Testing Goodness of Fit of Conditional Density Models with Kernels

Wittawat Jitkrittum\textsuperscript{1}, Heishiro Kanagawa\textsuperscript{2}, Bernhard Schölkopf\textsuperscript{1}

\textsuperscript{1}MPI for Intelligent Systems, Tübingen
\textsuperscript{2}Gatsby Unit, University College London

February 25, 2020

Abstract

We propose two nonparametric statistical tests of goodness of fit for conditional distributions: given a conditional probability density function \( p(y|x) \) and a joint sample, decide whether the sample is drawn from \( p(y|x)r_x(x) \) for some density \( r_x \). Our tests, formulated with a Stein operator, can be applied to any differentiable conditional density model, and require no knowledge of the normalizing constant. We show that 1) our tests are consistent against any fixed alternative conditional model; 2) the statistics can be estimated easily, requiring no density estimation as an intermediate step; and 3) our second test offers an interpretable test result providing insight on where the conditional model does not fit well in the domain of the covariate. We demonstrate the interpretability of our test on a task of modeling the distribution of New York City’s taxi drop-off location given a pick-up point. To our knowledge, our work is the first to propose such conditional goodness-of-fit tests that simultaneously have all these desirable properties.

1 INTRODUCTION

Conditional distributions provide a versatile tool for capturing the relationship between a target variable and a conditioning variable (or covariate). The last few decades has seen a broad range of modeling applications across multiple disciplines including econometrics in particular \cite{30, 42}, machine learning \cite{14, 40}, among others. In many cases, estimating a conditional density function from the observed data is a one of the first crucial steps in the data analysis pipeline. While the task of conditional density estimation has received a considerable attention in the literature, fewer works have investigated the equally important task of evaluating the goodness of fit of a given conditional density model.

Several approaches that address the task of conditional model evaluation take the form of a hypothesis test. Given a conditional model, and a joint sample containing realizations of both target variables and covariates, test the null hypothesis stating that the model is correctly specified, against the alternative stating that it is not. The model does not specify the marginal distribution of the covariates. We refer to this task as \textit{conditional goodness-of-fit testing}. One of the early nonparametric tests is \cite{1}, which extended the classic Kolmogorov test to the conditional case. Zheng \cite{42} considered the first-order linear expansion of the Kullback-Leibler divergence as the test statistic, and showed that the resulting test is
consistent against any fixed alternative under technical assumptions. The conditional Kolmogorov test however requires estimation of the cumulative distribution function (CDF), and may only be applied to data of low dimension. Zheng’s test involves density estimation as part the test statistic, and test consistency is only guaranteed with a decaying smoothing bandwidth whose rate can be challenging to control. While there are other tests which are more computationally tractable, these tests are only designed for conditional models from a specific family: [30] for structural equation models, [36] for generalized linear models, for instance.

Another line of work which is prominent in econometrics is based on the conditional moment restrictions (CMR). In CMR based tests, the conditional model is specified by a conditional moment function which has an important property that its conditional expectation under the data distribution is zero if the model is correct. This formulation is general, and in fact nests testing a conditional mean regression model as a special case [39, 4]. To guarantee consistency, Bierens and colleagues [6, 5] use a class of weight functions indexed by a continuous nuisance parameter so that an infinite number of moment conditions can be considered, resulting in a powerful test which detects any departure from the null model. Although the conditional moment function can be set to the squared loss between the model output and the target variable in the case of testing the conditional mean of a regression model, specifying the conditional moment function for a conditional density model is challenging, especially for a complex model whose normalizing constant is intractable.

A related thread of development of omnibus tests for model goodness of fit has arisen in the machine learning community recently through the use of kernel methods and Stein operators. The combination of Stein’s identity and kernel methods was investigated in [31] for the purpose of reducing the variance of Monte Carlo integration. Chwialkowski et al. [13] and Liu et al. [28] independently extended [31] to construct a consistent, nonparametric test of goodness of fit of a marginal density model known as the Kernel Stein Discrepancy (KSD) test. The KSD test has proved successful in many applications and has spawned a number of further studies including [15] which considered the KSD for checking the convergence of an MCMC procedure, [41] which extended the KSD test to a discrete domain, and [20, 23] which developed linear-time variants of the KSD. While proven to be powerful, an issue with the KSD is that it is only applicable to marginal (unconditional) density models. To our knowledge, there has been no attempt of extending the KSD test to handle conditional density models.

In the present work, we are interested in constructing omnibus statistics which can detect any departure from the specified conditional model in the null hypothesis. We propose two nonparametric, general conditional goodness-of-fit tests which require no density estimation as an intermediate step. Our first test, the Kernel-Smoothed Stein Discrepancy (KSSD, described in Section 3), generalizes the KSD to conditional goodness-of-fit testing. Briefly, we consider the KSD’s Stein witness function conditioned on the covariate. The KSSD statistic is defined as the norm, in a vector-valued reproducing kernel Hilbert space (RKHS), of the smoothed witness function where the smoothing is over the domain of the covariate. The smoothing operation guarantees that the discrepancy between the conditional model and the data can be detected for any realization of the conditioning variable. We prove that the KSSD test is consistent against any fixed alternative conditional model, for any $C_0$-universal positive definite kernels used; importantly, in the case of Gaussian kernels, the consistency holds regardless of the bandwidth parameter (not necessarily decaying in contrast to [42]).

Our second proposed test, referred to as the Finite Set Conditional Discrepancy (FSCD, described in Section 4), further extends the KSSD test to also return test locations (a set of points) that indicate realizations of the covariate for which the conditional model does not fit well. The FSCD test thus offers an interpretable indication of where the conditional model fails as evidence for rejecting the null hypothesis. Thanks to the Stein operator, our proposed tests do not require the normalizing constant of the conditional model. In experiments on both homoscedastic and heteroscedastic models, we show that the smoothing operation in the KSSD test allows it to detect global differences, whereas the use of test locations in the FSCD makes it more sensitive to local departure from the null model.
2 BACKGROUND

This section gives background materials which will be needed when we propose our new tests: the Kernel-Smoothed Stein Discrepancy (KSSD, Section 3) and the Finite Set Conditional Discrepancy (FSCD, Section 4). We describe two known (unconditional) goodness-of-fit tests: the Kernel Stein Discrepancy (KSD) test of Chwialkowski et al. [13], Liu et al. [28] in Section 2.1, and the Finite Set Stein Discrepancy (FSSD) of Jitkrittum et al. [24] in Section 2.2. We will see in Sections 3 and 4 that our proposed KSSD and FSCD are generalizations of KSD and FSSD, respectively, to the conditional goodness-of-fit testing problem.

2.1 KERNEL STEIN DISCREPANCY (KSD)

Consider probability distributions supported on an open subset $\mathcal{X} \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$. The Kernel Stein Discrepancy (KSD) between probability distributions $P$ and $R$ is a divergence measure defined as $S_P(R) := \sup_{\|f\|_{\mathcal{F}} \leq 1} |E_{x \sim R} T_p f(x) - E_{x \sim P} T_p f(x)|$, where $f \in \mathcal{F}^d$, $\mathcal{F}^d = \mathbb{R}^d \times \mathcal{F}$, and $\mathcal{F}$ is the reproducing kernel Hilbert space (RKHS, [3]) associated with a positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

Key to the KSD is $T_p$, a Stein operator constructed such that the expectation under the distribution $P$ vanishes, i.e., $E_{x \sim P} T_p f(x) = 0$, for any function $f \in \mathcal{F}^d$. For a distribution $P$ admitting a differentiable, strictly positive density $p : \mathcal{X} \to (0, \infty)$, the Langevin Stein operator of differentiable functions defined by $T_p f(x) = s_p(x)^T f(x) + \nabla_x f(x) \in \mathbb{R}^d$ satisfies the aforementioned condition, where $s_p(x) := \nabla_x \log p(x)$ is the score function (under suitable boundary conditions [31, Assumption A2']). Thus, the KSD can be equivalently written as $\sup_{\|f\|_{\mathcal{F}} \leq 1} |E_{x \sim P} T_p f(x)|$.$^1$ It can be shown that if the kernel $k$ is $C_0$-universal [34], and $R$ has a density $r$ such that $E_{x \sim r} \|\nabla_x \log p(x) - \nabla_x \log r(x)\|^2 < \infty$, then $S_p(r) = 0$ if and only if $p = r$ [12, Theorem 2.2].

