Fast and Scalable Score-Based Kernel Calibration Tests

Pierre Glaser[1] **David Widmann**[2] **Fredrik Lindsten**[3] **Arthur Gretton**[1]

[1]Gatsby Computational Neuroscience Unit, University College London, United Kingdom
[3]Division of Statistics and Machine Learning, Linköping University, Sweden
[2]Department of Information Technology, Uppsala University, Sweden

Abstract

We introduce the *Kernel Calibration Conditional Stein Discrepancy test* (KCCSD test), a non-parametric, kernel-based test for assessing the calibration of probabilistic models with well-defined scores. In contrast to previous methods, our test avoids the need for possibly expensive expectation approximations while providing control over its type-I error. We achieve these improvements by using a new family of kernels for score-based probabilities that can be estimated without probability density samples, and by using a conditional goodness-of-fit criterion for the KCCSD test's U-statistic. We demonstrate the properties of our test on various synthetic settings.

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. This property is essential in safety-critical applications such as autonomous driving. Unfortunately, empirical studies revealed that popular machine learning models such as deep neural networks tend to trade off calibration for accuracy [Guo et al., 2017]. This has lead to an increased interest in the study of calibrated models in recent years.

Calibration has been studied in the meteorological and statistical literature for many decades [e.g., Murphy and Winkler, 1977, DeGroot and Fienberg, 1983]. For a long time, research on calibration has been focused on different notions of calibration for probabilistic classifiers [e.g., Murphy and Winkler, 1977, DeGroot and Fienberg, 1983, Platt, 2000, Zadrozny and Elkan, 2001, Bröcker, 2009, Naeini et al., 2015, Guo et al., 2017, Kull et al., 2017, Kumar et al., 2018, Kull et al., 2019, Vaicenavicius et al., 2019, Widmann et al., 2019] and on calibration of quantiles and confidence intervals for real-valued regression problems [e.g., Ho and Lee, 2005, Rueda et al., 2006, Taillardat et al., 2016, Song et al., 2019, Fasiolo et al., 2020]. Regarding the calibration of classification models, different hypothesis tests have been proposed [e.g., Cox, 1958, Bröcker and Smith, 2007, Vaicenavicius et al., 2019, Widmann et al., 2019, Gweon, 2022, Lee et al., 2022]. Given a predictive model and a validation dataset, these tests output whether a model is likely to be uncalibrated. The recent work of Widmann et al. [2021] generalized the calibration-framework introduced for classification in [Widmann et al., 2019] to (possibly multi-dimensional) continuous-valued predictive models. In particular, Widmann et al. [2021] introduced a kernel-based hypothesis test for such general classes of models.

An important potential consumer of calibration tests is Bayesian inference, and in particular simulation-based inference (SBI), for which miscalibration is particularly undesirable. SBI [Cranmer et al., 2020] lies at the intersection of machine learning and domain sciences, and refers to the set of methods that train probabilistic models to estimate the posterior over scientific parameters of interest given some observed data. The models are trained using pairs composed of parameters drawn from a prior distribution, and their associated "synthetic" observed data, obtained by running a probabilistic program called the *simulator*, taking a parameter value as input, and that faithfully mimics the physical generative process of interest. The increasing number of use cases combined with advances in probabilistic modeling has elevated SBI to a critical role in solving complex scientific problems such as particle physics [Gilman et al., 2018] and neuroscience [Glöckler et al., 2022, Glaser et al., 2022]. However, as discussed in [Hermans et al., 2021], overconfidence in SBI models can conceal credible alternative scientific hypotheses, and result in incorrect discoveries [Hermans et al., 2021], highlighting the need for

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

principled and performant calibration tests suitable for such models.

While the theoretical framework of Widmann et al. [2021] describes the calibration of any probabilistic model, applying its associated calibration test to Bayesian inference remains challenging: indeed, the test statistics require computing expectations against the probabilistic models of interest, for reasons bearing both to the calibration setting, and to the limitations of currently available kernel-based tools for probabilistic models. Although such expectations can be computed exactly for classification models, expectations against generic probabilistic models are usually intractable and must be approximated. In cases where the models are *unnormalized*, these approximations are both computationally expensive—sometimes, prohibitively—and biased, thereby compromising theoretical guarantees of the calibration tests of Widmann et al. [2021], including type-I error control.

