Machine Learning with Applications 12 (2023) 100463

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Gradient-Free Kernel Conditional Stein Discrepancy goodness of fit testing

Elham Afzali *, Saman Muthukumarana

Department of Statistics, University of Manitoba, Winnipeg, R3T 2N2, Manitoba, Canada

ARTICLE INFO

Keywords:
Goodness-of-fit testing
Kernel Stein Discrepancy
Reproducing Kernel Hilbert Spaces
Kernel Stein Discrepancy for conditional density
Gradient-Free Kernel Stein Discrepancy
Importance sampling

ABSTRACT

In this study, we propose a gradient-free statistical goodness-of-fit test for determining if a joint sample (x_i, y_i) is drawn from $p(y|x)\pi_x$ for some density π_x given a conditional distribution. This test is an alternative to Kernel Conditional Stein Discrepancy, which require the computation of model derivatives and are therefore impractical for complex statistical models. Our method, known as Gradient-Free Kernel Conditional Stein Discrepancy, does not require the calculation of derivatives, this makes it a great tool for tackling difficult problems such as evaluating the performance of generative models. It is able to detect convergence and divergence with the same level of accuracy as the gradient-based method. We also discuss the application of this test in importance sampling and compare its performance with two other conventional methods.

1. Introduction

The goodness-of-fit test is a statistical hypothesis test that determines how closely observed data matches expected data. Goodness-of-Fit tests can assist identify whether or not two random samples are drawn from the same distribution. Classical methods as in Kolmogorov (1933), Smirnov (1948) entail comparing the models' cumulative distribution functions or their likelihoods. However, recent statistical and machine learning methods including graphical models (Koller & Friedman, 2009) or deep generative models (Salakhutdinov, 2015), intensely rely on complicated probabilistic models with difficult-to-compute likelihoods or cumulative distribution functions. To tackle this challenge, Gorham and Mackey (2015) introduced a measure that determines the maximum discrepancy between the expectation over a class of test functions and the empirical expectation of a given sample. The class of test functions is chosen in such a way that applying the Stein operator to them yields zero expectation over the target distribution. The closed-form of integrals over the target distribution is not required since the Stein operator merely requires the derivative of the logarithm of the target distribution's likelihood.

The implementation of the Stein operator on the Sobolev space, which increases the complexity of the test function class, is a barrier to adopting (Gorham & Mackey, 2015)'s approach. To deal with this challenge, Chwialkowski, Strathmann, and Gretton (2016), Liu, Lee, and Jordan (2016) introduced a likelihood-free method that is applicable to high-dimensional distributions. Their work were inspired by Oates et al. (2017) which was the first to investigate the use of Stein's identity in combination with kernel techniques to reduce the variance of Monte Carlo integration. It did not, however, specify the goodness-of-fit test. Chwialkowski et al. (2016), Liu et al. (2016) independently presented a nonparametric goodness-of-fit test as Kernel Stein Discrepancy (KSD) in such a way that instead of using a class of test functions in Sobolev space, they compute the Stein discrepancy in a ball Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite kernel. For discrete distributions, (Yang, Liu, Rao, & Neville, 2018), extended the kernel Stein discrepancy test in such a way that can be used for the discrete domain. Jitkrittum, Kanagawa, and Schölkopf (2020) proposed an extension of the KSD test called the goodness-of-fit test of Kernel Conditional Stein Discrepancy (KCSD) for conditional density models. They successfully implemented a kernel operator on the conditional witness function.

It has been demonstrated that the proposed discrepancies are effective statistical tools with a wide range of applications like parameter inference, and sampling (Barp, Briol, Duncan, Girolami, & Mackey, 2019; Chen et al., 2019; Chen, Mackey, Gorham, Briol, & Oates, 2018; Fisher, Nolan, Graham, Prangle, & Oates, 2021; Hodgkinson, Salomone, & Roosta, 2020; Liu & Lee, 2017; Matsubara et al., 2021; Riabiz et al., 2022). In real-world situations, however, gradient information of the target distribution is not always accessible. In some circumstances, the gradient cannot be determined analytically since the target distribution is only provided up to the normalization constant or as in other scenarios, computing the gradient could be too costly (Andrieu & Roberts, 2009; Filippone & Girolami, 2014).

To overcome this problem, Han and Liu (2018), Liu and Wang (2016) proposed the Stein variational gradient descent by employing Gradient-Free Stein operators. Fisher, Oates, et al. (2022) also introduced the Gradient-Free Kernel Stein Discrepancy (GF-KSD), a computational method, that is the result of integrating gradient-free Stein operators and reproducing kernels. The GF-KSD has been shown

* Corresponding author.
E-mail addresses: afzalie@myumanitoba.ca (E. Afzali), Saman.Muthukumarana@umanitoba.ca (S. Muthukumarana).

https://doi.org/10.1016/j.mlwa.2023.100463
Received 18 December 2022; Received in revised form 26 March 2023; Accepted 26 March 2023
Available online 1 April 2023

to be effective, but one drawback is that it only applies to models with marginal density and it is not defined for the conditional density.

The objective of this present work is on developing a comprehensive test statistic that is capable of detecting any divergence from the given conditional density model in the null hypothesis. We deliver a nonparametric, generalized gradient-free conditional goodness-of-fit test that in the intermediary step, does not require a gradient computation and density estimation. This test is an alternative to Kernel Conditional Stein Discrepancy (KCSD), which requires the computation of model derivatives and is therefore impractical for complex statistical models. The main advantage of the proposed approach is that, unlike other methods in conditional distributions, this approach does not require gradients of the target distribution, which makes it more efficient and cost-effective. In models where certain derivatives of the target distribution are either prohibitively expensive or intractable, this proposed method is useful. As a result, this approach can be used to address challenging tasks such as evaluating the performance of generative models and latent variable models.

The proposed method can detect convergence and divergence with the same level of accuracy as the gradient-based method. For the practical part, the application of this test in importance sampling is checked, and its performance is compared with two other conventional methods. In conditional goodness-of-fit testing, the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD, discussed in Section 2.2), generalizes the kernel conditional Stein discrepancy (KCSD) (Jitkrittum et al., 2020) and the Gradient-free kernel Stein discrepancy (GF-KSD) (Fisher et al., 2022).

Paper outline Section 2 contains two parts, the first part presents the background method for the kernel Stein Discrepancy, KSD, and the second part defines our proposed method, GF-KCSD, and all its essential properties, in Section 3 we provide experiments using GF-KCSD. Section 4 contains the application of this method in importance sampling and the conclusion of the paper is discussed in Section 5.

2. Background methods

Comparing distributions with discrepancy measures is a key component of statistical and machine learning models. For the purpose of proposing our new test statistics; the Gradient Free Kernel Conditional Stein Discrepancy (GF-KCSD), the fundamental materials are provided in the first part of this section. It explains the Kernel Stein Discrepancy (KSD) test by Chwialkowski et al. (2016), Liu et al. (2016), which is the base well-known goodness-of-fit test we will discuss in Section 2.1. In Section 2.2, we introduce a goodness-of-fit test using GF-KCSD to show the efficacy of GF-KCSD in distinguishing distributions. The GF-KCSD is an extension of both the Kernel Conditional Stein Discrepancy from Jitkrittum et al. (2020) and the Gradient-Free Kernel Stein Discrepancy from Fisher et al. (2022). This technique is specifically used to evaluate conditional distributions without the use of gradients.