The KSD can be rewritten in a form that can be estimated easily. Assume that the kernel $k$ is differentiable. Then, for any function $f \in \mathcal{F}^d$, we have $T_p f(x) = \langle f, \xi_p(x, \cdot) \rangle_{\mathcal{F}^d}$ where $\xi_p(x, \cdot) := s_p(x)k(x, \cdot) + \nabla_x k(x, \cdot)$, due to the reproducing property of $k$, where $\langle f, g \rangle_{\mathcal{F}^d} = \sum_{j=1}^d \langle f_j, g_j \rangle_{\mathcal{F}}$ is the inner product on $\mathcal{F}^d$. Assuming Bochner integrability of $\xi_p(x, \cdot)$ as in [13, 28], it follows that

$$S_p(r) = \sup_{f \in \mathcal{F}^d} |\langle f, E_{x \sim r} \xi_p(x, \cdot) \rangle_{\mathcal{F}^d}| = \|g_{p, r}\|_{\mathcal{F}^d},$$

where $g_{p, r} (\cdot) = E_{x \sim r} \xi_p (x, \cdot) \in \mathcal{F}^d$ is the function that achieves the supremum, and is known as the Stein witness function [23]. The squared KSD admits the expression $S_p^2(r) = \|g_{p, r}\|_{\mathcal{F}^d}^2 = E_{x, x' \sim r} h_p(x, x')$ where

$$h_p(x, x') := k(x, x') s_p^T(x) s_p(x') + \sum_{i=1}^d \frac{\partial^2 k(x, x')}{\partial x_i \partial x'_i} + s_p^T(x) \nabla_x k(x, x') + s_p^T(x') \nabla_x k(x, x').$$

Given a sample $\{x_i\}_{i=1}^n \sim r$, the squared KSD has an unbiased estimator $\hat{S}_p^2(r) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j)$, which is a U-statistic [33]. Since the KSD only depends on $p$ through $\nabla_x \log p(x)$, the normalizing constant of $p$ is not required. The squared KSD has been successfully used in [13, 28] as the test statistic for goodness-of-fit testing: given a marginal density model $p$ (known up to the normalizing constant), and a sample $\{x_i\}_{i=1}^n \sim r$, test whether $p$ is the correct model.

$^1$Note that the definition of the KSD does not depend on the existence of a density, as it is defined in terms of an expectation.
2.2 FINITE SET STEIN DISCREPANCY (FSSD)

The Finite Set Stein Discrepancy (FSSD, [23]) is one of several extensions of the original KSD aiming to construct a goodness-of-fit test of an unconditional density model that runs in linear time (i.e., $O(n)$ runtime complexity), and that offers an interpretable test result. Key to the FSSD is the observation that the KSD $S_p(r) = 0$ if and only if $p = r$, assuming conditions described in Section 2.1. As a result, $g_{p,r}$ is a zero function if and only if $p = r$, implying that the departure of $g_{p,r}$ from the zero function can be used to determine whether $p$ and $r$ are the same. In contrast to the KSD which relies on the RKHS norm $\| \cdot \|_{L^2}$, the FSSD statistic evaluates the Stein witness function to check this departure. Specifically, given a finite set $V := \{v_1, \ldots, v_J\} \subset \mathcal{X}$ (known as the set of test locations), the squared FSSD is defined as $\text{FSSD}_p^2(r) := \frac{1}{J} \sum_{j=1}^{J} \| g_{p,r}(v_j) \|^2_2$. It is shown in [23] that if $V$ is drawn from a distribution with a density supported on $\mathcal{X}$, then $\text{FSSD}_p^2(r) = 0$ if and only if $p = r$. The squared FSSD can be estimated in linear time.

An interesting property of the FSSD is that the test locations can be automatically optimized by maximizing the test power of the FSSD statistic. [23] showed that there are two advantages by doing so: firstly, optimizing the test locations increases the test power; secondly, the optimized test locations reveal where $p$ and $r$ differ in the domain $\mathcal{X}$. The latter advantage is what gives FSSD the ability to offer an interpretable test result and justify the rejection of the null hypothesis $H_0 : p = r$.

3 THE KERNEL-SMOOTHED STEIN DISCREPANCY (KSSD)

In this section, we propose our first statistic called the Kernel-Smoothed Stein Discrepancy (KSSD) for distinguishing two conditional probability density functions. In Section 4, we will extend the KSSD test to construct our second test, the Finite Set Conditional Discrepancy (FSCD), that returns an interpretable result indicating a realization $v \in \mathcal{X}$ for which the difference between $p(\cdot | v)$ and $r(\cdot | v)$ can be detected with high probability. We first start with a formal description of the conditional goodness-of-fit testing.

**Problem Setting** Let $X$ and $Y$ be two random vectors taking values in $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. Let $p = p(y|x)$ be a conditional density function of $y$ given $x$ representing a candidate model for modeling the conditional distribution of $y$ given $x$.\(^2\) Given a joint sample $Z_n = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} r_{xy}$ where $r_{xy}(x,y) = r(y|x)r_x(x)$ defined on $\mathcal{X} \times \mathcal{Y}$, the goal of conditional goodness-of-fit testing is to test

$$H_0 : p \overset{r_x}{=} r \quad \text{vs} \quad H_1 : p \overset{r_x}{\neq} r,$$

where we write $p \overset{r_x}{=} r$ if for $r_x$-almost all $x$ and for all $y \in \mathcal{Y}$, $p(y|x) = r(y|x)$. Note that $r_{xy}$ is only observed through the joint sample $Z_n$; and $p$ only specifies the conditional model. That is, $p$ does not specify a marginal model for $x$. Hence, only the difference between $p$ and $r$ (i.e., the conditional densities) can be the basis for a rejection of $H_0$. This subtlety is what distinguishes the conditional goodness-of-fit testing from testing the difference between two joint distributions.

**Vector-valued reproducing kernels** We will require vector-valued reproducing kernels for the construction of our new tests. We briefly give a brief introduction to this concept here. For further details, please see Section 2.2 of Carmeli et al. [10] and Carmeli et al. [11], Caponnetto et al. [9], Kadri et al. [26], Sriperumbudur et al. [34], Szabó and Sriperumbudur [38]. Let $\mathcal{L}(\mathcal{H}; \mathcal{H}')$ be the Banach space of bounded operators from a Hilbert space $\mathcal{H}$ to $\mathcal{H}'$ endowed with the uniform norm. We write $\mathcal{L}(\mathcal{H})$ for $\mathcal{L}(\mathcal{H}; \mathcal{H})$. A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Z})$ is said to be a $\mathcal{Z}$-reproducing kernel if

\(^2\)Note that $p$ and $r$ are conditional density functions from Section 3 onwards, unlike in Section 2 where they are unconditional density functions.
Then we will assume that \((x, z) \in X \times Z\) and \((y, z) \in Z\). The proof can be found in Section A.1.

Let \(L^2(\mathcal{X}; \mathcal{Z})\) be the vector space of continuous functions mapping from \(\mathcal{X}\) to \(\mathcal{Z}\). In this work, we will assume that \(\mathcal{X}\) and \(\mathcal{Z}\) are Banach spaces. Let \(C_0(\mathcal{X}; \mathcal{Z}) \subset C(\mathcal{X}; \mathcal{Z})\) denote the subspace of continuous functions that vanish at infinity i.e., \(\|f(x)\|_Z \to 0\) as \(\|x\| \to \infty\). A \(Z\)-reproducing kernel \(K: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{Z})\) is said to be \(C_0\) if \(F_K\) is a subspace of \(C_0(\mathcal{X}; \mathcal{Z})\) [10, Section 2.3, Definition 1]. A \(C_0\)-kernel \(K\) is said to be universal if \(F_K\) is dense in \(L^2(\mathcal{X}; \mu; \mathcal{Z})\) for any probability measure \(\mu\) [10, Section 4.1].