Contributions In this paper, we introduce the kernel calibration-conditional Stein discrepancy (or KCCSD) test, a new nonparametric, score-based kernel calibration test which addresses the limitations of existing methods. The KCCSD test builds on the insight that the definition of calibration given by Vaicenavicius et al. [2019] is a conditional goodness of fit property, as we remark in Section 3. This fact allows us to leverage the kernel conditional goodness of fit test proposed by Jitkrittum et al. [2020] as the backbone of the KCCSD test. Unlike the test-statistics of Widmann et al. [2021], the KCCSD test statistic does not contain *explicit* expectations against the probabilistic models; however, as in [Widmann et al., 2021], it requires evaluating a kernel between probabilities densities, which in most cases of interest introduces an (intractable) expectation against the densities. To eliminate this limitation, we construct two new kernels between probability distributions that do not involve expectations against its input distributions, while remaining suitable for statistical testing. These kernels rely on a generalized version of the Fisher divergence and are of independent interest. We investigate a connection between these kernels and diffusion, akin to Stein methods, and discuss the relationships with other kernels on distributions. By using such kernels in the KCCSD test statistic, we obtain a fast and scalable calibration test that remains consistent and calibrated for unnormalized models, answering the need for such tests discussed above. We confirm in Section 6 the properties and benefits of the KCCSD test against alternatives on synthetic experiments.

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*, e.g *maps* $P_{|\cdot} \colon x \in \mathcal{X} \longmapsto P_{|x}(\cdot) \in \mathcal{P}(\mathcal{Y})$ approximating the unknown conditional distribution of Y given $X = x$, e.g $P_{|x}(\cdot) \simeq \mathbb{P}(Y \in \cdot \mid X = x)$. The target variable Y is typically a parameter of a probabilistic system of interest—like synaptic weights in biological neural networks—while the 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 [Bröcker, 2008, Vaicenavicius et al., 2019, Widmann et al., 2021] 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 right-hand side exist.

Hypothesis Testing: Kernel Calibration Error There are multiple ways to test whether a given predictive probabilistic model is calibrated. In this section, we introduce the kernel-based tests of Widmann et al. [2019] and their later generalization [Widmann et al., 2021], since our KCCSD test is built on these approaches. These tests turn the equality between conditional distributions present in Equation (1) into a more classical equality between two joint distributions. The transformation is achieved by noting that

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

where Z is an "auxiliary" variable such that $Z \mid P_{|X} \sim P_{|X}(\cdot)$. This identity was used by Widmann et al. [2021] to construct an MMD-type calibration test based on the (squared) kernel calibration error (SKCE) criterion

$$\sup_{h \in \mathcal{B}(0_{\mathcal{H}}, 1)} \mathbb{E}_{(x,y,z) \sim \mathbb{P}(X,Y,Z)} \left[h(P_{|x}, y) - h(P_{|x}, z)\right]. \tag{2}$$

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 $k_{\mathcal{H}} \colon (P_{|\mathcal{X}} \times \mathcal{Y})^2 \to \mathbb{R}$. As noted by Widmann et al. [2021], the SKCE generalizes the (squared) kernel classification calibration error (SKCCE) defined for the special case of discrete output spaces $\mathcal{Y} = \{1, \ldots, d\}$ [Widmann et al., 2019], to continuous ones. Given n pairs of samples $\{(P_{|x^i}, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$, Widmann et al. [2021] consider the following SKCE estimator

$$\widehat{\mathrm{SKCE}} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} G((P_{|x^i}, y^i), (P_{|x^j}, y^j)) \tag{3}$$

where

$$\begin{aligned} G((p, y), (p', y')) \coloneqq & k((p, y), (p', y')) \\ & - \mathbb{E}_{z \sim p}\, k((p, z), (p', y')) \\ & - \mathbb{E}_{z' \sim p'}\, k((p, y), (p', z')) \\ & + \mathbb{E}_{z \sim p}\, \mathbb{E}_{z' \sim p'}\, k((p, z), (p', z')). \end{aligned} \tag{4}$$

For a target false rejection rate $\alpha \in (0,1)$, the test of Widmann et al. [2021] follows standard methodology in recent nonparametric testing [Gretton et al., 2012, 2007, Chwialkowski et al., 2016] by rejecting the null hypothesis (e.g, that the model is calibrated) if $\widehat{\mathrm{SKCE}} > \gamma_{1-\alpha}$, where $\gamma_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of $\widehat{\mathrm{SKCE}}$ under the null. While various methods are available to estimate this quantile, all experiments in this paper use a *bootstrap* approach [Arcones and Gine, 1992]. As discussed by Widmann et al. [2021], Equation (4) contains 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 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, taking as input two probability distributions p and p', are more limited and require expensive approximations methods when working with unnormalized models.

