Approx. $\widehat{K_\nu(p,q)}$ of $K_\nu(p,q)$ in ??
Parameters: Base measure $\nu$, num. of base samples $m$
for $i \leftarrow 1$ to $m$ do
    Draw $Z^i \sim \nu$
end
return $\exp\left(-\frac{1}{2m\sigma^2}\sum_{i=1}^m \|s_p(Z^i) - s_q(Z^i)\|^2\right)$
```

In the GFD case, the error arises from discretizing ν whereas in the second case the error is caused by approximating p and q by samples, i.e., by setting $\widehat{K_{\text{MMD}}(p,q)} := K_{\text{MMD}}(\widehat{p}, \widehat{q})$.

Kernelizing the Generalized Fisher Divergence While the recipe given above suffices to obtain a valid kernel on the space of probability densities, the approximation error arising from the discretization of the base measure ν may scale unfavorably with the dimension of the underlying space $\mathcal{X}$. To address this issue, it is possible to apply a kernel-smoothing step to the GFD feature map $\phi(p)$ by composing it with an integral operator $T_{K,\nu}$ associated with a $\mathcal{X}$-vector-valued kernel K and its RKHS $\mathcal{H}_K$

$$T_{K,\nu} \colon f \in \mathcal{L}(\mathcal{X}, \mathbb{R}^d) \longmapsto \int_{\mathcal{X}} K_x f(x)\, \nu(\mathrm{d}x) \in \mathcal{H}_K$$

and comparing the difference in feature map using the squared RKHS norm $\|\cdot\|^2_{\mathcal{H}_K}$. This choice of feature map yields another metric, which we call the "kernelized" GFD:

$$\text{KGFD}(p,q) := \|T_{K,\nu}s_p - T_{k,\nu}s_q\|^2_{\mathcal{H}_K}\,.$$

It admits the following sample-based estimator:

$$\frac{1}{m^2}\sum_{i,j=1}^m \left\langle K(Z^i, Z^j)(s_p - s_q)(Z^i), (s_p - s_q)(Z^j)\right\rangle_{\mathcal{X}}\,.$$

The kernelized GFD is also a Hilbertian metric. Moreover, for characteristic kernels K, the integral operator $T_{K,\nu}$ is a Hilbertian isometry between $\mathcal{L}_2(\nu)^{\otimes d}$ and $\mathcal{H}_K$, making the exponentiated KGFD kernel positive definite.

4.1 A DIFFUSION INTERPRETATION OF THE GENERALIZED FISHER DIVERGENCE

In this section, we further analyze the properties of the GFD by establishing a link with diffusion processes. This link further anchors the GFD to the array of previously known divergences, while opening the door for possible refinements and generalizations of the GFD.

A Dissipation Inequality Diffusion processes (?) are well-known instances of stochastic processes $(X_t)_{t\geq 0}$ that evolve from some initial distribution μ_0 towards a target distribution p according to the differential update rule

$$\mathrm{d}X_t = s_p(X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0 \sim \mu_0.$$

For any time $t \geq 0$, the probability density of X_t is the solution $\mu_{\mu_0,p}(\cdot,t)$ of the so-called Fokker-Planck equation

$$\frac{\partial\mu(x,t)}{\partial t} = \operatorname{div}(-\mu(x,t)s_p(x)) + \Delta_x\mu(x,t) \qquad (11)$$

with initial condition $\mu(\cdot,0) = \mu_0$. **??** establishes a link between these solutions and the GFD:

Proposition 4.3 (Diffusion interpretation of the GFD)**.** *Let $\mu_{\nu,p}$ (resp. $\mu_{\nu,q}$) be the solution of* **??** *with initial condition ν and target p (resp. q). Let k be a* real-valued*, twice-differentiable kernel. Then, we have that*

$$\lim_{t\to 0}\frac{1}{t}\,\text{MMD}(\mu_{\nu,p}(\cdot,t), \mu_{\nu,q}(\cdot,t)) = \sqrt{\text{KGFD}(p,q)}$$

where the MMD *is w.r.t. the kernel k, and the* KGFD *is with respect to the matrix-valued kernel $\nabla_x\nabla_y k(x,y)$.*

Proof. See Section D of the Appendix. □

?? frames the exponentiated KGFD kernel as the $t \to 0$ limit of the kernel obtained by setting

$$\phi_t \colon p \longmapsto \nabla_x \log\mu_{\nu,p}(\cdot,t)$$

which is the score of the solution of the Fokker-Planck equation **??** with target p and initial measure ν, and setting $H = \mathcal{H}$. Interestingly, the other limit case $t \to \infty$ recovers the exponentiated MMD kernel. Indeed, under mild conditions, the Fokker-Planck solution converges to the target and thus we have that $\lim_{t\to\infty}\phi_t(p) = p$: the feature map converges to the identity. Thus, the diffusion framework introduced above allows to recover both the KGFD and the MMD as special cases. However, while the limit $t \to 0$ and $t \to \infty$ yield both Hilbertian metrics, it is an open question whether for a given time $0 < t < \infty$, ϕ_t is also Hilbertian. A positive answer to this question would allow to construct positive definite kernels that can possibly overcome the pitfalls of score-based tools (**??**), while being computable in finite time.