Let \(l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\) be a positive definite kernel, and \(F_l\) be the associated reproducing kernel Hilbert space (RKHS). Write \(F_l^{d_y} := \times_{y \in Y} F_l\) and define \(\langle a, b \rangle_{F_l^{d_y}} := \sum_{y \in Y} \langle a_y, b_y \rangle_{F_l}\) to be the inner product on \(F_l^{d_y}\) for \(a := (a_1, \ldots, a_{d_y}), b := (b_1, \ldots, b_{d_y}) \in F_l^{d_y}\). Let \(K: \mathcal{X} \times \mathcal{X} \rightarrow F_l^{d_y}\) be a \(F_l^{d_y}\)-reproducing kernel. Let \(k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\) be a real-valued positive definite kernel, and \(F_k\) be the associated RKHS. For brevity, we write \(E_{x,y}\) for \(E_{(x,y) \sim \rho_{x,y}}\). In what follows, we will interchangeably write \(p_{|x}\) and \(p(\cdot|x)\) i.e., the density of \(y\) given a realization \(x\).

**Proposed statistic** Consider the following population statistic defining a discrepancy between \(p\) and \(r\):

\[
D_p(r) := \|E_{(x,y) \sim \rho_{x,y}} K_x \xi_{p|y}(y, \cdot)\|^2_{F_k},
\]

where \(\xi_{p|y}(y, \cdot) := l(y, \cdot) \nabla_y \log p(y|x) + \nabla_y l(y, \cdot) \in F_l^{d_y}\). We refer to \(D_p(r)\) as the Kernel-Smoothed Stein Discrepancy (KSSD). Our first result in Theorem 1 shows that the KSSD is zero if and only if \(p \equiv r\).

**Theorem 1** \((D_p(r)\) distinguishes conditional density functions). Let \(K: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(F_l^{d_y})\) and \(l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\) be positive definite kernels. Define \(g_{p,r}(w|x) := E_{y \sim \rho_{y|x}} \xi_{p|y}(y, w) \in F_l^{d_y}\). Assume the following assumptions:

1. \(K\) and \(l\) are \(C_0\)-universal;
2. \(r_x\)-ess \(\sup_x E_{y \sim r(x|y)} \|\nabla_y \log \frac{p(y|x)}{r(y|x)}\|^2 < \infty\);
3. \(\int_X \|g_{p,r}(\cdot|x)\|^2_{F_l^{d_y}} r_x(x) \, dx < \infty\);
4. \(E_{x,y} \|K_x \xi_{p|y}(y, \cdot)\|^2_{F_k} < \infty\);

Then \(D_p(r) = 0\) if and only if \(p \equiv r\) i.e., for \(r_x\)-almost all \(x \in \mathcal{X}\), \(p(\cdot|x) = r(\cdot|x)\).

**Proof (sketch).** The idea is to rewrite (1) into a form that involves the Stein witness function (as described in Section 2) \(g_{p,r}(\cdot|x)\) between \(p(\cdot|x)\) and \(r(\cdot|x)\). It then amounts to showing that \(g_{p,r}(\cdot|x)\) is a zero function for \(r_x\)-almost all \(x\). This is done by applying the integral operator \(f_x \rightarrow \int K_x f_x r_x(x) \, dx\) on \(g_{p,r}(\cdot|x)\) to incorporate \(r_x\)-almost all \(x\). The result is \(G_{p,r} = \int K_x g_{p,r}(\cdot|x) r_x(x) \, dx\). Since \(K\) is \(C_0\)-universal, this operator is injective, implying \(G_{p,r}\) is zero if and only if \(g_{p,r}(\cdot|x)\) is a zero function for \(r_x\)-almost all \(x\). But, \(G_{p,r} = E_{(x,y) \sim \rho_{x,y}} K_x \xi_{p|y}(y, \cdot)\). Thus, taking the norm gives (1). The complete proof can be found in Section A.1.
In the proof sketch, one can see the application of the integral operator as smoothing the Stein witness function \( g_{p,r}(\cdot|x) \) with the kernel \( K \), giving rise to the name Kernel-Smoothed Stein Discrepancy. Theorem 1 states that the population statistic in (1) distinguishes two conditional density functions under regularity conditions given above. In particular, it is required that the two kernels \( K \) and \( l \) are \( C_0 \)-universal. Examples of a real-valued \( C_0 \)-universal kernels are the Gaussian kernel \( l(y,y') := \exp \left( -\frac{\|y-y'\|^2}{2\sigma^2} \right) \in \mathbb{R} \), Laplace kernel, and the inverse multiquadrics kernel [34, p. 2397]. An example of a \( \mathcal{F}_d^{I} \)-reproducing kernel \( K \) which is \( C_0 \)-universal is given by \( K(x,x') = k(x,x')I \) where \( k \) is a real-valued \( C_0 \)-universal kernel, and \( I \in \mathcal{L}(\mathcal{F}_d^{I}) \) is the identity operator.\(^3\) For simplicity, in this work, we will assume a kernel \( K \) that takes this form.

### 3.1 HYPOTHESIS TESTING WITH KSSD

To construct a statistical test for conditional goodness of fit, we start by rewriting \( D_p(r) \) in (1) in a form that can be estimated easily. Assume that \( K(x,x') := k(x,x')I \) for a positive definite kernel \( k: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \). Define \( s_p(y|x) := \nabla_y \log p(y|x) \). We have

\[
D_p(r) = \left\langle E_{xy}K_{x}\xi_{p|x}(y,\cdot), E_{x'y}K_{x'}\xi_{p|x'}(y',\cdot) \right\rangle_{\mathcal{F}_K}
\]

\[= \left\langle E_{xy}E_{x'y} K_{x}\xi_{p|x}(y,\cdot), E_{x'y}K_{x'}\xi_{p|x'}(y',\cdot) \right\rangle_{\mathcal{F}_K}
\]

\[= \left\langle E_{xy}E_{x'y} K_{x}K_{x'}\xi_{p|x}(y,\cdot), \xi_{p|x'}(y',\cdot) \right\rangle_{\mathcal{F}_d^{I}}
\]

\[= E_{xy}E_{x'y} h_p((x,y),(x',y')), \quad (2)
\]

where at \((a)\) the expectation and the inner product commute because of Bochner integrability of \((x,y) \to K_x\xi_{p|x}(y,\cdot)\) (see assumption 4 in Theorem 1, and Steinwart and Christmann [35, Definition A.5.20]), at \((b)\) we use the adjoint \( K^*_{x'} \) and the reproducing property i.e., \( K^*_{x'}K_x = K(x,x') = k(x,x')I \),

\[h_p((x,y),(x',y')) := \left\langle \xi_{p|x}(y,\cdot), \xi_{p|x'}(y',\cdot) \right\rangle_{\mathcal{F}_d^{I}}
\]

\[= l(y,y')s_{p}^\top(y|x)s_{p}(y'|x') + \sum_{i=1}^{d_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(y,y')
\]

\[+ s_{p}^\top(y|x) \nabla_y l(y,y') + s_{p}^\top(y'|x') \nabla_y l(y,y'), \quad (3)
\]

and \( s_{p}(y|x) := \nabla_y \log p(y|x) \). With \( H_p((x,y),(x',y')) := k(x,x')h_p((x,y),(x',y')) \), given an i.i.d. sample \( \{ (x_i,y_i) \}_{i=1}^{n} \sim r_{xy} \), an unbiased, consistent estimator for (2) is given by

\[\hat{D}_p := \frac{1}{n(n-1)} \sum_{i \neq j} H_p((x_i,y_i),(x_j,y_j)), \quad (4)
\]

which is a second-order U-statistic with \( H_p \) as the U-statistic kernel [33, Section 5], and can be computed easily. It is clear from (3) that the KSSD statistic (both population and its estimator) depends on the model \( p \) only through \( \nabla_y \log p(y|x) = \nabla_y \log p(y,x) \) which is independent of the normalizer \( p(x) \). The fact that the KSSD does not require the normalizer is a big advantage since modern conditional models tend to be complex and their normalizers may not be tractable. A consequence of being a U-statistic is that its asymptotic behaviors can be derived straightforwardly, as given in Proposition 2.