A popular approach to design kernels on distributions [Szabó et al., 2015, 2016] is to first embed the probability distributions in a 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 inverse 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 $\|\phi(\mu)-\phi(\nu)\|^2_\mathcal{H} \coloneqq \mathrm{MMD}^2(\mu,\nu)$, are often referred to as MMD-type kernels [Meunier et al., 2022]. Other distances, like the Wasserstein distance in 1 dimension or the sliced Wasserstein distance [Bonneel et al., 2015] 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 [Meunier et al., 2022]. In general, however, computing $k_P(p,p')$ 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 Equation (2) would require an estimator that does not require samples from p and p'.

2.2 KERNEL CONDITIONAL GOODNESS-OF-FIT TEST

We briefly introduce the background on goodness-of-fit methods relevant to our new test. *Conditional goodness-of-fit* (or CGOF) testing adapts the familiar goodness of fit tests to the conditional case. In particular, CGOF tests whether

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

given a candidate $Q_{|z}$ for the conditional distribution $\mathbb{P}(Y \in \cdot \mid Z = z)$ and samples $\{(z^i,y^i)\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathbb{P}(Z,Y)$. This problem was studied by Jitkrittum et al. [2020] for the case $\mathcal{Z}\times\mathcal{Y} \subset \mathbb{R}^{d_z}\times\mathbb{R}^{d_y}$ and models $Q_{|z}$ with a differentiable, strictly positive density $f_{Q_{|z}}$. They proposed a kernel CGOF test for Equation (5) based on the (squared) kernel conditional Stein discrepancy (KCSD)

$$D_{Q_{|\cdot}}(\mathbb{P}) \coloneqq \left\|\mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)}\left[K_z\xi_{Q_{|z}}(y,\cdot)\right]\right\|^2_{\mathcal{F}_K} \tag{6}$$

Here, $\mathcal{F}_K$ is an $\mathcal{F}_l^{d_y}$ (e.g $\overbrace{\mathcal{F}_l\times\cdots\times\mathcal{F}_l}^{d_y \text{ times}}$) -vector-valued RKHS with kernel $K\colon \mathcal{Z}\times\mathcal{Z}\to\mathcal{L}(\mathcal{F}_l^{d_y},\mathcal{F}_l^{d_y})$, K_z is its associated linear operator on $\mathcal{F}_l^{d_y}$ with $K_z g \coloneqq K(z,\cdot)g \in \mathcal{L}(\mathcal{Z},\mathcal{F}_l^{d_y})$ for $g\in\mathcal{F}_l^{d_y}$, $\mathcal{F}_l$ is an RKHS on $\mathcal{Y}$ with kernel $l\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ and $\xi_{Q_{|z}}$ is the "kernelized score":

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

We refer to Jitkrittum et al. [2020, Section 2 and 3] for an intuition behind the KCSD formula, and its relationship to the more familiar Kernel Stein Discrepancy Chwialkowski et al. [2016], Gorham and Mackey [2017]. Under certain assumptions, the null hypothesis in Equation (5) is true if and only if $D_{Q_{|\cdot}}(\mathbb{P}) = 0$. In particular, as shown in [Jitkrittum et al., 2020, Theorem 1], the latter will hold if $\mathcal{X}$ is compact, and the kernels K and l are universal, meaning that F_K (resp. $\mathcal{F}_l$) is dense w.r.t $\mathcal{C}(\mathcal{Z},\mathcal{F}_l^{d_y})$, the space of continuous functions from $\mathcal{Z}$ to $\mathcal{F}_l^{d_y}$ (resp. $C(\mathcal{Y},\mathbb{R})$). An instance of a universal $\mathcal{F}^{d_y}$-reproducing kernel is given by

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

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 universal kernel [Carmeli et al., 2010]. Jitkrittum et al. [2020] showed that the CGOF statistic $D_{Q_{|\cdot}}(\mathbb{P})$ admits an unbiased consistent estimator and used it to construct hypothesis tests of Equation (5) with operator-valued kernels of the form in Equation (7).