5 FAST AND SCALABLE CALIBRATION TESTS

The framing of the calibration testing problem of **??** alongside with the GFD-based kernels of **??** allows us to design a fast and scalable alternative to the pioneering tests of **?**. The full testing procedure is outlined in **??**.

```
Algorithm 3: CGOF Calibration Test (GFD Kernel)
Data: Pairs $\{(P_{|x^i}, y^i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$
Result: Whether to reject $H_0$: “model is calibrated”
Parameters: Base measure $\nu$, number of base samples
            $m$, number of data samples $n$, kernel
            $l \colon \mathcal{Y}^2 \to \mathbb{R}$, set of indices pairs
            $R \subset [\{1, \ldots, n\}]^2$, significance level $\alpha$
for $i \leftarrow 1$ to $m$ do
    Draw $z^i \sim \nu$
end
for $(i, j) \in R$ do
    /* Use ?? with base samples $\{z^k\}_{k=1}^{m}$ */
    $\kappa^{i,j} \leftarrow K_\nu(\widehat{P_{|x^i}, P_{|x^j}})$
end
Run ?? with kernel $k(P_{|x^i}, P_{|x^j}) := \kappa^{i,j}$
```

Use as a Regularizer We additionally propose to use this statistic as a calibration regularizer during the training procedure of a probabilistic model. Because the plug-in estimator of D_p using the proposed kernels is cheap to compute and differentiable, adding this regularizer adds little overhead to the entire procedure and allows us to train calibrated models instead of simply diagnosing calibration of existing models.

6 RELATED WORK

The conservativeness of SBI posterior models was first investigated by **?**. The authors introduced the notion of *expected coverage probability*, which can be seen as a consequence of calibration in the sense of **?**, as shown in **??**. **?** highlighted the computational cost of assessing the conservativeness of such SBI posterior models, which is particularly high for sequential inference methods. **?** introduced a calibration test for predictive models of the form of $g \colon x \mapsto \widehat{p}(\cdot \,|\, x)$, which could be used to assess the calibration of such posterior models more efficiently than in **?**. However, since calibration implies conservativeness, using this test likely generates *false positives*.

Related work on assessing calibration for SBI models can be divided into 2 categories, depending on whether they investigate *frequentist* calibration **??** or *Bayesian* calibration **??**.

Generally, there are two main lines of research on calibration: How to assess model calibration and how to improve it. Typically, calibration of classification models is quantified with the expected calibration error (ECE), using different notions of calibration **?????**. Common ECE estimators are histogram-regression based and require binning the predictions into different clusters. Unfortunately, these estimators are usually biased and inconsistent **?**. The maximum calibration error **?** is another measure of calibration. More recently, different kernel-based statistics such as the maximum mean calibration error **?** and the more general SKCE **??** were introduced. In contrast to the ECE and MCE, unbiased and consistent estimators exist for the kernel-based calibration measures. Moreover, they are differentiable and hence, as the KCCSD, they can be used as regularizer in gradient-based training algorithms **?**.

Suggestion: Shorten dicussion below, move partly to the appendix and integrate with the first paragraph in this section

We consider the case of Simulation-Based Inference, for which reliability is a crucial property for trustworthy scientific discovery as discussed by **??**. **?** use the notion of expected posterior coverage to quantify and check if posterior approximations $\hat{p}(\theta \,|\, x)$ obtained with common methods in simulation-based inference are too overconfident. Here we use θ instead of y to denote the target in order to comply with the conventions of Bayesian inference.