\(^3\)By Carmeli et al. [10, Example 14], if \( k \) is a \( C_0 \)-scalar reproducing kernel, and \( B \) is a positive operator, then the kernel \( K = kB \) is \( C_0 \)-universal if and only if \( k \) is \( C_0 \)-universal and \( B \) is injective. See also Carmeli et al. [10, Section 3.3, Proposition 9].
Proposition 2 (Asymptotic distributions of $\hat{D}_p$). Assume all conditions in Theorem 1 and assume that $E_{x,y}E_{x',y'} H_p^2((x, y), (x', y')) < \infty$. Then,

1. Under $H_0$, $n\hat{D}_p \overset{d}{\to} \sum_{j=1}^{\infty} \lambda_j \chi_j^2 - 1$, where $\{\chi_j^2\}_j$ are independent $\chi_1^2$ random variables, $\lambda_j$ are eigenvalues of the operator $A$ defined as $A\varphi(z) = \int H_p(z', z') \varphi(z') dz'$ for non-zero $\varphi$, $z := (x, y)$ and $z' := (x', y')$;

2. Under $H_1$, $\sqrt{n}(\hat{D}_p - D_p(r)) \overset{d}{\to} N(0, \sigma^2_{H_1})$ where $\sigma^2_{H_1} := 4V[E_{x,y}H_p((x, y), (x', y'))]$.

A proof of Proposition 2 can be found in Section A.2 (appendix). Proposition 2 suggests that under $H_0$, $n\hat{D}_p$ converges to a limit distribution given by an infinite weighted sum of chi-squared random variables. Under $H_1$, for any fixed $p$ and $r$, we have $n\hat{D}_p = \mathcal{O}(\sqrt{n})$, which diverges to $+\infty$. The behaviors are common in many recently developed nonparametric tests [41, 13, 28, 16, 18]. A consistent test that has an asymptotic false rejection rate no larger than a specified significance level can be constructed by setting the rejection threshold (critical value) to be $\gamma_{1-\alpha}$.

In our work, we use the bootstrap procedure of Arcones and Gine [2], Huskova and Janssen [21] (with a V-statistic estimator). To generate a bootstrap sample, we draw $w_1, \ldots, w_n \sim \text{Multinomial}(n; \frac{1}{n}, \ldots, \frac{1}{n})$, define $\tilde{w}_i := \frac{1}{n}(w_i - 1)$, and compute

$$\hat{D}_p^* = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{w}_i \tilde{w}_j H((x_i, y_i), (x_i, y_j)). \quad (5)$$

By bootstrapping $m$ times to generate $\hat{D}_{p1}^*, \ldots, \hat{D}_{pm}^*$, the test threshold can be estimated by computing the empirical $(1 - \alpha)$-quantile of these bootstrapped samples. The overall computational cost of this bootstrap procedure is $O(mn^2)$, which is the same cost as testing a marginal probability model in Chwialkowski et al. [13], Liu et al. [28], Yang et al. [41].

4 THE FINITE SET CONDITIONAL DISCREPANCY (FSCD)

In this section, we extend the KSSD statistic presented in Section 3 to enable it to also pinpoint the location(s) in the domain of $X$ that best distinguish $p(\cdot|x)$ and $r(\cdot|x)$. The result is a goodness-of-fit test for conditional density functions which can have an interpretative output (in terms of locations in $X$) to justify a rejection of the null hypothesis.

We start by noting that Theorem 1 and (1) implies that $G_{p,r}: \mathcal{X} \to \mathcal{F}_{\mathcal{I}}^{d_r'}$ defined as $G_{p,r}(v) := [E_{(x,y) \sim r_{x,y}} K_{x} \ell_{p,r}(y, \phi)](v) \in \mathcal{F}_{\mathcal{I}}^{d_r'}$ is a zero function if and only if $p \overset{d}{=} r$, under the conditions described in the theorem statement. Note that the KSSD $D_p(r) = \|G_{p,r}\|_{\mathcal{F}_{\mathcal{I}}}^{d_r'}$. For a fixed $v \in \mathcal{X}$, the function $v \mapsto \frac{1}{d_v} \|G(v)\|_{\mathcal{F}_{\mathcal{I}}}^{d_r'} \geq 0$ can be seen as quantifying the extent to which $p$ and $r$ differ, as measured at $v \in \mathcal{X}$; that is, the higher $\frac{1}{d_v} \|G(v)\|_{\mathcal{F}_{\mathcal{I}}}^{d_r'}$, the larger the discrepancy between $p|_v$ and $r|_v$. Inspired by Jitkrittum et al. [24], one can thus construct a variant of the KSSD statistic as follows. Given a set of $J$ test locations $V := \{v_j\}_{j=1}^{J} \subset \mathcal{X}$, we evaluate $G_{p,r}(v)$ at these locations instead of taking the
norm $\| \cdot \|_{\mathcal{F}}$ [22, 23, 24, 32]. More formally, we propose a statistic defined as

$$T_p^V(r) := \frac{1}{Jd_y} \sum_{i=1}^{J} \|G_{p,r}(v_i)\|_{\mathcal{F}}^2,$$  \hspace{1cm} (6)

which we refer to as the Finite Set Conditional Discrepancy (FSCD). Later in Section 4.2, we will describe how $V$ can be automatically optimized by maximizing the test power of the FSCD test. The optimized test locations in $V$ are interpretable in the sense that they specify points $\{v_j\}_{j=1}^J$ in $\mathcal{X}$ that best reveal the differences between the two conditional density functions. For the purpose of describing the statistic in this section, we assume that $V$ is given. We first show in Theorem 3 that the FSCD almost surely distinguishes two conditional probability density functions.

**Theorem 3.** Assume all conditions in Theorem 1. Further assume that $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is a connected open set, and $K(x,x') = k(x,x')I$ where $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a real analytic kernel i.e., for any $x \in \mathcal{X}, v \mapsto k(x,v)$ is a real analytic function. Then, for any $J \in \mathbb{N}$, the following statements hold:

1. Under $H_0$, $T_p^V(r) = 0$ for any $V = \{v_j\}_{j=1}^J \subseteq \mathcal{X}$.

2. Under $H_1$, if $v_1, \ldots, v_J$ in $V$ are drawn from a probability density $\eta$ whose support is $\mathcal{X}$, then $\eta$-almost surely $T_p^V(r) > 0$.

**Proof.** A proof can be found in Section A.3. \hspace{1cm} \Box

Theorem 3 states that given $p$ and $r$, $T_p^V(r) = 0$ if and only if $p \equiv r$ when $V$ is drawn from any probability density supported on $\mathcal{X}$. The core idea is that $\|G_{p,r}(v)\|_{\mathcal{F}}^2$ is a real analytic function of $v$ if $k$ is a real analytic kernel. It is known that the set of roots of a non-zero real analytic function has zero Lebesgue measure [29]. So, pointwise evaluations at the $J$ random test locations suffice to check whether $G_{p,r}$ is a zero function, and the result follows. The FSCD statistic in (6) can thus be seen as quantifying the average discrepancy between $p(\cdot | x)$ and $r(\cdot | x)$ as measured at the locations $x \in V$.