3 KERNEL CALIBRATION-CONDITIONAL STEIN DISCREPANCY

Calibration testing in the sense of Equation (1) is an instance of *conditional goodness-of-fit* testing of Equation (5) with input $Z = P_{|X}$, target Y, and models $Q_{|z} = z = P_{|x}$. Assuming that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ and that distributions $P_{|x}$ have a differentiable, strictly positive density $f_{P_{|x}}$. In that case, the (squared) kernel conditional Stein discrepancy in Equation (6) becomes

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

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 Equation (7), statistic $C_{P_{|\cdot}}(\mathbb{P})$ has an unbiased consistent estimator

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

where

$$H((p, y), (p', y')) \coloneqq k(p, p') h((p, y), (p', y')) \quad (9)$$

with

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

where $s_p(y) \coloneqq \nabla_y \log f_p(y)$ (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 Equation (7).

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.

Proposition 3.1 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 Equation* (9).

The full testing procedure is outlined in Algorithm 1. The computations can be performed with kernels K of the form in Equation (7) 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.

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 $k\colon (P_{|\mathcal{X}})^2 \to \mathbb{R}$, level α

```
/* Estimate KCCSD using Equation (10) or (A.1)    */
```
$\widehat{C} \leftarrow \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} H((P_{|x^i}, y^i), (P_{|x^j}, y^j))$
```
/* Use e.g. bootstrap [Arcones and Gine, 1992]    */
```
$\widehat{C}_\alpha \leftarrow$ 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

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 more general setting, including Bayesian inference.

General Recipe As in prior work on kernels for distributions [Meunier et al., 2022, Szabó et al., 2016], our proposed kernels take the form of exponentiated Hilbertian metrics

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

for two probability densities p and q, defined on some set $\mathcal{X} \subset \mathbb{R}^d$, where H is some separable Hilbert space, $\phi\colon p \mapsto \phi(p) \in H$ is a feature map, and $\sigma > 0$ is a bandwidth parameter. Our contributions in this section consist in pairs of carefully designed ϕ and H that will allow approximating k easily.

4.1 THE GENERALIZED FISHER DIVERGENCE (KERNEL)

Our starting point is the *Fisher Divergence* [Lyu, 2012, Sriperumbudur et al., 2017, Hyvärinen, 2005], also known as the *Relative Fisher Information* [Otto and Villani, 2000], between two probability densities p and q, which is given by

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

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

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

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 to the (squared) Fisher Divergence 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 [Hyvärinen, 2005], 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 unnormalized models. 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

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

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 $\phi \colon p \mapsto s_p(\cdot) \in \mathcal{L}_2(\nu)$ can be checked to be injective. 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

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

Since the (square root of the) GFD is a Hilbertian metric, K_ν is positive definite [Meunier et al., 2022], 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 Algorithm 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 Definition 4.2
Parameters: Base measure ν, num. of base samples m

```
for i ← 1 to m do
    Draw Z^i ~ ν
end
return exp(−(1/(2mσ²)) Σ_{i=1}^m ‖s_p(Z^i) − s_q(Z^i)‖²)
```

Use in hypothesis testing In addition to being tractable to estimate, we show that when $\mathcal{X}$ is compact (for instance, a bounded subset of $\mathbb{R}^d$), the exponentiated GFD kernels K_ν are *universal*. As a consequence, our KCCSD test, which is an instance of a KCSD test, will be able to distinguish the null-hypothesis from *any* alternative satisfying mild smoothness assumptions, as guaranteed by Jitkrittum et al. [2020, Theorem 1].

Proposition 4.3. *Assume that $\mathcal{X}$ is compact, ν has full support on $\mathcal{X}$, and let $\mathcal{P}_\mathcal{X}$ be the set of twice-differentiable probability densities on $\mathcal{X}$ equipped with the norm $\|p\|^2 = \|p\|^2_{\mathcal{L}_2(\nu)} + \sum_{i=1}^d \|\partial_i p\|^2_{\mathcal{L}_2(\nu)} + \sum_{i,j=1}^d \|\partial_i \partial_j p\|^2_{\mathcal{L}_2(\nu)}$. Then K_ν is universal for any bounded subset of $\mathcal{P}_\mathcal{X}$.*

Proof. The proof is given in Section D.2. □

4.2 THE KERNELIZED GENERALIZED FISHER DIVERGENCE (KERNEL)

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:

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

which, like the GFD, admits a tractable, unbiased estimator:

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

Since the KGFD is also a Hilbertian metric, we build upon it to construct our second proposal kernel:

Definition 4.4 (Exponentiated KGFD Kernel)**.** Consider the setting of Definition 4.2, and let k be a bounded positive definite kernel. The *exponentiated KGFD kernel* is given by:

$$K_{K,\nu} \coloneqq e^{-\mathrm{KGFD}(p,q)/(2\sigma^2)}$$

For characteristic kernels K, the integral operator $T_{K,\nu}$ is a Hilbertian isometry between $\mathcal{L}_2(\nu, \mathbb{R}^d)$ and $\mathcal{H}_K$, making the exponentiated KGFD kernel positive definite. Additionally, $K_{K,\nu}$ enjoys a similar universality property as its GFD analogue, as discussed in the next proposition.