7 EXPERIMENTS

We validate the properties of our proposed calibration tests with synthetic data and compare them with existing tests based on the SKCE. We compare the impact of different kernels for $P_{|\cdot}$: The exponentiated GFD kernel with a standard Gaussian as base measure, the exponentiated KGFD with a standard Gaussian as base measure and a vector-valued kernel $K(x, x') = k(x, x') I_{\mathcal{X}}$ with real-valued Gaussian kernel k, the exponentiated MMD kernel with a Gaussian kernel on the ground-space, and, for isotropic Gaussian distributions, the exponentiated Wasserstein kernel with closed-form expression

$$\begin{aligned} &k_W\big(\mathcal{N}(\mu, \sigma^2 I_d), \mathcal{N}(\mu', {\sigma'}^2 I_d)\big) \\ &\qquad = \exp\big(-(\|\mu - \mu'\|_2^2 + d(\sigma^2 - {\sigma'}^2))/(2\ell^2)\big). \end{aligned}$$

On $\mathcal{Y}$, we study the Gaussian and the inverse multi-quadric (IMQ) kernel.

We repeated all experiments with 100 resampled datasets and used 500 bootstrap iterations for approximating the quantiles of the test statistic with a prescribed significance level of $\alpha = 0.05$. The bandwidths of the kernels, including the ground-space kernels of the KGFD and the exponentiated MMD kernel, are selected with the median heuristic.

7.1 FALSE REJECTION RATE

First we check empirically that the type I error (false rejection rate) of the proposed tests is asymptotically upper bounded by the prescribed significance level. To that end, we generate calibrated data in a two-step procedure: First we sample distributions $P_{|x^i}$ and then we draw a corresponding target y^i for each $P_{|x^i}$.

We compare two different models for sampling Gaussian distributions $P_{|x^i}$:

Mean Gaussian Model (MGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $\mathbb{P}(X) = \mathcal{N}(0, I_d)$, and $P_{|x} = \mathcal{N}(x, I_d)$. A similar model was used by **?**.

Linear Gaussian Model (LGM) Here $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_d)$, and $P_{|x} = \mathcal{N}(\sum_{i=1}^d ix_i, 1)$. This model was used by **?**.

Figures **??** and F.1 demonstrate that the proposed KCCSD tests are calibrated and their type I errors do not exceed the set significance level, apart from sampling noise. We see that this holds both when we evaluate the kernels exactly, exploiting closed-form expressions for Gaussian distributions, and when we approximate the kernel evaluations using samples from the base measure. The comparison with SKCE (Figures F.2 and F.3) highlights that for the SKCE exact computations of the test statistic are only possible for specific combinations of models and kernels (in this example, Gaussian distributions with Gaussian kernels).

7.2 TEST POWER

We compare the test power of the proposed tests in different scenarios with the test power of SKCE-based tests. We generate uncalibrated data with the following models:

Miscalibrated Mean Gaussian Model (MMGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $\mathbb{P}(X) = \mathcal{N}(0, I_d)$, $\mathbb{P}(Y \,|\, X = x) = \mathcal{N}(x, I_d)$, and $P_{|x} = \mathcal{N}(x + \delta, I_d)$ for some $\delta \neq 0$. We consider perturbations δ of the form $\delta_0 \mathbf{1}_d$ (miscalibration of all dimensions) and $\delta_0 e_1$ (miscalibration of only the first dimension) with different scales $\delta_0 > 0$. A similar model was used by **?**.

Heteroscedastic Gaussian Model (HGM) Here $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_d)$, $\mathbb{P}(Y \,|\, X = x) = \mathcal{N}(\sum_{i=1}^d x_i, 1)$, and $P_{|x} = \mathcal{N}(\sum_{i=1}^d x_i, \sigma^2(x))$ where $\sigma^2(x) = 1 + 10 \exp\left(-\|x - c\|_2^2/(2 \times 0.8^2)\right)$ for $c = 2/3\, \mathbf{1}_d$. This model was used by **?**.

Quadratic Gaussian Model (QGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = U(-2, 2)$, $\mathbb{P}(Y \,|\, X = x) = \mathcal{N}(0.1x^2 + x + 1, 1)$, and $P_{|x} = \mathcal{N}(x + 1, 1)$. This model was used by **?**.