### 4.1 HYPOTHESIS TESTING WITH FSCD

To estimate $T_p^V(r)$ in (6), we first rewrite $\|G_{p,r}(v)\|_{\mathcal{F}}^2$ as

$$\|G_{p,r}(v)\|_{\mathcal{F}}^2 = \| [E(x,y)\rightarrow_{rv} K_{x}(x,y) ] (v) \|_{\mathcal{F}}^2$$

$$\overset{(a)}{=} \left\| E(x,y)\rightarrow_{rv} K_{x}(x,y) \xi_{p,a}(y,\cdot) \right\|_{\mathcal{F}}^2$$

$$\overset{(b)}{=} E_{xy} E_{x'y'} k(x,v) k(x',v) h_{p}((x,y),(x',y'))$$ \hspace{1cm} (7)

where at (a) we use $(K_x f)(v) = K(x,v) f$ for $f \in \mathcal{F}^0_{x}$, and at (b) we use $h_{p}((x,y),(x',y')) = \left\langle \xi_{p,a}(y,\cdot), \xi_{p,a}(y',\cdot) \right\rangle_{\mathcal{F}}$ as in (2). It follows that

$$T_p^V(r) = E_{xy} E_{x'y'} H_p^V((x,y),(x',y'))$$ \hspace{1cm} (8)

where $(x,y),(x',y')$ are i.i.d. random variables following $r_{xy}$,

$$H_p^V((x,y),(x',y')) := \frac{1}{d_y} \bar{K}_V(x,x') h_{p}((x,y),(x',y'))$$,

where $\bar{K}_V(x,x')$ is an arbitrary function of $x$ and $x'$.
and $k_V(x, x') := \frac{1}{n} \sum_{i=1}^{n} k(x, y_i) k(x', y_i)$ is a kernel that depends on $V$. Similarly to (4), an unbiased estimator of $T_p^V$ is given by a second-order U-statistic:

$$
\hat{T}_p^V := \frac{1}{n(n-1)} \sum_{i \neq j} \Pi_p^V((x_i, y_i), (x_j, y_j)).
$$

It is clear from (8) and the definition of $\hat{T}_p^V$ that the FSCD statistic is in fact a special case of the KSSD with the kernel $k$ in (2) replaced with $k_V$. For this reason, the asymptotic distributions of $\hat{T}_p^V$ under both $H_0$ and $H_1$ are almost identical to those of the KSSD. We omit the result here and present it in Proposition 7 in the appendix. Since $\hat{T}_p^V$ is also a degenerate U-statistic, the test threshold can be obtained by bootstrapping with weights drawn from the multinomial distribution as in the case of the KSSD.

### 4.2 FSCD WITH OPTIMIZED TEST LOCATIONS

While Theorem 3 guarantees that the FSCD can distinguish two conditional density functions with any $V$ drawn from any probability density supported on $\mathcal{X}$, in practice, optimizing $V$ will further increase the power of the test, and allow us to interpret $V$ as the locations in $\mathcal{X}$ for which the difference between $p(\cdot | x)$ and $r(\cdot | x)$ can be detected with largest probability. Inspired by the recent approaches of [22, 37, 25, 32], we propose optimizing the test locations in $V$ by maximizing the asymptotic test power of the test statistic $\hat{T}_p^V$. The test power is defined as the probability of rejecting $H_0$ when it is false. We start by giving the expression for the asymptotic test power of $\hat{T}_p^V$ in Corollary 4. For brevity, we write $T_p^V$ for $\hat{T}_p^V(r)$.

**Corollary 4.** Assume that $H_1$ holds. Given a set $V$ of test locations, and a rejection threshold $\gamma \in \mathbb{R}$, the test power of the FSCD test is

$$
P\left(\hat{T}_p^V > \gamma\right) \approx \Phi\left(\frac{\sqrt{n}T_p^V}{\sigma_V} - \frac{\gamma}{\sqrt{n} \sigma_V}\right)
$$

for sufficiently large $n$, where $\Phi$ is the CDF of the standard normal distribution, and $\sigma_V = \sqrt{4V[\mathbb{E}_{xy}[\Pi_p^V((x, y), (x', y'))]]}$ is the standard deviation of the distribution of $\hat{T}_p^V$ under $H_1$.

The result directly follows from the fact that $\hat{T}_p^V$ is asymptotically normally distributed (see Proposition 7 in the appendix). Following the same line of reasoning as in Jitkrittum et al. [24], Sutherland et al. [37], for large $n$, the power expression is dominated by $T_p^V / \sigma_V$, which is called the power criterion [24].

Assume that $n$ is sufficiently large. It follows that finding the test locations $V$ which maximize the test power amounts to finding $V^* = \arg\max_V P\left(\hat{T}_p^V > \gamma\right) \approx \arg\max_V T_p^V / \sigma_V$. We also use the same objective function to tune the two kernels $k$ and $l$.

To optimize, we split the data into two independent sets: training and test sets. We then optimize this ratio with its consistent estimator $\hat{T}_p^V / \hat{\sigma}_V$ estimated from the training set. The hypothesis test is performed on the test set using the optimized parameters. Indeed, this data splitting scheme has also been used in several modern statistical tests [22, 37, 25, 32]. There are two reasons for doing so: firstly, conducting a test on an independent test set avoids overfitting to the training set — the false rejection rate of $H_0$ may be higher than the specified significance level $\alpha$ otherwise; secondly, for the statistic to be a U-statistic, its U-statistic kernel (i.e., $\hat{T}_p^V$) must be independent of the samples used to estimate the summands (see (9)). In Section 5, we shall see that finding $V$ in this way leads to a higher test power when the difference between $p$ and $r$ is localized.
Figure 1: The power criterion of FSCD as a function of $x$ is high where the difference between $p(y|x)$ and $r(y|x)$ can be best detected.

5 EXPERIMENTS

In this section, we empirically investigate the two proposed tests on a number of problems. We divide the empirical study into three parts: 1) illustration of the FSCD power criterion (Section 4.2), 2) test power, and 3) interpretable FSCD power criterion applied to the New York City taxi data modeling problem.

1. Illustration of the FSCD power criterion Our first task is to illustrate that the power criterion of the proposed FSCD test reveals where $p$ and $r$ differ in the domain of the conditioning variable ($x$).

We consider a simple univariate problem where the model is $p(y|x) := \mathcal{N}(x/2, 1)$, the data generating distribution is $r(y|x) := \mathcal{N}(x, 1)$, and $r_x(x) = \mathcal{N}(0, 1)$. We use Gaussian kernels for both $k$ and $l$. The power criterion function is shown in Figure 1.

2. Test power We investigate the test power of the following methods:

- **KSSD**: our proposed KSSD test using Gaussian kernels where the bandwidths are chosen by the median heuristic. Specifically, we use $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma_x^2}\right)$ and $l(y, y') = \exp\left(-\frac{\|y-y'\|^2}{2\sigma_y^2}\right)$ where the bandwidths are set with $\sigma_x := \text{median} \left(\|x_i - x_j\|^2\right)_{i,j=1}^n$ and $\sigma_y := \text{median} \left(\|y_i - y_j\|^2\right)_{i,j=1}^n$. This heuristic has been used to set the bandwidth in many existing kernel-based tests [18, 8, 28, 13].

- **FSCD**: our proposed FSCD test using Gaussian kernels for $k$ and $l$. There are two variations of the FSCD. In FSCD-rand, the $J$ test locations are randomly drawn from a Gaussian distribution fitted to the data with maximum likelihood. In the second variant FSCD-opt, 30% of the observed data are used for optimizing the two bandwidths and the $J$ test locations by maximizing the power criterion, and the rest 70% of the data are used for testing. The data splitting is to guarantee the independence between the training and test sets and is a standard procedure for optimizing hyperparameters of a test [19, 37]. All parameters of FSCD-opt are optimized jointly with Adam [27] with default parameters implemented in Pytorch. We consider $J \in \{1, 5\}$.

- **MMD**: the Maximum Mean Discrepancy (MMD) test [18]. The MMD test was originally created for two-sample testing. Here, we adapt it to conditional goodness-of-fit testing by splitting the data into two disjoint sets $\{(x_i^{(1)}, y_i^{(1)})\}_{i=1}^{n/2}$ and $\{(x_i^{(2)}, y_i^{(2)})\}_{i=1}^{n/2}$ of equal size $n/2$. We then
We report the rejection rates of these tests on all the three problems in Figure 2, where we observe that all the tests correctly have their false rejection rates no larger than $\alpha$ (up to sampling noise). While FSCD-opt has less test data since 30% of locations are randomized, and may be far from $c$ performance of FSCD-rand (random test locations) is significantly lower than FSCD-opt, since the test are effective in identifying where to pinpoint to difference in $X$. In this case, $H_0$ is true. In Figure 2a, we observe that all the tests correctly have their false rejection rates no larger than $\alpha = 0.05$. (b) $H_1$ is true. FSCD-opt is good for detecting local difference. (c) KSSD is good for detecting global difference.