Proposition 4.5. *Assume that $\mathcal{X}$ is compact, ν has full-support on $\mathcal{X}$, and let $\mathcal{P}_{\mathcal{X}}$ be the set of twice-differentiable probability densities equipped with the norm $\|p\|^2 = \|p\|^2_{\mathcal{L}_2(\nu)} + \sum_{i=1}^d \|\partial_i p\|^2_{\mathcal{L}_2(\nu)} + \sum_{i,j=1}^d \|\partial_i \partial_j p\|^2_{\mathcal{L}_2(\nu)}$. Then $K_{K,\nu}$ is universal for any bounded subset of $\mathcal{P}_{\mathcal{X}}$.*

A diffusion interpretation of the KGFD In this section, we establish a relationship between the KGFD and diffusion processes [Rogers and Williams, 2000], further anchoring the KGFD to the array of previously known divergences while opening the door for possible refinements and generalizations. 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

$$\mathrm{d}X_t = s_p(X_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 \sim \mu_0,$$

where W_t is a standard Brownian motion. 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

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

with initial condition $\mu(\cdot, 0) = \mu_0$. Proposition 4.6 establishes a link between these solutions and the KGFD:

Proposition 4.6 (Diffusion interpretation of the KGFD)**.** *Let $\mu_{\nu,p}$ (resp. $\mu_{\nu,q}$) be the solution of Equation* (11) *with initial condition ν and target p (resp. q). Let k be a* real-valued, *twice-differentiable kernel. Then, we have that*

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

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. □

Proposition 4.6 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 Equation (11) 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 $t \to \infty$ both yield 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 [Wenliang and Kanagawa, 2020, Zhang et al., 2022], while being computable in finite time.

5 FAST AND SCALABLE CALIBRATION TESTS

The framing of the calibration testing problem of Section 3 alongside with the GFD-based kernels of Section 4 allows us to design a fast and scalable alternative to the pioneering tests of Widmann et al. [2019]. The full testing procedure is outlined in Algorithm 3.

Calibration tests as a reliability tests in Bayesian inference As one main motivation for studying calibration of generic probabilistic models is Bayesian inference, it is important to note that reliability metrics traditionally used 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 ν, num. of base samples m, number of data samples n, kernel $l\colon \mathcal{Y}^2 \to \mathbb{R}$, significance level α

for $i \leftarrow 1$ **to** m **do**
  Draw $z^i \sim \nu$
end
for $1 \leq i < j \leq n$ **do**
  /* Use Algorithm 2 with base samples $\{z^k\}_{k=1}^m$ */
  $\kappa^{i,j} \leftarrow K_\nu(\widehat{P_{|x^i}, P_{|x^j}})$
end
Run Algorithm 1 with kernel $k(P_{|x^i}, P_{|x^j}) := \kappa^{i,j}$

Bayesian inference such as conservativeness [Hermans et al., 2021] differ from the notion of calibration in Equation (1). We first briefly recall the notion of posterior coverage:

Definition 5.1 (Conservativeness of a Bayesian model [Hermans et al., 2021])**.** Let $P_{|x}(\cdot)$ be a conditional distribution model for $\mathbb{P}(Y \in \cdot \mid X = x)$, and assume that $P_{|x}$ has a density $f_{P_{|x}}$ for $\mathbb{P}(X)$-almost every x. For level $1 - \alpha \in [0, 1]$, let $\Theta_{P_{|x}}(1 - \alpha)$ be the highest density region of $P_{|x}$.[1] Then $P_{|\cdot}$ is said to be *conservative* if

$$\mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)}\, \mathbb{1}_{\Theta_{P_{|x}}(1-\alpha)}(y) \geq 1 - \alpha.$$

In the following proposition, we show that a probabilistic model that is calibrated according to Equation (1) is also conservative in the sense of Hermans et al. [2021], grounding the use of our tests in Bayesian inference.