8 CONCLUSION

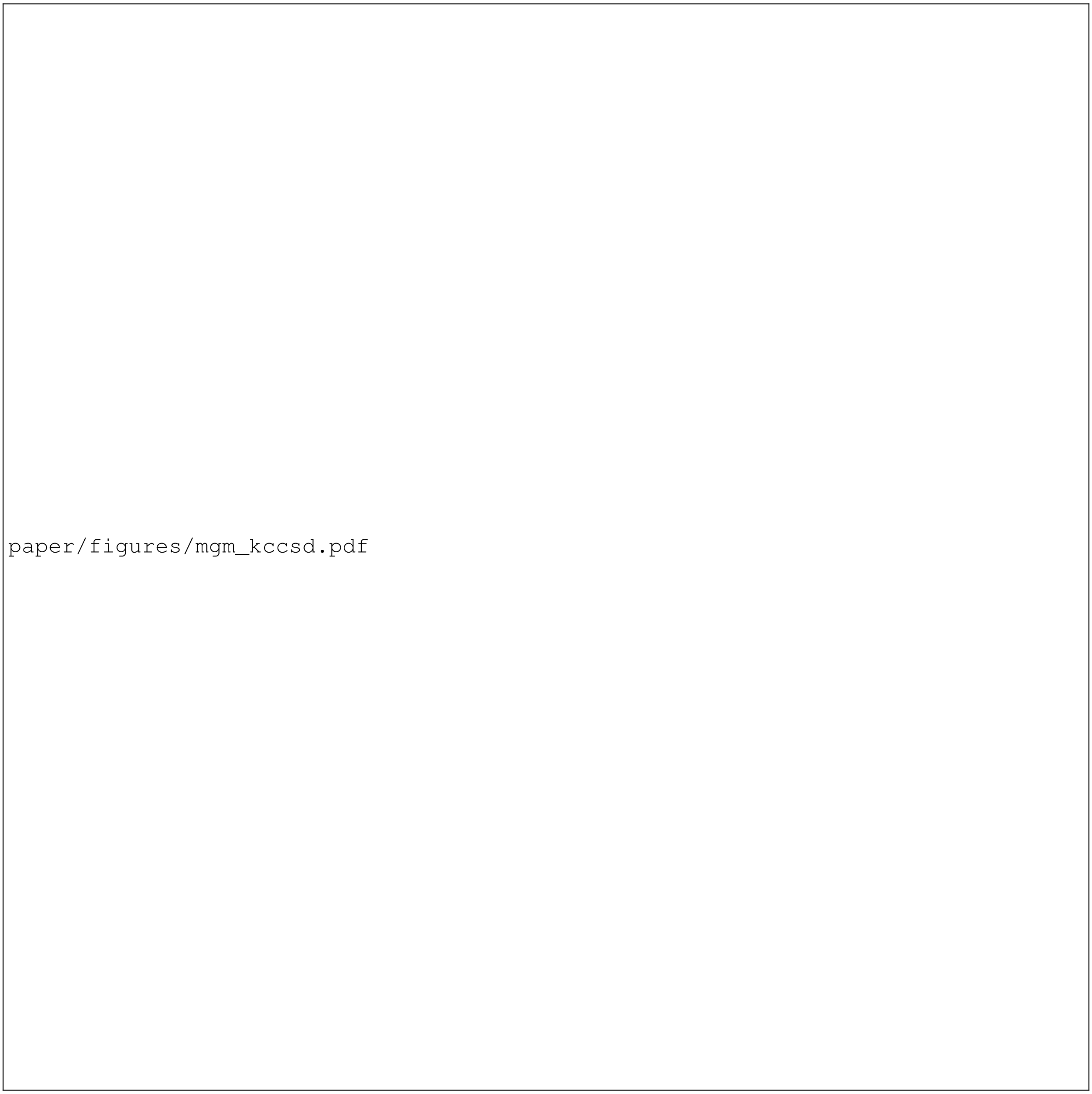

Figure 1: False rejection rate of the KCCSD for MGM using exact and approximate (10 base samples) evaluations.

paper/figures/pmgm_kccsd_delta=0.1_shift-dim=1.pdf

Figure 2: Rejection rate of the KCCSD for MMGM ($\delta = 0.1e_1$) using exact and approximate (10 base samples) evaluations.