2.1. Kernel Stein Discrepancy (KSD)

A significant part of machine learning and statistical modeling involves the use of discrepancy measures to compare different distributions. These measures allow us to quantitatively assess the similarity or difference between two or more distributions, and they play a crucial role in many common statistical techniques such as hypothesis testing, model selection, and goodness-of-fit testing. One conventional approach for finding this measure in probability theory is Stein's method (Stein, 1972). The primary idea underlying Stein's method is to utilize a Stein operator instead of the characteristic function that is generally used to demonstrate distributional convergence.

Consider two probability distributions P and Π supported on an open subset or any convex set $\mathcal{X} \subseteq \mathbb{R}^D, (D \in \mathbb{N})$, and their continuous differentiable densities, $p(\mathbf{x})$ and $\pi(\mathbf{x})$. Under suitable boundary conditions, we can define an integrable score function, $\mathbf{s}_p(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x})$, (Oates, Girolami, & Chopin, 2017), which corresponds to the gradient of the log-likelihood with regard to the input (Hyvärinen & Dayan, 2005), rather than the commonly defined score function which is the gradient of the log-likelihood with respect to the parameter. For any test function $\mathbf{g} \in \mathcal{G}^D$, $g : \mathcal{X} \to \mathbb{R}^D$, the Langevin Stein operator of differentiable functions is defined as:

$$\mathcal{T}_p \mathbf{g}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})^T \mathbf{g}(\mathbf{x}) + \nabla_{\mathbf{x}}^T \mathbf{g}(\mathbf{x}) \tag{1}$$

where $\mathcal{T}_p \mathbf{g}(\mathbf{x}) \in \mathbb{R}^D$. In this case, these functions are vector-valued functions for convenience and it is going to output a scalar-valued function that is always going to be mean-zero under the target distribution. Stein's method shows how to yield expectation zero test functions for distributions that are known up to a normalization constant. Applying the Stein operator $\mathcal{T}_p$ on the test functions, meets regularity and boundary conditions (Gorham & Mackey, 2015, Proposition 1) such that:

$$\mathbb{E}_{\mathbf{x}\sim p}[\mathcal{T}_p \mathbf{g}(\mathbf{x})] = 0 \tag{2}$$

Eq. (2), known as Stein's identity, is used to construct Stein discrepancies which are a type of Integral Probability Metric (IPM) that do not require explicit integration under P. This can make the Stein discrepancies more amenable to estimate, since expectations are often easier to compute than integrals. The Stein discrepancy, as described in the work of Gorham and Mackey (2015), Liu et al. (2016), is as follows:

$$S_p(\pi, \mathcal{T}_p, \mathcal{G}) = sup_{\|\mathbf{g}\|_{\mathcal{G}^D} \leq 1} \left| \mathbb{E}_{\mathbf{x}\sim\pi} \mathcal{T}_p \mathbf{g}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p} \mathcal{T}_p \mathbf{g}(\mathbf{x}) \right| \tag{3}$$

The Stein discrepancy thus emphasizes the score difference between p and π. To make it easier to compute and analyze, Chwialkowski et al. (2016), Gorham and Mackey (2017), Liu et al. (2016), Oates et al. (2017), have applied the use of Reproducing Kernel Hilbert Spaces (RKHS, Berlinet & Thomas-Agnan, 2011) to the Stein discrepancy, resulting in the Kernelized Stein Discrepancy. The Kernelized Stein Discrepancy limits the test functions $\mathbf{g}$ to be in an RKHS $\mathcal{G}^D$, associated with a positive definite kernel $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, and their norms $\|\mathbf{g}\|_{\mathcal{G}^D}$ are jointly bounded by one. Under the condition that the kernel k is C_o-universal (Sriperumbudur, Fukumizu, & Lanckriet, 2011), as long as $\mathbb{E}_{\mathbf{x}\sim\pi} \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x}) \right\|_2^2 < \infty$, based on the theorem 2.2 in Chwialkowski, Ramdas, Sejdinovic, and Gretton (2015), $S_p(\pi, \mathcal{T}_p, \mathcal{G}) = 0$ if and only if $p = \pi$.

Rewriting the KSD in a way that makes estimation straightforward can be a useful technique for simplifying calculations and improving the accuracy and efficiency of statistical analysis. Consider the kernel $k(\mathbf{x}, \mathbf{x}')$ is in the Stein class of π. Define $b_p(\mathbf{x}, \mathbf{x}')$ as follows:

$$\begin{aligned} b_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \mathbf{s}_p^T(\mathbf{x}) \mathbf{s}_p(\mathbf{x}') + \sum_{i=1}^{D} \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_i \partial x_i'} \\ + \mathbf{s}_p^T(\mathbf{x}) \nabla_{x'} k(\mathbf{x}, \mathbf{x}') + \nabla_x k(\mathbf{x}, \mathbf{x}')^T \mathbf{s}_p(\mathbf{x}') \end{aligned} \tag{4}$$

where the $\sum_{i=1}^{D} \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_i \partial x_i'} = trace(\nabla_{x,x'} k(\mathbf{x}, \mathbf{x}'))$. Subsequently, given a sample $\{x_i\}_{i=1}^n \sim \pi$, the U-statistic (Serfling, 2009) and the V-Statistics of the squared KSD are defined in Eqs. (5) and (6), respectively.

U-statistics:

$$\hat{S}_p^2(\pi) = \frac{1}{n(n-1)} \sum_{i \neq j} b_p(x_i, x_j) \tag{5}$$

V-statistics:

$$\hat{S}_p^2(\pi) = \frac{1}{n^2} \sum_{i=1}^{n} b_p(x_i, x_j) \tag{6}$$

where n is the sample size. In Chwialkowski et al. (2016), Liu et al. (2016), the authors successfully use the KSD's U-statistics to evaluate whether a given density model p is a proper model for a given sample $\{x_i\}_{i=1}^n \sim \pi$.

The KSD has several attractive properties, such as being a valid metric in the space of probability distributions and being able to capture a wide range of different types of discrepancies between distributions. It has been used in a variety of applications, including density estimation, hypothesis testing, and generative modeling.

2.2. Gradient Free Kernel Conditional Stein Discrepancy (GF-KCSD)

The goal of this section is to propose the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) as a method for detecting the divergence between two conditional density functions. In the first part, we explain how to construct the Gradient-Free Stein Discrepancy in Fisher et al. (2022) for the conditional density functions. All the problem setting in the conditional part is similar to that introduced in Jitkrittum et al. (2020). All densities, from this onwards, are considered as conditional density functions.

Problem Design Let $\mathbf{x} \in \mathbb{R}^{D_x}$ and $\mathbf{y} \in \mathbb{R}^{D_y}$ be two random vectors taking values in $\mathcal{X} \times \mathcal{Y} \subset R^{D_x} \times R^{D_y}$. Let $\mathcal{P}(\mathbb{R}^D)$ be a set of probability distributions on $\mathbb{R}^D$. For $q, q_0 \in \mathcal{P}(\mathbb{R}^D)$, take that q is absolutely continuous with respect to q_0 (denoted by $q \ll q_0$) it means that if for any measurable set A, $q_0(A) = 0$ implies $q(A) = 0$. Let $p = p(\mathbf{y}|\mathbf{x}) \in \mathcal{P}(\mathbb{R}^D)$ be a target model for the conditional density function of $\mathbf{y}|\mathbf{x}$. Given a joint sample $\mathcal{Z}_n = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \pi_{xy}$ from a joint density π_{xy} defined on $\mathcal{X} \times \mathcal{Y}$ we want to define the conditional goodness-of-fit test. It should be noted that π_{xy} is only be seen in the joint sample $\mathcal{Z}_n$, and according to the Bayes rule, the joint density function is defined as $\pi_{xy}(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}|\mathbf{x})\pi_x(\mathbf{x})$. Thus the hypothesis test for conditional goodness-of-fit is then defined as follows:

$$H_0 : p_{\mathbf{xy}} \overset{\pi_x}{=} \pi_{\mathbf{xy}} \qquad vs \qquad H_1 : p_{\mathbf{xy}} \overset{\pi_x}{\neq} \pi_{\mathbf{xy}} \tag{7}$$