Proposition 5.2 (Calibrated models are conservative)**.** *If a model $P_{|\cdot}$ is calibrated in the sense of Equation* (1)*, then it is conservative.*

The proof is given in Section C of the appendix.

6 EXPERIMENTS

We validate the properties of our proposed calibration tests with synthetic data and compare them with existing tests based on the SKCE.[2] More concretely, we run KCCSD tests using either a exponentiated GFD kernel or kernelized exponentiated GFD kernel with a matrix-valued kernel of the form in Equation (7) with real-valued Gaussian kernel k; and compare them with SKCE tests using two already investigated kernels on distributions: 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

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

We set the base measure ν of the GFD and kernelized GFD kernels to be a standard Gaussian. On $\mathcal{Y}$, we study the Gaussian and the inverse multi-quadric (IMQ) kernel.

We repeated all experiments with 100 resampled datasets and used a wild bootstrap with 500 samples for approximating the quantiles of the test statistic with a prescribed significance level of $\alpha = 0.05$. The bandwidths of the kernels are selected with the median heuristic. A "second-order" median heuristic is used for the ground-space kernels of the KGFD and the exponentiated MMD kernel: For each pair of distributions, we compute the median distance between samples from an equally weighted mixture of these distributions (numerically for tractable cases such as Gaussian distributions and using samples otherwise), and then the bandwidth of the kernel is set to the median of these evaluations.

We repeatedly generate datasets $\{(P_{|x^i}, y^i)\}_i$ 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 different setups of targets Y and Gaussian distributions $P_{|X}$ with varying degree $\delta \geq 0$ of miscalibration (models are calibrated for $\delta = 0$ and miscalibrated otherwise):[3]

Mean Gaussian Model (MGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^5$, $\mathbb{P}(X) = \mathcal{N}(0, I_5)$, $\mathbb{P}(Y \mid X = x) = \mathcal{N}(x, I_5)$, and $P_{|x} = \mathcal{N}(x + \delta c, I_5)$ for $c \in \{\mathbf{1}_5, e_1\} \subset \mathbb{R}^5$ (miscalibration of all dimensions or only the first one).

Linear Gaussian Model (LGM) Here $\mathcal{X} = \mathbb{R}^5$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_5)$, and $P_{|x} = \mathcal{N}(\delta + \sum_{i=1}^5 i x_i, 1)$.

Heteroscedastic Gaussian Model (HGM) Here $\mathcal{X} = \mathbb{R}^3$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_3)$, $\mathbb{P}(Y \mid X = x) = \mathcal{N}(m(x), 1)$, and $P_{|x} = \mathcal{N}(m(x), \sigma^2(x))$ with $m(x) = \sum_{i=1}^3 x_i$ and $\sigma^2(x) = 1 + 10\delta \exp\left(-\|x - c\|_2^2/(2 \cdot 0.8^2)\right)$ for $c = 2/3\, \mathbf{1}_3$.

Quadratic Gaussian Model (QGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{U}(-2, 2)$, $\mathbb{P}(Y \mid X = x) = \mathcal{N}(0.1x^2 + x + 1, 1)$, and $P_{|x} = \mathcal{N}(0.1(1 - \delta)x^2 + x + 1, 1)$.

Figure 1 demonstrates that the proposed KCCSD tests are calibrated: The false rejection rates (type I errors) of the

[1] The highest density region of a probabilistic model $P_{|x}$ with density $f_{P_{|x}}$ is defined [see, e.g., Hyndman, 1996] by $\Theta_{P_{|x}}(1 - \alpha) := \{y\colon f_{P_{|x}}(y) \geq c_{P_{|x}}(1 - \alpha)\}$ where $c_{P_{|x}}(1 - \alpha) := \sup\{c\colon \int \mathbb{1}_{[c,\infty)}(f_{P_{|x}}(y))\, P_{|x}(\mathrm{d}y) \geq 1 - \alpha\}$.

[2] The code to reproduce the experiments is available at `https://github.com/pierreglaser/kccsd`.

[3] MGM is adapted from a model used by Widmann et al. [2021], and LGM, HGM, and QGM were used by Jitkrittum et al. [2020].