These methods are tested on the following problems:

- **Linear Gaussian Model (LGM):** In this problem, $(x, y) \in \mathbb{R}^3 \times \mathbb{R}$ and we set $p(y|x) = \mathcal{N}\left(\sum_{i=1}^5 ix_i, 1\right)$, set $r := p$ and $r_x(x) = \mathcal{N}(0, I)$. This is a problem where $H_0$ is true.

- **Heteroscedastic Gaussian Model (HGM):** $(x, y) \in \mathbb{R}^3 \times \mathbb{R}$ and $p(y|x) = \mathcal{N}\left(\sum_{i=1}^3 x_i, \sigma^2(x)\right)$ where $\sigma^2(x) := 1 + 10\exp\left(-\frac{\|x-c\|^2}{2\times0.8}\right)$ and $c = \frac{2}{3}1$. We set the observation model to be $r(y|x) = \mathcal{N}\left(\sum_{i=1}^3 x_i, 1\right)$ and set $r_x(x) = \mathcal{N}(0, I)$. In this problem, the observations are drawn from $r$ given by a linear Gaussian model with unit variance. The model $p$ is heteroscedastic (i.e., the noise depends on $x$) where the variance function is created such that it is roughly 1 everywhere in the domain of $x$, except in the region near $c$. This problem is challenging since the difference is local in the domain of $x$. In this case, $H_1$ is true.

- **Quadratic Gaussian Model (QGM):** $(x, y) \in \mathbb{R} \times \mathbb{R}$ and we define $p(y|x) = \mathcal{N}\left(x + 1, 1\right)$, $r(y|x) = \mathcal{N}\left(0.1x^2 + x + 1, 1\right)$, and $r_x(x) = \text{Uniform}(-2, 2)$. Here, the conditional mean of the true distribution $r$ is given by a quadratic function, whereas the model $p$ is linear. This simulates a typical scenario where the model is too simplistic to model the data. Note that the the quadratic term carries a small weight of 0.1, making the difference between $p$ and $r$ challenging to detect. In this case, $H_1$ is true.

We report the rejection rates of these tests on all the three problems in Figure 2, where we conduct 200 independent trials for each experiment with the significance level set to $\alpha = 0.05$. In Figure 2a, we observe that all the tests correctly have their false rejection rates no larger than $\alpha = 0.05$ (up to sampling noise) since $H_0$ is true in the LGM problem. In the HGM problem (Figure 2b) where the difference between $p$ and $r$ is local in the domain $X$, we observe the optimized test locations of FSCD-opt are effective in identifying where to pinpoint to difference in $X$. This can be seen by noting that the performance of FSCD-rand (random test locations) is significantly lower than FSCD-opt, since the test locations are randomized, and may be far from $c$ which specifies the neighborhood that reveals the difference (see the specification of the HGM problem). While FSCD-opt has less test data since 30% of
3. Interpretable test with FSCD  In our final experiment, we show with real data that the power criterion of the FSCD, as a function of $v \in \mathcal{X}$ offers an interpretable indication of where the conditional model does not fit the data well. We train a Mixture Density Network (MDN, Bishop [7, Section 5.6]) on the New York City (NYC) taxi dataset. The dataset contains millions of trip records that include pick-up locations, drop-off locations, time, etc. The MDN models the conditional probability of the drop-off location $y$ given a pick-up location $x$, expressed as a latitude/longitude coordinate (i.e., $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^2$). We train the model on five million trip records of yellow cabs from January 2015 using 20 Gaussian components, and a ReLU-based architecture for the mean, mixing proportion, and variance functions.
For simplicity, only trips with pick-up and drop-off locations within or close to Manhattan are used. More details on the MDN and its training can be found in Section B in the appendix.

We use Gaussian kernels for both $k$ and $l$ with their bandwidths chosen by the median heuristic, and compute the power criterion of the FSCD test at two separate test locations, using a held-out data of size 4000. The results are shown in Figure 3 where blue points indicate observed drop-off locations conditioned on the pick-up location denoted by $\triangle$. We consider conditioning separately on two pick-up locations $\triangle_1$ and $\triangle_2$, shown in Figure 3a and Figure 3b, respectively.

In Figure 3a, $p(y|x = \triangle_1)$ fits relatively well to the data compared to $p(y|x = \triangle_2)$ shown in 3b. In Figure 3b, the observed data (blue) do not respect the multimodality suggested by the model. As a result, the power criterion evaluated at $\triangle_2$ is higher, indicating a poorer fit at $\triangle_2$. This suggests that the power criterion function of the FSCD gives an interpretable indication for where the conditional model does not fit well.

6 CONCLUSION

We have proposed two novel conditional goodness-of-fit tests: the Kernel-Smoothed Stein Discrepancy (KSSD), and the Finite Set Conditional Discrepancy (FSCD). We prove that the population statistics of the two test define a proper divergence measure between two conditional density functions. In experiments, we show that the test locations found by optimizing the test power of the FSCD are interpretable and useful for identifying the region in the domain of the covariate for which the model does not fit well. There are several possible future directions. Both KSSD and FSCD can be extended to handle a discrete domain $Y$ by considering a Stein operator defined in terms of forward and backward differences as in [41]. Further, our two tests can be sped up to have a runtime complexity linear in the sample size (instead of quadratic in the current version) by considering random Fourier features as in [20]. We leave these research directions as future work.

Acknowledgment

We thank Patsorn Sangkloy for helping us with the experiment on the NYC taxi dataset. Heishiro Kanagawa thanks the Gatsby Charitable Foundation for the financial support.

References

[1] Donald W. K. Andrews. A conditional Kolmogorov test. Econometrica, 65(5):1097–1128, 1997.

[2] Miguel A Arcones and Evarist Gine. On the bootstrap of U and V statistics. The Annals of Statistics, pages 655–674, 1992.

[3] Alain Berlinet and Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media, 2011.

[4] Herman J Bierens. Consistent model specification tests. Journal of Econometrics, 20(1):105 – 134, 1982. ISSN 0304-4076.

[5] Herman J Bierens. A consistent conditional moment test of functional form. Econometrica: Journal of the Econometric Society, pages 1443–1458, 1990.

[6] Herman J Bierens and Werner Ploberger. Asymptotic theory of integrated conditional moment tests. Econometrica: Journal of the Econometric Society, pages 1129–1151, 1997.

[7] Christopher M Bishop. Pattern recognition and machine learning. Springer, 2006.
[8] Wacha Bounliphone, Eugene Belilovsky, Matthew B Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. In ICLR, 2015.

[9] Andrea Caponnetto, Charles A Micchelli, Massimiliano Pontil, and Yiming Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(Jul):1615–1646, 2008.

[10] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *arXiv e-prints*, Jul 2008.

[11] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

[12] Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

[13] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, pages 2606–2615, 2016.

[14] Vincent Dutordoir, Hugh Salimbeni, James Hensman, and Marc Deisenroth. Gaussian process conditional density estimation. In *NeurIPS*, pages 2385–2395, 2018.

[15] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, pages 1292–1301, 2017.

[16] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *NeurIPS*, pages 585–592, 2008.

[17] Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *NeurIPS*, pages 673–681. 2009.

[18] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[19] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *NeurIPS*, pages 1205–1213, 2012.

[20] Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *NeurIPS*, pages 1899–1909, 2018.

[21] Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate u-statistics. *Ann. Statist.*, 21(4):1811–1823, 12 1993.

[22] Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *NeurIPS*, pages 181–189, 2016.

[23] Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *ICML*. 2017.

[24] Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *NeurIPS*, 2017.

[25] Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In *NeurIPS*, pages 808–819, 2018.

[26] Hachem Kadri, Emmanuel Duflot, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.
[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[28] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In ICML, pages 276–284, 2016.

[29] Boris Mityagin. The zero set of a real analytic function. arXiv e-prints, art. arXiv:1512.07276, Dec 2015.