The null hypothesis means that for almost all $\mathbf{x}$ and for all $\mathbf{y}$, the target model $p(\mathbf{y}|\mathbf{x})$, is a proper model for the conditional density $\pi(\mathbf{y}|\mathbf{x})$. However, the alternative hypothesis states that there exists a set in $\mathcal{X}(S \subseteq \mathcal{X})$ such that

$$\forall \mathbf{x} \in S, \quad p_{(.|\mathbf{x})} \neq \pi_{(.|\mathbf{x})}$$

The goodness-of-fit test in Eq. (7), is useful in a variety of machine learning and statistical conditional models such as Bayesian classifier models in discrete random vectors and regression models (both with heteroscedasticity and homoscedasticity noises) in continuous random vectors. Insofar as the score function ($\nabla_{\mathbf{y}} \log p(\mathbf{y}|\mathbf{x})$) is differentiable, the target model distribution can be any nonlinear function such as a neural network. In order to use the Stein discrepancy (with gradient or gradient-free) we need to define the Stein operator. In our case, since we want to work with conditional distributions, we need to explain a bit about the reproducing kernels and some of their properties in advance.

Reproducing kernels are widely used in machine learning and statistics because they provide a natural way to define inner products between data points. In machine learning and statistics, inner products are often used to compare data points and measure the similarity or dissimilarities between them. Reproducing kernels also provides a framework for defining feature maps, which are important tools in deep learning. The development of the new test statistic will require vector-valued reproducing kernels. We provide a quick overview of this idea. For advance details, please refer to Carmeli, De Vito, and Toigo (2006), Carmeli, De Vito, Toigo, and Umanitá (2010), Sriperumbudur et al. (2011), Szabó and Sriperumbudur (2017). Let $\mathcal{X}$ be a nonempty set, $\mathcal{W}$ be a real Hilbert space with the inner product $\langle ., . \rangle_{\mathcal{W}}$, and $\mathcal{L}(\mathcal{W})$ be the Banach space (a complete normed vector space) of bounded linear operators on $\mathcal{W}$. Bounded linear operators are linear transformations that map elements of $\mathcal{W}$ to elements of $\mathcal{W}$, and are bounded in the sense that their operator norm is finite.

An operator-valued positive definite kernel ($\mathcal{W}$-reproducing kernel) is a function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{W})$ that satisfies certain conditions. In this context, $\mathcal{X}$ is a set and $\mathcal{L}(\mathcal{W})$ is the space of bounded linear operators on a Hilbert space $\mathcal{W}$. For each pair $(x, y) \in \mathcal{X} \times \mathcal{X}, K(x, y) \in \mathcal{L}(\mathcal{W})$ is a self-adjoint operator (an operator that is equal to its own adjoint), and we have:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \langle K(\mathbf{x}_i, \mathbf{x}_j)\mathbf{w}_i, \mathbf{w}_j, \rangle_{\mathcal{W}} \geq 0$$

for every set of $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and $\{\mathbf{w}_i\}_{i=1}^n \subset \mathcal{W}$ and $n \in \mathbb{N}$. Let $\mathcal{W}^{\mathcal{X}}$ be a vector space of all functions from $\mathcal{X}$ to $\mathcal{W}$ under the assumption that $\mathcal{X}, \mathcal{W}$ are Banach spaces. For each $\mathbf{x} \in \mathcal{X}$ and $\mathbf{w} \in \mathcal{W}$, the linear operator function $K_x\mathbf{w} = K(., \mathbf{x})$ where $w \in \mathcal{W}^{\mathcal{X}}$ defined as $(K_x\mathbf{w})(\mathbf{z}) = K(\mathbf{z}, \mathbf{x})\mathbf{w}$ for all $\mathbf{z} \in \mathcal{X}$. If there is a $\mathcal{W}$-reproducing kernel K in real valued reproducing kernel, then there exists a unique Reproducing Kernel Hilbert Space (RKHS) $\mathcal{G}_K$ as $K_{\mathbf{x}} \in \mathcal{L}(\mathcal{W}; \mathcal{G}_K)$ and under the reproducing property, $g(\mathbf{x}) = K_{\mathbf{x}}^* g$ for all $g \in \mathcal{G}_K$, $\mathbf{x} \in \mathcal{X}$. Under the aforementioned conditions, the adjoint operator of $K_{\mathbf{x}}$ defined as $K_{\mathbf{x}}^* : \mathcal{G}_K \to \mathcal{W}$.

Take into account that $\mathbb{C}(\mathcal{X}; \mathcal{W}) : \mathcal{X} \to \mathcal{W}$ as a vector space of continuous functions, and $C_0(\mathcal{X}; \mathcal{W})$ is a subspace of the $\mathbb{C}(\mathcal{X}; \mathcal{W})$ such that vanishes with the uniform norm at infinity. This means that the elements of $C_0(\mathcal{X}; \mathcal{W})$ become arbitrarily close to 0 as the input values get farther and farther away from the origin i.e. $\|g(\mathbf{x})\|_{\mathcal{W}} \to 0$ as $\|x\| \to \infty$. We can consider a $\mathcal{W}$-reproducing kernel K to be a C_0 if $\mathcal{G}_K \subset C_0(\mathcal{X}; \mathcal{W})$. It means that if all of the functions generated by the $\mathcal{W}$-reproducing kernel K are themselves continuous, then we can consider the kernel to be a C_0 function. Moreover, if $\mathcal{G}_K$ is dense in $L^2(\mathcal{X}, \mu, \mathcal{W})$ for any probability measure μ, then a C_0-kernel K is referred to as universal. This is because the kernel is able to generate a wide range of continuous functions that are dense in the space of square-integrable functions, which means that it can be used to approximate many different types of functions in that space.

Assume that the positive definite kernel $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be connected with the RKHS $\mathcal{G}_l$. For two vectors $\mathbf{a} = (a_1, \ldots, a_{D_y})$, $\mathbf{m} = (m_1, \ldots, m_{D_y})$ that are the elements of the set $\mathcal{G}_l^{D_y}$ where $\mathcal{G}_l^{D_y} = \prod_{i=1}^{D_y} \mathcal{G}_l$, the inner product of two vectors a and m on $\mathcal{G}_l^{D_y}$ is defined as the sum of the inner products of their individual elements $\langle \mathbf{a}, \mathbf{m} \rangle = \sum_{i=1}^{D_y} \langle a_i, m_i \rangle$.