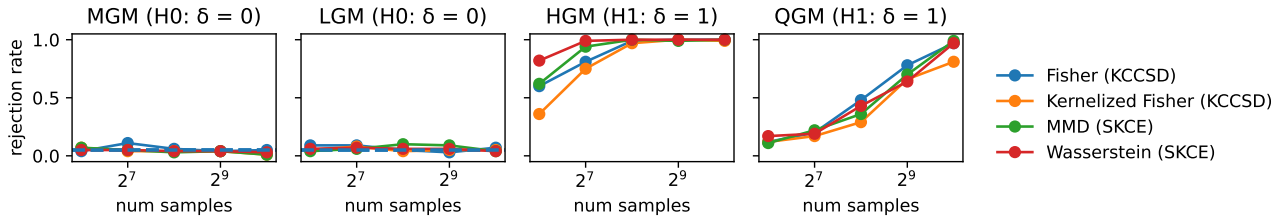

Figure 1: Rejection rates of the KCCSD and SKCE tests with a Gaussian kernel on the target space $\mathcal{Y}$ (significance level $\alpha = 0.05$). All kernels and test statistics are evaluated exactly using closed-form expressions.

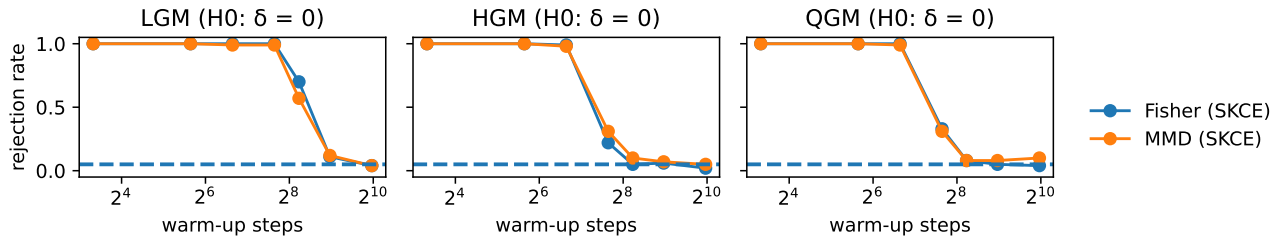

Figure 2: False rejection rates of the SKCE tests for the calibrated LGM, HMC, and QGM ($n = 200$ data points, significance level $\alpha = 0.05$). The expectations in the test statistic are estimated with 2 samples obtained with the Metropolis-adjusted Langevin algorithm (MALA) without step size tuning.

calibrated MGM and LGM do not exceed the set significance level, apart from sampling noise. Figures F.1 and F.7 in the suppplementary material confirm empirically that this is the case also when we approximate the Fisher and MMD kernels using samples.

Moreover, we see in Figure 1 that for the miscalibrated HGM the SKCE tests exhibit larger rejection rates, and hence test power, than the KCCSD tests in the small sample regime, regardless of the kernel choice. This specific setting with Gaussian distributions and a Gaussian kernel on the target space $\mathcal{Y}$ is favourable to the SKCE test as both its test statistic, as well as the exponentiated MMD or Wasserstein kernel evaluations are available in closed-form. In such analytical scenarios we expect the score-based KCCSD tests to perform worse [Wenliang and Kanagawa, 2020, Zhang et al., 2022]. However, the KCCSD tests present themselves as a practically useful alternative even in this example: For the miscalibrated HGM their rejection rates are close to 100% with ≥ 256 data points, and for the miscalibrated QGM they show very similar performance as the SKCE tests. Overall, as expected, we see in Figure 1 that for all studied tests rejection rates for the miscalibrated models increases with increasing number of samples.

One main advantage of the KCCSD over the SKCE is that it has first-class support for unnormalized models for which only the score function is available: In contrast to the SKCE its test statistic only involves scores but no expectations. In principle, for unnormalized models these expectations in the test statistic of the SKCE can be approximated with, e.g., MCMC sampling. However, Figure 2 shows that there is a major caveat: If the MCMC method is not tuned sufficiently well (e.g., if the chain is too short or the proposal step size is not tuned properly), it might return biased samples which causes the SKCE tests to be miscalibrated. On the other hand, increasing the number of MCMC samples increases the computational advantage of the KCCSD even more.

Another difference between the KCCSD and SKCE is highlighted in Figures F.1 and F.2: The number of combinations of kernels for which the test statistic can be evaluated exactly is smaller for the SKCE (in these Gaussian examples, it requires Gaussian kernels on the target space).