[30] Marcelo J. Moreira. A conditional likelihood ratio test for structural models. Econometrica, 71(4):1027–1048, 2003.

[31] Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695–718, 2017. doi: 10.1111/rssb.12185.

[32] Meyer Scetbon and Gael Varoquaux. Comparing distributions: L1 geometry improves kernel two-sample testing. In NeurIPS, pages 12306–12316. 2019.

[33] Robert J. Serfling. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, 2009.

[34] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. Journal of Machine Learning Research, 12:2389–2410, 2011.

[35] Ingo Steinwart and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

[36] Winfried Stute and Li-Xing Zhu. Model checks for generalized linear models. Scandinavian Journal of Statistics, 29(3):535–545, 2002. ISSN 03036898, 14679469.

[37] Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In ICLR. 2016.

[38] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. Journal of Machine Learning Research, 18(233):1–29, 2018.

[39] Gautam Tripathi, Yuichi Kitamura, et al. Testing conditional moment restrictions. The Annals of Statistics, 31 (6):2059–2095, 2003.

[40] Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. The Journal of Machine Learning Research, 17(1):7184–7220, 2016.

[41] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In ICML, pages 5561–5570, 2018.

[42] John Xu Zheng. A consistent test of conditional parametric distributions. Econometric Theory, 16(5):667–691, 2000.
A PROOFS

This section contains proofs of the theoretical results we gave in the main text. We first give two known lemmas that will be needed.

**Lemma 5** (10, Theorem 2b, Section 4 (rephrased)). Let $\mathcal{X}$ be a locally compact second countable topological space, and $\mathcal{Z}$ be a complex separable Hilbert space. Let $K: \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Z})$ be a $C_0$ universal kernel associated with the vector-valued RKHS $\mathcal{F}_K$, where $\mathcal{L}(\mathcal{Z})$ denotes the Banach space of bounded operators from $\mathcal{Z}$ to $\mathcal{Z}$. Let $P$ be a probability measure on $\mathcal{X}$. Then, the operator $L_P: L^2(\mathcal{X}, P; \mathcal{Z}) \to \mathcal{F}_K$ given by $(L_P f)(t) = \int_{\mathcal{X}} K(t, x) f(x) \, dP(x)$ is injective, for all $f \in L^2(\mathcal{X}, P; \mathcal{Z})$.

**Lemma 6** (Chwialkowski et al. 12, Lemma 1). If $k: \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is a bounded, real analytic kernel (i.e., for any $v \in \mathcal{X}$, $x \mapsto k(x, v)$ is a real analytic function), then all functions in the RKHS defined by $k$ are real analytic.

A.1 PROOF OF THEOREM 1

Recall the theorem:

**Theorem 1** ($D_P(r)$ distinguishes conditional density functions). Let $K: \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F}_I^{d_y})$ and $l: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be positive definite kernels. Define $g_{p,r}(w|x) := \mathbb{E}_{y \sim r_x} \xi_{p|x} (y, w)$, $w \in \mathbb{R}^{d_y}$. Assume the following assumptions:

1. $K$ and $l$ are $C_0$-universal;
2. $r_x$-ess sup $\mathcal{X} \mathbb{E}_{y \sim r(x|\cdot)} \left\| \nabla_y \log p(y|x) \right\|^2 < \infty$;
3. $\int_{\mathcal{X}} ||g_{p,r}(\cdot|x)||^2 \, d\nu(x) < \infty$.
4. $\mathbb{E}_x \mathbb{E}_{y|x} ||K_x \xi_{p|x}(y, \cdot)||_{\mathcal{F}_K} < \infty$.

Then $D_P(r) = 0$ if and only if $p \equiv r$ i.e., for $r_x$-almost all $x \in \mathcal{X}$, $p(\cdot|x) = r(\cdot|x)$

**Proof.** We first rewrite the statistic as

$$D_P^2(r) = \left\| \mathbb{E}_{x,y} K_x \xi_{p|x}(y, \cdot) \right\|^2_{\mathcal{F}_K}$$

$$= \left\| \mathbb{E}_{x,y} K_x \xi_{p|x} \mathbb{E}_{r(x,y)}(y, \cdot) \right\|^2_{\mathcal{F}_K}$$

$$= \left\| \mathbb{E}_{x,y} K_x g_{p,r}(\cdot|x) \right\|^2_{\mathcal{F}_K},$$

where $g_{p,r}(w|x) := \mathbb{E}_{y \sim r_x} \xi_{p|x} (y, w), w \in \mathbb{R}^{d_y}$ is the Stein witness function between $p|x$ and $r|x$, and $\xi_{p|x}(y, \cdot) := l(y, \cdot) \nabla_y \log p(y|x) + \nabla_y l(y, \cdot) \in \mathcal{F}_I^{d_y}$ for $r_x$-almost all $x$ [13, 28, 24]. By Chwialkowski et al. [13, Theorem 2.2], for $r_x$-almost all $x \in \mathcal{X}$, the Kernel Stein Discrepancy (KSD) between the two probability density functions $p|x$ and $r|x$ is 0 if and only if they coincide. That is, given $x \sim r_x$,
\[ KSD_{p,x}^2 (r|x) = 0 = \| \mathbf{g}_{p,r}(\phi|x) \|_{L^2_{p,x}}^2 \] if and only if \( p|x = r|x \). Thus, proving the claim amounts to showing \( \mathbf{g}_{p,r}(\phi|x) = 0 \) for \( r|x \)-almost all \( x \) if and only if \( p \leq r \). Since \( \mathbf{g}_{p,r} \in L^2(\mathcal{X}, r_x; F_{p,x}^d) \) and \( K \) is \( C_0 \)-universal, Lemma 5 (by setting \( Z = F_{p,x}^d \)) implies that the map \( \mathbf{g}_{p,r} \mapsto E_{x \sim r_x} K_x \mathbf{g}_{p,r}(\phi|x) \) is injective. As a result of the injectivity and the fact that \( A0 = 0 \) if \( A \) is a linear operator, we have \( E_{x \sim r_x} K_x \mathbf{g}_{p,r}(\phi|x) = 0 \) if and only if \( \mathbf{g}_{p,r} = 0 \) or equivalently \( \mathbf{g}_{p,r}(\phi|x) = 0 \) for all \( r|x \)-almost all \( x \).

\[ \square \]

### A.2 PROOF OF PROPOSITION 2

Define \( \zeta_1 := \mathbb{V} \left[ E_{(x,y) \sim r_x} H_p((x,y), (x', y')) \right] \). We only need to show that under \( H_0, \hat{D}_p \) is a degenerate \( U \)-statistic i.e., \( \zeta_1 = 0 \), and under \( H_1, \hat{D}_p \) is non-degenerate i.e., \( \zeta_1 > 0 \). Then, the asymptotic distributions in the two cases follow from Serfling [33, Section 5.5].

**Case: \( H_0 \) is true**

\[ E_{(x,y) \sim r_x} H_p((x,y), (x', y')) = E_{(x,y) \sim r_x} \left( K_x \xi_{p,x}(y, \phi), K_x \xi_{p,x}(y', \phi) \right)_{F_K} \]

\[ \equiv \left( E_{(x,y) \sim r_x} K_x \xi_{p,x}(y, \phi), K_x \xi_{p,x}(y', \phi) \right)_{F_K} \]

where the interchange of the inner product and the expectation is justified since \( E_{x,y} \| K_x \xi_{p,x}(y, \phi) \|_{F_K} < \infty \) (Bochner integrability). But by Theorem 1 and (1), we have that \( G_{p,r} := E_{(x,y) \sim r_x} K_x \xi_{p,x}(y, \phi) = 0 \).

So, \( \zeta_1 = 0 \) and the result under \( H_0 \) follows from Serfling [33, Section 5.5.2].

**Case: \( H_1 \) is true**

From (2), it can be seen that

\[ (10) = E_{x,y} k(x,x') h_p((x,y), (x', y')) := t(x', y'). \]

Since \( \zeta_1 = \mathbb{V}[t(x,y)] \), it suffices to show that \( t \) is not a constant function. To see this, note that the kernel \( k \) is \( C_0 \)-universal and cannot be a constant function. The function \( h_p \) (see (3)) includes the kernel \( l \) which is also \( C_0 \)-universal. Therefore, \( t \) is not a constant function and \( \zeta_1 > 0 \). We get the asymptotic normality from the result in Serfling [33, Section 5.5.1].