Consider a $\mathcal{G}_l^{D_y}$-reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{G}_l^{D_y}$ that is $\mathcal{W} = \mathcal{G}_l^{D_y}$. Assume a real-valued kernel connected with RKHS $\mathcal{G}_k$ as $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Definition 1. Let the gradient-free Stein operator be the one defined in Definition 1 by Fisher et al. (2022), but instead of the marginal distribution in the definition, consider all densities to be conditional. For $p, \rho \in \mathcal{P}(\mathbb{R}^D)$ when ρ is absolutely continuous with respect to p and given the fact that the score function of ρ is well-defined and $\int \|\nabla \log \rho\| \, d\rho < \infty$, the gradient-free conditional Stein operator on a differentiable function $g : \mathbb{R}^D \to \mathbb{R}^D$ in a way that its first derivatives are bounded, is defined as:

$$\mathcal{T}_{p,\rho} g = \frac{\rho}{p}(\nabla . g + g . \nabla \log \rho)$$

When $\rho(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$ the Langevin Stein operator is recovered. However, when $\rho(\mathbf{y}|\mathbf{x}) \neq p(\mathbf{y}|\mathbf{x})$ the dependency on the derivatives of p would be eliminated. $\mathcal{T}_{p,\rho} g$ can still be identified as a diffusion Stein operator, in this case, (Fisher et al., 2022). We begin by noting that the gradient-free conditional Stein operator is defined from the gradient-free Stein operator in Fisher et al. (2022) by replacing the marginal distribution with the conditional distribution. The presence of ρ adds an additional degree of freedom to Stein operators and gives them a greater range of flexibility in their operations to explore the search space. It is allowing them to capture more information about the function and thus reduces the number of independent variables required to describe the function. The vanishing integral property in C_o-universal kernels is essential in Stein operators that form the basis of Stein's approach (Stein, 1972), and helps to define the Stein identity. This property states that the integral of any C_0-universal kernel over the entire domain of the kernel is equal to zero. This property is useful for ensuring that the integral of the kernel over any given region is independent of the kernel's domain and can be used to analyze the behavior of the kernel over different domains.

Proposition 1. *In the case presented in Definition 1, the discrepancy between two conditional distributions p and π is defined as:*

$$D_{p,\rho}^2(\pi) = \left\| \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \pi_{xy}} K_{\mathbf{x}} \zeta_{\rho(.|\mathbf{x})}(\mathbf{y}, .) - \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim p_{xy}} K_{\mathbf{x}} \zeta_{\rho(.|\mathbf{x})}(\mathbf{y}, .) \right\|_{\mathcal{G}_K}^2 \tag{8}$$

where $\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.) = \frac{\rho}{p}\left[l(\mathbf{y},.)\nabla_{\mathbf{y}}\log\rho(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{y}}l(\mathbf{y},.)\right] \in \mathcal{G}_l^{D_y}$. $D^2_{p,\rho}(\pi)$ *is referred to as Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). The second expectation term in* $D_{p,\rho}(\pi)$ *(under the distribution P) is zero by considering the Stein operator as a non-standard instance of the density method in Stein, Diaconis, Holmes, and Reinert (2004). In this case, the first expectation term in* $D^2_{p,\rho}(\pi)$*, which is the expectation with respect to the distribution* $\pi \in \mathcal{P}(\mathbb{R}^D)$ *is zero if and only if* $p \overset{\pi_x}{=} \pi$*; thus, the value of this expectation can be used to determine how far* π *and* p *diverge. By stating which test functions* (g) *are taken into account and then performing a supremum over the related expectations, a discrepancy is obtained. In this study, for the sake of computational simplicity, we assume that* g *is contained in the unit ball of a Reproducing Kernel Hilbert Space.*

Theorem 1. *Let* $l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ *and* $K : \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{G}_l^{D_y})$ *(where* $\mathcal{L}$ *is a Banach space) be two positive definite* C_0*-universal kernels. Consider the conditional witness function defined as* $\mathbf{r}_{p,\pi}(\boldsymbol{\omega}|\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\pi_{(.|x)}}\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},\boldsymbol{\omega}) \in \mathbb{R}^{D_y}$ *and for all* $\mathbf{x}$, $\mathbf{r}_{p,\pi}(.|\mathbf{x}) \in \mathcal{G}_l^{D_y}$*, suppose that*

1. $\int(\frac{\rho}{p})^2\, d\pi < \infty$;
2. *for* r_x*-almost all* $\mathbf{x}$, $sup_{\mathbf{x}}\left\{\int\left\|\nabla_{\mathbf{y}}\log\frac{\rho(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y}|\mathbf{x})}\right\|_2^2 d\pi\right\} < \infty$;
3. $\int_{\mathcal{X}}\left\|\mathbf{r}_{p,\pi}(.|\mathbf{x})\right\|^2_{\mathcal{G}_l^{D_y}}\pi_x(\mathbf{x})\, d\mathbf{x} < \infty$;
4. $\int\left\|K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y}|.)\right\|_{\mathcal{G}_K} d\pi_{xy} < \infty$;

$D^2_{p,\rho}(\pi) = 0$ *if and only if* $p \overset{\pi_x}{=} \pi$*, for almost all* π_x, $\mathbf{x}\in\mathcal{X}$, $p(.|\mathbf{x}) = \pi(.|\mathbf{x})$

Appendix A.1 includes the proof in details.

The majority of the conventional kernels comply with Theorem 1's requirements which needs kernel K and l to be C_0-universal. The Inverse Multi Quadratic kernel(IMQ), Gaussian kernels, Laplace kernel, etc, are some real-valued C_0-universal kernels (Sriperumbudur et al., 2011). According to Theorem 1, we need a $\mathcal{G}_l^{D_y}$-reproducing, C_0-universal kernel such as $K(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}')I$ where $I \in \mathcal{L}(\mathcal{G}_l^{D_y})$ is an identity operator (Carmeli et al., 2010), and k is a real-valued C_0-universal kernel. In this study we use IMQ kernel described as follows to implement the GF-KCSD:

$$l(\mathbf{y},\mathbf{y}') = (\sigma_y^2 + \|\mathbf{y}-\mathbf{y}'\|^2)^{-\beta}, \qquad \beta\in(0,1), \sigma_y\in(0,\infty) \tag{9}$$

2.3. Explicit form of GF-KCSD

In order to construct the statistical test *for*gradient-free conditional goodness-of-fit, we need to rewrite the $D^2_{p,\rho}(\pi)$ in a form that can be estimated easily as shown in Proposition 2.

Proposition 2. *For a positive definite kernel* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$*, assume that* $K(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}')I$*. Define the score function as* $\mathbf{s}_\rho(\mathbf{y}|\mathbf{x}) = \nabla_{\mathbf{y}}\log\rho(\mathbf{y}|\mathbf{x})$*. Then the GF-KCSD can be written as*

$$D^2_{p,\rho}(\pi) = \iint k(\mathbf{x},\mathbf{x}')\frac{\rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{y}'|\mathbf{x}')}{p(\mathbf{y}|\mathbf{x})p(\mathbf{y}|\mathbf{x})}b_\rho((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))\, d\pi_{\mathbf{xy}}\, d\pi_{\mathbf{x}'\mathbf{y}'} \tag{10}$$

where

$$\begin{aligned} b_\rho((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')) = l(\mathbf{y},\mathbf{y}')\mathbf{s}_\rho^T(\mathbf{y}|\mathbf{x})\mathbf{s}_\rho(\mathbf{y}'|\mathbf{x}') + \sum_{i=1}^{D_y}\frac{\partial^2 l(\mathbf{y},\mathbf{y}')}{\partial\mathbf{y}_i\partial\mathbf{y}_i'} \\ + \mathbf{s}_\rho^T(\mathbf{y}|\mathbf{x})\nabla_{\mathbf{y}'}l(\mathbf{y},\mathbf{y}') + \nabla_{\mathbf{y}}l(\mathbf{y},\mathbf{y}')^T\mathbf{s}_\rho(\mathbf{y}'|\mathbf{x}') \end{aligned} \tag{11}$$