One limitation of the (kernelized) exponentiated GFD Kernel is that it necessitates setting an additional hyperparameter: the base measure ν, which weights the score differences between its two input distributions p and q at all points of the ground space $\mathcal{X}$. While our experiments have set ν to be a Gaussian measure in order to obtain closed-form expressions for Gaussian p, q, other choices may be more adequate depending on the problem at hand. For instance, when p and q are posterior models for a given prior π, we hypothesize that setting ν to π constitutes a better default choice.

7 CONCLUSION

In this paper, we introduced the Kernel Calibration Conditional Stein Discrepancy test, a fast and reliable alternative to prior calibration tests for general, density-based probabilistic models, thereby addressing an important need in the Bayesian inference community. In doing so, we introduced

kernels for density-based inputs, which we believe are of independent interest and could be used in other domains such as distribution regression [Szabó et al., 2016] or meta-learning [Denevi et al., 2020]. Moreover, while the set of experiments conducted in this paper focused on "offline" calibration testing, its low computational cost opens the door to promising new use cases. One particularly interesting avenue would consist in using the KCCSD test criterion as a regularizer directly within the training procedure of a probabilistic model, allowing not only to detect miscalibration but also to prevent it in the first place. We look forward to seeing extensions and applications of the tools introduced in this paper.

Acknowledgements

This research was financially supported by the Centre for Interdisciplinary Mathematics (CIM) at Uppsala University, Sweden, by the project *NewLEADS - New Directions in Learning Dynamical Systems* (contract number: 621-2016-06079), funded by the Swedish Research Council, and by the *Kjell och Märta Beijer Foundation*. Pierre Glaser and Arthur Gretton acknowledge support from the Gatsby Charitable Foundation.

References

M. A. Arcones and E. Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674, 1992.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vis.*, 51(1):22–45, 2015.

J. Bröcker. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 2008.

J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 2009.

J. Bröcker and L. A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 2007.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 2010.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

D. R. Cox. Two Further Applications of a Model for Binary Regression. *Biometrika*, 45(3/4):562, Dec. 1958.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 1983.

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. In *NeurIPS*, 2020.

M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 2020.

D. Gilman, S. Birrer, T. Treu, C. R. Keeton, and A. Nierenberg. Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 2018.

P. Glaser, M. Arbel, A. Doucet, and A. Gretton. Maximum likelihood learning of energy-based models for simulation-based inference. *arXiv e-prints*, 2022.

M. Glöckler, M. Deistler, and J. H. Macke. Variational methods for simulation-based inference. *arXiv preprint arXiv:2203.04176*, 2022.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, 2017.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

H. Gweon. A power-controlled reliability assessment for multi-class probabilistic classifiers. *Advances in Data Analysis and Classification*, 2022.

J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, and G. Louppe. Averting a crisis in simulation-based inference, 2021.

Y. H. S. Ho and S. M. S. Lee. Calibrated interpolated confidence intervals for population quantiles. *Biometrika*, 2005.

R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 1996.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6(24): 695–709, 2005.

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

M. Kull, T. S. Filho, and P. Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, 2017.

M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *NeurIPS*, pages 12316–12326, 2019.

A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, 2018.

D. Lee, X. Huang, H. Hassani, and E. Dobriban. T-Cal: An optimal test for the calibration of predictive models, 2022.

S. Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

D. Meunier, M. Pontil, and C. Ciliberto. Distribution regression with sliced Wasserstein kernels. In *ICML*, 2022.

A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 1977.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 2000.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

L. C. Rogers and D. Williams. *Diffusions, Markov processes, and martingales: Volume 1, foundations*. Cambridge University Press, 2000.

M. Rueda, S. Martinez-Puertas, H. Martinez-Puertas, and A. Arcos. Calibration methods for estimating quantiles. *Metrika*, 2006.

H. Song, T. Diethe, M. Kull, and P. Flach. Distribution calibration for regression. In *ICML*, 2019.

B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.*, 2017.

Z. Szabó, A. Gretton, B. Póczos, and B. K. Sriperumbudur. Two-stage sampled learning theory on distributions. In *AISTATS*, 2015.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *J. Mach. Learn. Res.*, 2016.

M. Taillardat, O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 2016.

J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. Evaluating model calibration in classification. In *AISTATS*, 2019.

L. K. Wenliang and H. Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests in multi-class classification: A unifying framework. In *NeurIPS*, 2019.

D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests beyond classification. In *ICLR*, 2021.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML*, 2001.

M. Zhang, O. Key, P. Hayes, D. Barber, B. Paige, and F.-X. Briol. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.