### A.3 PROOF OF THEOREM 3

Recall the proposition from the main text:

**Theorem 3.** Assume all conditions in Theorem 1. Further assume that \( \mathcal{X} \subseteq \mathbb{R}^{d_x} \) is a connected open set, and \( K(x,x') = k(x, x') I \) where \( k: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is a real analytic kernel i.e., for any \( x \in \mathcal{X} \), \( v \mapsto k(x,v) \) is a real analytic function. Then, for any \( J \in \mathbb{N} \), the following statements hold:

1. Under \( H_0 \), \( T_{p}^V (r) = 0 \) for any \( V = \{v_j\}_{j=1}^J \subset \mathcal{X} \).
2. Under \( H_1 \), if \( v_1, \ldots, v_J \) in \( V \) are drawn from a probability density \( \eta \) whose support is \( \mathcal{X} \), then \( \eta \)-almost surely \( T_{0}^V (r) > 0 \).

**Proof.** Recall that \( T_{p}^V (r) := \frac{1}{d_x} \sum_{i=1}^J \| G_{p,v_i} (v_i) \|_{L^2_{p,x}}^2 \). If \( H_0 \) is true, then \( G_{p,v_i} = 0 \) by Theorem 1. As a result, \( T_{0}^V (r) = 0 \). Now suppose that \( H_1 \) is true. We first show that \( \bar{G}(v) := \| G_{p,v} (v) \|_{L^2_{p,x}}^2 \) is a real analytic function. Consider

\[ \bar{G}(v, v') = E_{x,y} E_{x',y'} k(x, v) k(x', v') h_p((x,y), (x', y')) \]

17
\[
= \mathbb{E}_{x,x'} \hat{k}((x,x'),(v,v')) \hat{h}_p(x,x'),
\]
where \( \hat{k}(x,x',(v,v')) = k(x,v)k(x',v') \) and \( \hat{h}_p(x,x') = \mathbb{E}_{y|x} \mathbb{E}_{y'|x} h_p((x,y),(x',y')) \). Note that \( \hat{h}_p(x,x') = \langle g_{p,r}(\cdot|x),g_{p,r}(\cdot|x') \rangle_{\mathcal{F}^d_p} \), and thus we have
\[
\mathbb{E}_{x,x'} \langle g_{p,r}(\cdot|x),g_{p,r}(\cdot|x') \rangle_{\mathcal{F}^d_p}^2 \leq \left( \mathbb{E}_x \|g_{p,r}(\cdot|x)\|_{\mathcal{F}^d_p}^2 \right)^2.
\]

The RHS is finite by Assumption 3 in Theorem 1, and so \( \hat{h}_p \in L^2(\mathcal{X} \times \mathcal{X},r_x \otimes r_x) \). Therefore, \( \hat{G} \) is given by the integral transform of \( \hat{h}_p \) with respect to the kernel \( \hat{k} \), which implies that \( \hat{G} \) is an element of the RKHS of \( \hat{k} \) [35, Theorem 4.26]. Since the product of real analytic functions is real analytic, consequently for any \((v,v'),(z,z') \mapsto \hat{k}(z,z') \) is real analytic and bounded by our assumption. Thus, by Lemma 6, \( \hat{G}(v,v') \) is analytic. From (7), we have \( \hat{G}(v,v) = G(v,v) \); hence \( \hat{G} \) is analytic and not a zero function by Theorem 1. Since the zero set of \( \hat{G}(v) \), \( \{v' \in \mathcal{X} | \hat{G}(v') = 0 \} \), has zero Lebesgue measure [29], we have that \( \eta \)-almost surely \( \hat{G}(v) > 0 \) for any \( v \sim \eta \), and the result follows.

\begin{proposition}[Asymptotic distributions of \( \hat{T}_p^V \)] Assume that \( \mathbb{E}k_{V}(x,x')h_p^2((x,y),(x',y')) < \infty \). The following statements hold.
\begin{enumerate}
\item If \( \sigma_{V}^2 := 4V[\mathbb{E}_{xy}[\mathcal{P}_p^V((x,y),(x',y'))]] > 0 \), then \( \sqrt{n} \left( \hat{T}_p^V - T_p^V(r) \right) \overset{d}{\to} N(0,\sigma_{V}^2); \)
\item If \( \sigma_{V}^2 = 0 \), then \( n\hat{T}_p^V \overset{d}{\to} \sum_{j=1}^{\infty} \lambda_j(\chi_{1j}^2 - 1) \), where \( \{\chi_{1j}\}_j \) are independent \( \chi_1^2 \) random variables, \( \lambda_j \) are eigenvalues of the operator \( A \) defined as \( (A\varphi)(z) = \int \mathcal{T}_p^V(z,z')\varphi(z')r_{xy}(z') \, dz' \) for non-zero \( \varphi, z := (x,y) \) and \( z' := (x',y') \).
\end{enumerate}
\end{proposition}

\section{NYC Taxi Data Experiment}
Here, we describe technical details of the Mixture Density Network (MDN) used in the NYC taxi data experiment\textsuperscript{4} for estimating the conditional probability of a drop-off location given a pick-up location. The NYC taxi dataset is available at \url{https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page}. An MDN specifies a conditional density model of the form
\[
p(y|x) = \sum_{i=1}^{C} \pi_i(x) N \left( y \mid \mu_i(x), \text{diag} \left( \sigma_{i,1}^2(x), \ldots, \sigma_{i,d_y}^2(x) \right) \right),
\]
where \( C \) is the number of Gaussian components, \( x \in \mathbb{R}^{d_x} \), \( y \in \mathbb{R}^{d_y} \) and \( \text{diag}(s) \) constructs a diagonal matrix with the diagonal entries given by \( s \). In our problem, \( x \) (pick-up location) and \( y \) (drop-off location) contain latitude/longitude coordinates; so, \( d_x = d_y = 2 \). The mixing proportion function \( \pi(x) := (\pi_1(x), \ldots, \pi_C(x)) \), the mean function \( \mu(x) := (\mu_1(x), \ldots, \mu_C(x))^\top \in \mathbb{R}^{C \times d_y} \), and the variance function \( \sigma^2(x) := (\sigma_{i,j}^2(x))_{i,j} \in \mathbb{R}_+^{C \times d_y} \) for \( i \in \{1, \ldots, C\}, j \in \{1, 2\} \) depend on \( x \) and are specified by neural networks. The network architecture is as follows:

\textsuperscript{4}Our implementation is lightly based on public code at \url{https://github.com/sagelywizard/pytorch-mdn}. 

\begin{figure}[ht]
\centering
\includegraphics[width=\textwidth]{network架构.png}
\caption{Network Architecture}
\end{figure}
Table 1: Network architecture for $\pi$.

| Layer ↓ | Input | Output |
|---------|-------|--------|
| Linear  | $d_x = 2$ | 128 |
| Batch normalization | - | - |
| ReLU activation | - | - |
| Linear  | 128 | 64 |
| Batch normalization | - | - |
| ReLU activation | - | - |
| Linear  | 64 | $C = 20$ |
| Softmax | - | - |

Table 2: Network architecture for $\mu$.

| Layer ↓ | Input | Output |
|---------|-------|--------|
| Linear  | $d_x = 2$ | 128 |
| Batch normalization | - | - |
| ReLU activation | - | - |
| Linear  | 128 | 64 |
| Batch normalization | - | - |
| Linear  | 64 | $C \times d_y$ |

Table 3: Network architecture for $\sigma^2$.

We train the model on five million trip records of New York’s yellow cabs from January 2015 using $C = 20$ Gaussian components. Only trips with pick-up and drop-off locations within or close to Manhattan are used. The training objective function is the negative log likelihood. We train the model for three epochs using Adam [27] as the optimization procedure, with a minibatch size of 2000, and a learning rate of 0.001.