Define $B_\rho((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')) = k(\mathbf{x},\mathbf{x}')\frac{\rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{y}'|\mathbf{x}')}{p(\mathbf{y}|\mathbf{x})p(\mathbf{y}'|\mathbf{x}')}b_\rho((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))$. Having a joint sample $\{(\mathbf{x_i},\mathbf{y_i})\}_{i=1}^n \overset{i.i.d.}{\sim} \pi_{xy}$ an unbiased, consistent estimator for Eq. (10) is given by

$$\hat{D}^2_{p,\rho}(\pi) = \frac{1}{n(n-1)}\sum_{i\neq j}B_\rho((\mathbf{x}_i,\mathbf{y}_i),(\mathbf{x}_j,\mathbf{y}_j)) \tag{12}$$

The proof is in the Appendix A.2. From Eq. (11) the GF-KCSD is clearly dependent on the model p only through the density value, and is dependent on the model ρ only through the $\nabla_{\mathbf{y}}\log\rho(\mathbf{y}|\mathbf{x}) = \nabla_{\mathbf{y}}\log\rho(\mathbf{y},\mathbf{x})$. In this regard, the GF-KCSD is unaffected by the normalizer $p(\mathbf{x})$ and $\rho(\mathbf{x})$ and can be employed with p^* instead of $p \propto \frac{p^*}{C}$, where C is an intractable normalization constant. Due to this characteristic, GF-KCSD can be applied to posterior approximation problems. One critical aspect of posterior approximation problems is ensuring a consistent approximation of the target distribution. This can be addressed by convergence control; the next section takes this worry into account in GF-KCSD.

2.4. Convergence detection and control

Convergence for a broad class of target distributions is indisputably determined by the Langevin kernel Stein discrepancy for the kernels with slow decaying tails (Gorham & Mackey, 2017). We must deal with equivalent convergence detection in order to demonstrate that the proposed discrepancy(GF-KCSD) is consistent.

The initial step in setting the convergence detection properties of the GF-KCSD is to configure its convergence-detecting settings. Huggins and Mackey (2018) provide a basis for this configuration by defining the Lipschitz constant for a Lipschitz function, which is a mathematical function that satisfies the Lipschitz condition. This constant is used to measure the maximum difference between the values of the function at two different points.

The Lipschitz constant for a Lipschitz function $h : \mathbb{R}^D \to \mathbb{R}^D$ is defined as $L(h) = \underset{y\neq y'}{sup}\frac{\|h(y)-h(y')\|}{\|y-y'\|}$, where y and y' are two different points in the domain of the function. The well-defined tilted Wasserstein distance is then established between the Lipschitz function h and a measurable function $f : \mathbb{R}^D \to \mathbb{R}$ as

$$W_1(\pi,p;f) = \underset{L(h)\leq 1}{sup}\left|\int hf\, d\pi - \int hf\, dp\right| \tag{13}$$

where π and p are two different probability distributions and f is a measurable function. The well-defined tilted Wasserstein distance can be used to measure the difference between the two probability distributions. Once the Lipschitz constant and the well-defined tilted Wasserstein distance have been established, they can be used to detect convergence in the GF-KCSD. The convergence detection properties of the GF-KCSD are based on these two measures, which allow it to accurately determine when a target distribution has reached a stable state and is not changing significantly with further calculations. This is important for ensuring the validity of the results obtained from the GF-KCSD.

In order to demonstrate the consistency of the GF-KCSD, it is necessary to show that it is able to accurately detect convergence for a broad range of target distributions. This can be done by comparing the results of the GF-KCSD with known converged distributions and verifying that they match. By doing this, it is possible to confirm that the GF-KCSD is able to accurately detect convergence and produce reliable results.

Theorem 2. *Let* p,ρ *are conditional distributions in* $\mathcal{P}(\mathbb{R}^D)$ *and* ρ *is absolutely continuous with respect to p, the score function of* ρ *is well-defined (Lipschitz) and* $\int\|\nabla\log\rho\|^2\, d\rho < \infty$*. Let* $K_x = k(x,x')I$*, for a sequence* $\pi_n \in \mathcal{P}(\mathbb{R}^D)$*, we have* $\int\left\|K_x\nabla\log\rho\right\|^2 d\pi_n < \infty$, $\int\left(\frac{\rho}{p}\right)^2 d\pi_n \in (0,\infty)$, $\int\left\|K_x\nabla\log\rho\right\|^2\left(\frac{\rho}{p}\right)d\pi_n < \infty$, $\int\|y\|\frac{\rho(y)}{p(y)}\, d\pi_n(y) < \infty$*. Consider* l *and* k *as a kernels such that all their derivatives exsit and are bounded. Then*

$$W_1(\pi_n,p,\frac{\rho}{p}) \to 0 \quad \Rightarrow \quad D_{p,\rho}(\pi_n) \to 0.$$

According to Eq. (13), by considering the function f as the weighting function $\frac{\rho}{p}$, the gradient-free kernel conditional Stein discrepancy $D_{p,\rho}$ is able to detect the convergence of π_n to p.

Under the same theoretical foundations and assumptions as Gorham and Mackey (2017), the gradient-free kernel conditional Stein discrepancy generally does not give weak convergence control. It is important to note that the GF-KCSD requires a specific framework and is not simply an extension of the standard kernel Stein discrepancy. The framework configuration can be founded on meticulously recasting gradient-free kernels.

Proposition 3. *Let $\rho \in \mathrm{P}(\mathbb{R}^D)$, $p \in \mathcal{P}(\mathbb{R}^D)$, and $\inf_{y\in\mathbb{R}^D} \frac{\rho(y|x)}{p(y|x)} > 0$ such that p is continuous. Consider the sequence $\pi_n \subset \mathcal{P}(\mathbb{R}^D)$ which the preconditions of Theorem 2 are satisfied then we have*

$$D_{p,\rho}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{D} p.$$

The convergence detection necessitated extra conditions on ρ. Consider the weight matrix $\mathrm{M}(y) = (\frac{\rho(y)}{p(y)})I$ which is a position-dependent matrix and Lipschitz, explained in proposition 8 of Gorham, Duncan, Vollmer, and Mackey (2019). This means that the mathematical function used to calculate the weight matrix must satisfy the Lipschitz condition, which is a property of functions that limits the rate at which their values can change. In light of this circumstance, for ensuring that the convergence detection results are accurate and reliable, ρ needs to be equally concentrated as p.

In Section 3, we shall see that proper choices for kernel parameters and ρ lead to a higher test power. Note that in Sections 3 and 4 we consider all the kernels to be inverse multi-quadric (IMQ) kernels.

3. Experiments

In this section, we conduct an empirical investigation into the proposed tests using different kernel parameters and different ρ as an approximation of p. Our initial aim is to demonstrate how the proposed GF-KCSD can identify areas of difference between p and π in the domain of the conditioning variable (x). The experiments were conducted using Python and mainly the PyTorch library from Paszke et al. (2019).

In this study, we assume that kernel $k(x, x') = (\sigma_x^2 + \|x - x'\|^2)^{-\beta}$ and $l(y, y') = (\sigma_y^2 + \|y - y'\|^2)^{-\beta}$ to be the inverse multi-quadratic (IMQ) kernels. Generally speaking, the parameters of the IMQ needed to be chosen with caution and they should be specifically tuned to the problem's configuration. The shape parameter, σ, is taken into consideration as $\sigma = 1$. The other parameter is regarded to be $\beta > D/2$ as in Sriperumbudur et al. (2011), since we are considering univariate analysis ($D = 1$), then we assume $\beta = 0.5$. Although, we investigate different kernel parameters in detail.

The goal is to analyze the behavior of the proposed test (GF-KCSD). For this, a number of sequences as (π_n), $n \in \mathbb{N}$ are taken into consideration to evaluate the performance of test statistic. As it is displayed in Figure Fig. 1, some of which converge to a given distribution p (straight lines) and others (dash-dotted lines) converge to an alternative Gaussian target distribution. Since we are working with univariate distributions, we should keep in mind that the distribution of these sequences (s) be in one-dimensional probability space $\mathcal{P}(\mathbb{R})$. The sequences have been chosen to have a sloped pattern (shifting in mean and variance). The powerful discrepancy should be able to distinguish between the convergent and divergent nature of these sequences.

This section contains two scenarios with respect to the way of choosing the ρ distribution. In Section 3.1 we explore the behavior of the GF-KCSD based on the assumption that the target distribution is conditional Gaussian model, on the other hand, in Section 3.2 we analyze the behavior of the proposed test statistic under the assumption that the target distribution p, follows the conditional Gaussian mixture model. In order to inherit the desirable performance of the Langevin kernel Stein discrepancy for which ρ and p are equal, we represent ρ as an approximation of p in the discussion that follows. As a requirement in Proposition 3, the way to choose ρ should satisfy $\inf_{y\in\mathbb{R}^D} \frac{\rho(y)}{p(y)} > 0$, note that the distributions are conditional. Given this explanation, each target distribution has its own setting for selecting ρ as its approximation.

3.1. Conditional Gaussian model

Problem setting: We take into account a basic univariate problem where the data generating distribution is $\pi_x(x) = Uniform(-2, 2)$, and the target model is $p(y|x) = \mathcal{N}(y; 0.5(x+1), 1)$. To define the convergence sequences $\pi_n \sim s$, we put sequence distribution (s) equal to the target distribution p, while for the non-convergence sequences, the sequence distribution (s) is $\mathcal{N}(0, 0.64)$. In this configuration, we assume such as the Bayesian approach that, the target distribution p is the posterior distribution, taking ρ to be the prior distribution. We define two priors that both satisfy the requirements in Proposition 3: one has a tail that is heavier than the target distribution $\rho_{heavy} \sim \mathcal{N}(0.5, 1.75)$, while the other has a light tail $\rho_{light} \sim \mathcal{N}(0.5, 0.64)$.

3.2. Conditional Gaussian mixture model

Problem setting: We take into account that the data generating distribution is $\pi_x(x) = \mathrm{N}(0.1)$, and we assume the target distribution as follows

$$p(y|x) = \sum_{i=1}^{3} w_i \mathcal{N}(y; \mu_i, \sigma_i^2)$$

where $\mathcal{N}(y; \mu_y, 1)$ is the Normal density, $\mu_y = 0.5(x + 1)$. The weight vector is $\mathbf{w} = (0.3, 0.6, 0.1)$, the mean vector is as $\boldsymbol{\mu} = (-1, 0.5, 0.2)\mu_y$ and the variance vector is as $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.15, 0.2, 0.9)$.

To define the approximation distribution, in this configuration, we consider the Laplace approximation and the Gaussian mixture model. In problems where the target distribution p is differentiable yet computationally costly, it is practical to use its Laplace approximation ρ instead. Given the explanation regarding the target distribution, the computed Laplace approximation has $\rho \sim \mathcal{N}(0.3, 0.2040^2)$ distribution. Another scenario is when there is not any problem in derivatives of the target distribution p and samples from it can be obtained straightforwardly, one may consider ρ to be a Gaussian mixture model fitted to these samples which is a more suitable and flexible substitute for Laplace. The Gaussian mixture model here is computed using 100 samples from the target p. To choose the number of components in the mixture model we consider the minimum Akaike Information Criterion (AIC) and according to this criterion, the best number of components is 2.

The performance of our proposed test statistic for the different conditional target models and their ρ's (with mentioned settings) are shown in Figure Fig. 2. It is evident that two selected ρs (heavy-tailed and light-tailed) in the conditional Gaussian model, can identify when π_n converges or diverges from the target distribution p. As for the selection of ρ, it depends on the application and the type of data. The heavy-tailed ρ is suitable for capturing outliers in the data, while the light-tailed ρ is suitable for dealing with data with fewer outliers. In addition, it is worth noting that the selection of ρ also affects the convergence speed of the algorithm, the light-tailed ρ can accelerate the convergence speed while the heavy-tailed ρ can slow down the convergence speed. On the other hand, for the conditional Gaussian mixture target distribution, using the Gaussian mixture model as ρ, the convergence to the target distribution has a steeper slope than the Laplace approximation. It is clear from the figure that the Laplace transformation convergences after $n > 50$.

4. Application

To illustrate the practicality of gradient-free kernel conditional Stein discrepancy, we demonstrate how it can be applied to posterior approximation via Stein importance sampling (presented in Section 4.1). This extension expands the scope of existing algorithms to statistical models where certain derivatives of p are either prohibitively expensive or intractable.

4.1. Gradient-free conditional Stein importance sampling

Importance sampling is a method used in statistical sampling to improve the accuracy and efficiency of estimates. This is achieved by weighting the sample data according to their relative importance or relevance to the population being studied. By assigning higher

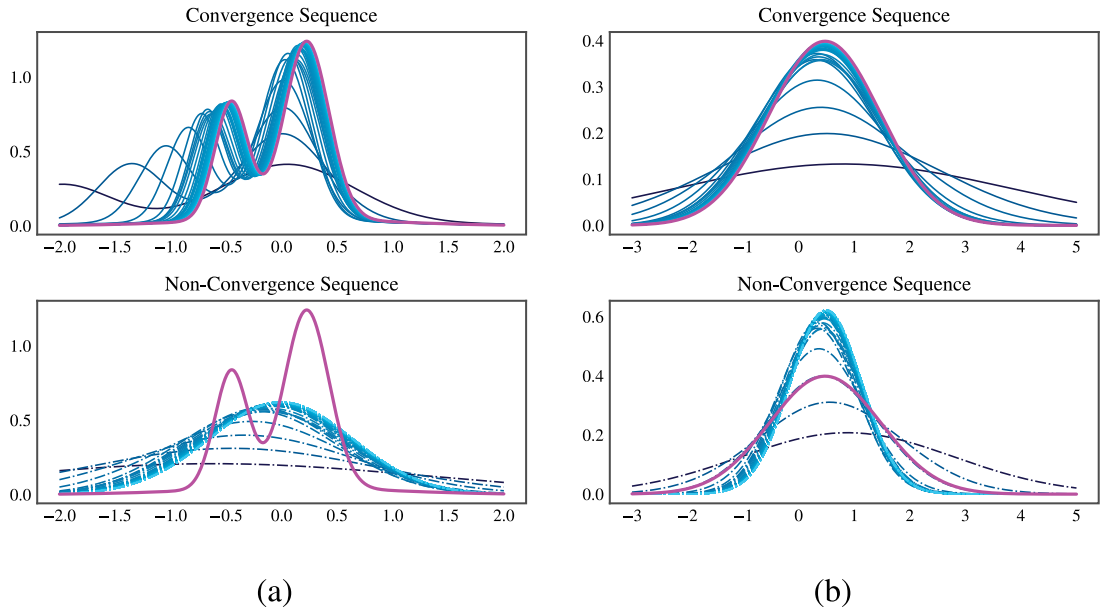
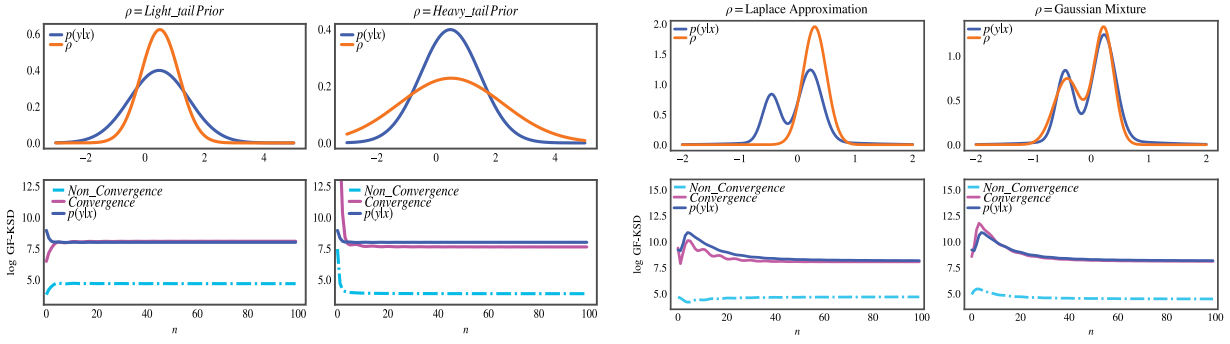


Fig. 1. Experimental slant sequences π_n with fixed kernel parameters($\sigma = 1$, $\beta = 0.5$). The first row shows the sequences (straight line) converge to the target distribution p (magenta), while the second row depicts the sequences (dash–dot) that converge to different Gaussian distribution. (a) The target distribution is conditional Gaussian mixture model. (b) The target distribution is conditional Gaussian model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Conditional Gaussian target

(b) Conditional Gaussian mixture target

Fig. 2. Experimental evaluation of the gradient-free kernel conditional Stein discrepancy on the sequence π_n. The first row displays the distribution of the target model p and its corresponding ρ. (a) displays the assessed GF-KCSD for the conditional Gaussian target. (b) depicts the evaluated GF-KCSD for the Conditional Gaussian mixture target. For the Laplace transformation. For $n > 50$, the convergence is evident. While for ρ as Gaussian mixture model, for $n > 10$ the convergence happens.

weights to more important or representative data points, the overall estimate is more reflective of the true population. Importance sampling is a widely utilized technique in machine learning and statistics, yet its effectiveness is limited as it requires the use of straightforward proposals with easily calculated importance weights. Stein importance sampling is an effective and efficient tool for sampling from a target probability distribution p, even when the distribution cannot be differentiated (Hodgkinson et al., 2020; Liu & Lee, 2017). It involves sampling from a tractable approximation of the target distribution ρ and then correcting the bias in the samples in order to obtain samples from the desired target distribution p. The technique has been successfully applied in instances where the statistical model p can be differentiated, but Fisher et al. (2022)'s contribution removed this requirement. They proposed a practical solution (gradient-free Stein importance sampling) to the problem of sampling from a target distribution p, even when the distribution cannot be differentiated. Our contribution is to apply gradient-free Stein importance sampling to conditional target distributions. We will analyze this method, which involves generating independent samples (y_n) from the same approximate distribution ρ as used in gradient-free kernel conditional Stein discrepancy (GF-KCSD).

To apply our proposed method, we need to extend Theorem 3 in Fisher et al. (2022) to the case of conditional target distributions. This will allow us to use gradient-free Stein importance sampling to accurately and efficiently estimate conditional probabilities, even when the target distribution cannot be differentiated. By doing this, we can extend the applicability of gradient-free Stein importance sampling and improve its usefulness for statistical estimation.

Theorem 3 (*Gradient-Free Conditional Stein Importance Sampling*)**.** *Let conditional distribution p and ρ, satisfying the preconditions of Proposition 3. Assume that $\int exp(\lambda \left\| K_x \nabla \log \rho \right\|^2) \, d\rho < \infty$ for some positive λ. For independent samples (y_n) from ρ, the optimal weights are as follows:*

$$\alpha^* \in argmin \left\{ D_{p,\rho} \left(\sum_{i=1}^{n} \alpha_i \, \delta(y_i) \right) : 0 \leq \alpha_1, \ldots, \alpha_n, \sum_{i=1}^{n} \alpha_i = 1 \right\}$$

where $\pi_n = \sum_{i=1}^{n} \alpha_i^* \, \delta(y_i)$, as $n \to \infty$, $\pi_n \xrightarrow{d} p$.

The weighted samples are generated by multiplying each sample by its corresponding weight, and then summing over all the samples. As the number of samples increases, the distribution of the weighted

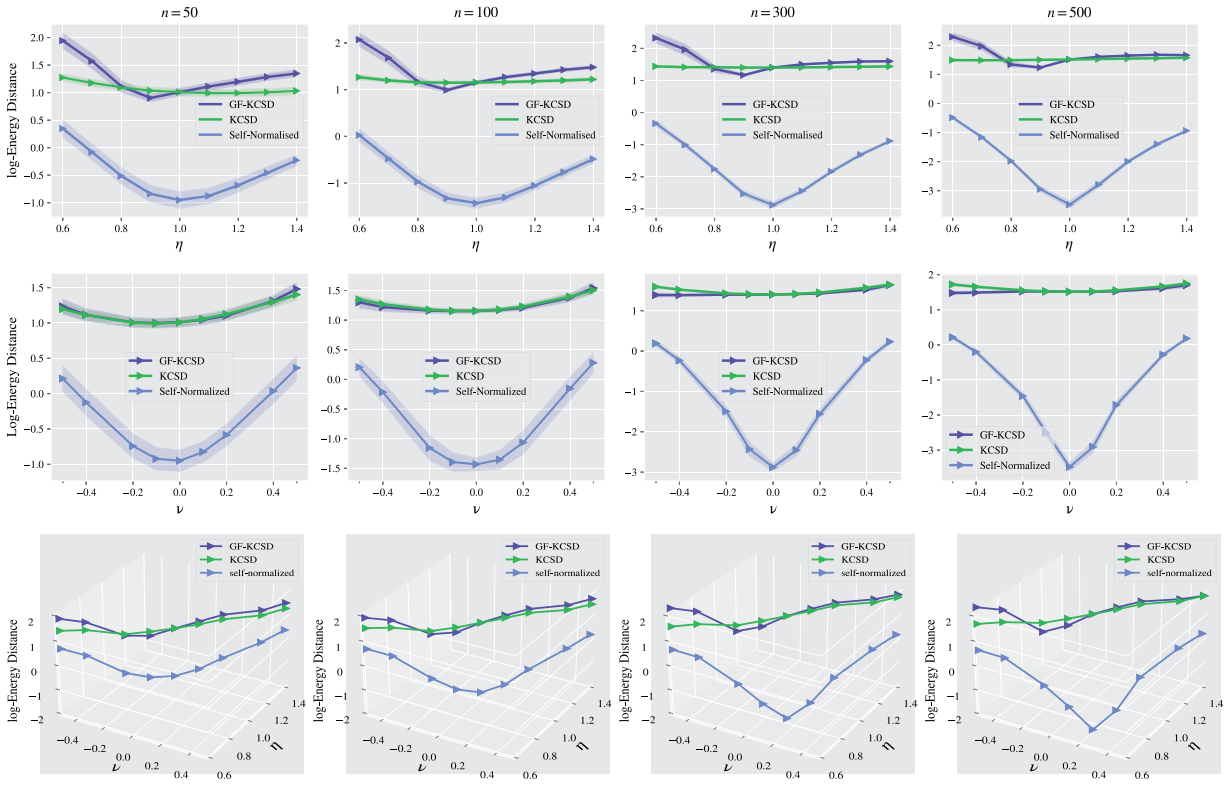


Fig. 3. Evaluation of gradient-free conditional importance sampling in different sample size and scenarios: First Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x, \eta)$, Second Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x + \nu, 1)$, Third Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x + \nu, \eta)$.

samples generated by the importance sampling method converges to the true target distribution almost surely. This means that the estimates obtained using importance sampling become more accurate and reliable as the number of samples increases.

To demonstrate the effectiveness of gradient-free conditional Stein importance sampling, we implemented it to approximate a posterior distribution. We analyze the efficacy of GF-KCSD importance sampling in comparison with the performance of two other importance sampling methods: KCSD importance sampling, and self-normalized importance sampling. The experiment is run by first sampling a set of points from the target and approximate distributions. These samples are then used to calculate the maximum mean discrepancy (MMD) between the two distributions using each of the three importance sampling methods. The three importance sampling methods differ in the specific techniques they use to correct for bias in the samples. We utilize the logarithm of the energy distance as a measure of the quality of the approximation and compare the efficacy of these three methods. In general, a lower logarithm of energy distance indicates that the two distributions are more similar, while a higher energy distance indicates that they are more different.

Problem Setting: We consider the performance of each method for a different number of samples n. As mentioned in Theorem 3, we are required to find the optimal weights for conditional Stein importance sampling (with gradient and without gradient). This optimization problem can be solved using a type of algorithm called a splitting conic solver, which was developed by O'donoghue, Chu, Parikh, and Boyd (2016). These types of algorithms are typically used to solve optimization problems that involve a combination of linear and convex quadratic constraints. There are three different scenarios:

- **Different scale scenario:** $\pi_x \sim \mathcal{N}(0, 1)$, $p(y|x) \sim \mathcal{N}(y; x, 1)$ and $\rho(y|x) \sim \mathcal{N}(y; x, \eta)$.
- **Different location scenario:** $\pi_x \sim \mathcal{N}(0, 1)$, $p(y|x) \sim \mathcal{N}(y; x, 1)$ and $\rho(y|x) \sim \mathcal{N}(y; x + \nu, 1)$.
- **Different scale-location scenario:** $\pi_x \sim \mathcal{N}(0, 1)$, $p(y|x) \sim \mathcal{N}(y; x, 1)$ and $\rho(y|x) \sim \mathcal{N}(y; x + \nu, \eta)$

where $0.6 \leq \eta \leq 1.4$ and $-0.5 \leq \nu \leq 0.5$. Figure Fig. 3 displays the result of each scenarios. These results indicate that the way that gradient-free conditional Stein sampling corrects the bias in the samples in comparison with the conditional Stein importance sampling performed almost similar to or even better.

In the different scale scenario, for $\eta > 0.8$, the proposed method performed better than the gradient-based approach. Moreover, When the number of samples increases, the proposed method performs almost similarly to the gradient-based importance sampling. In the different location scenario, for $\nu > -0.2$ the proposed method performed almost similar to or better than the gradient-based conditional Stein importance sampling. In the different scale-location scenario, for $-0.2 < \nu < 0.1$ and $0.8 < \eta < 1.2$ it is evident that the Gradient-Free Conditional Stein importance sampling outperformed gradient-based approach. And in other intervals, its performance was comparable to that of the gradient-based Stein importance sampling.

5. Conclusion

We propose Gradient-Free Conditional Stein Discrepancy which is a measure of the difference between two distributions, and it is often used in statistical testing and machine learning. Compared to existing gradient-based methods, such as Kernel Conditional Stein Discrepancy (KCSD), GF-KCSD does not require the computation of derivatives of the target distribution, thereby making it more efficient and cost-effective. Moreover, GF-KCSD is particularly useful in models where certain derivatives of the target distribution are either computationally expensive or intractable. We have done the experimental investigation

in a wide range of scenarios for checking its ability to identify the convergence and divergence from the target distribution. The result shows that this approach is scalable and stable. Besides, we utilized this method in importance sampling problem. The result evidently showed that Gradient-Free Conditional Stein importance sampling achieved similar to or better performance than gradient-base conditional Stein importance sampling in terms of the logarithm of the energy distance in variety of problem settings. Future work on the GF-KCSD can include extending it to deal with a discrete domain $\mathcal{Y}$ by applying a Stein operator based on forward and backward differences. These research ideas remain open for consideration.

CRediT authorship contribution statement

Elham Afzali: Conceptualization, Methodology, Software, Visualization, Writing – original draft, Data curation. **Saman Muthukumarana:** Reviewing, Supervision, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Appendix. Proofs

Proofs for all the theoretical findings discussed in the main article are included in this appendix.

A.1. Proof of Theorem 1

Proof. First we need to show that the second expectation in Eq. (8) is equal to zero. To do that in the first step we rewrite the expectation in the integral form ($\int |K_{\mathbf{x}}\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},.)|\,dp$) and show that it is well-defined. The function g and its first derivatives are assumed to be bounded, which means that the magnitude of $g(x)$ and its first derivatives are never too large for any point $x \in \mathbb{R}^d$. Now the quantities C_0 and C_1 are defined as the supremum (i.e., the smallest upper bound) of the magnitudes of $g(x)$ and its derivatives, respectively, over the entire space $\mathbb{R}^d$ as $C_0 = sup\left\{\|g(x)\| : x \in \mathbb{R}^D\right\}$ and $C_1 = sup\left\{|\nabla.g(x)| : x \in \mathbb{R}^D\right\}$. Also, Since K is C_0-universal, this operator is injective (Carmeli et al., 2010), as a result of the injectivity and having the linear operator we have the following relation:

$$\begin{aligned}\int K_{\mathbf{x}}|\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},.)|\,dp &= \int K_{\mathbf{x}}|\frac{\rho}{p}(\nabla.g + g.\nabla\log\rho)|\,dp \\ &= \int K_{\mathbf{x}}|(\nabla.g + g.\nabla\log\rho)|\,d\rho \le C_1 + C_0 * constant < \infty\end{aligned}$$

Now, let B be a ball in $\mathbb{R}^D$ centered at the origin with radius r. The set S_r is defined as the set of all points $y \in \mathbb{R}^d$ such that the Euclidean norm of y is less than or equal to r. On the other hand, let the set S be the surface of the ball B, which is defined as the set of all points y such that the Euclidean norm of y is equal to r. In other words, S is the set of points on the boundary of B. Then we define the quantity $m(r)$ as the smallest upper bound of the values of a given function ρ on the set S. In other words, $m(r)$ is the largest possible value of $\rho(y)$ for any point y on the surface of the ball B. We define the indication function $1_B(y)$ that is 1 if $y \in B$ and 0 if $y \notin B$. This function can be thought of as a "mask" that selects only the points in the set B and all other points to zero. By assumption, the quantity $r^{d-1}m(r)$ approaches zero as r approaches infinity. This means that the values of $\rho(y)$ on the surface of the ball B become arbitrarily small as the radius of B becomes arbitrarily large. Now we can define the expectation in terms of the sets B, S and 1_B functions as $\int K_{\mathbf{x}}\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},.)\,dp = \lim_{r\to\infty}\int 1_{B_r}K_{\mathbf{x}}\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},.)\,dp$. Then we have:

$$\begin{aligned}\int 1_{B_r}K_{\mathbf{x}}\zeta_{\rho_{(.|\mathbf{x})}}(\mathbf{y},.)\,dp &= \int 1_{B_r}K_{\mathbf{x}}\frac{\rho}{p}\left[(\nabla.g + g.\nabla\log\rho)\right]\,dp \\ &= \int_{B_r} K_{\mathbf{x}}\left[\rho\nabla.g + g.\nabla\rho\right]dy = \int_{B_r} K_{\mathbf{x}}\nabla.(\rho g)\,dy \\ &= \int_{S_r} K_{\mathbf{x}}\,\rho\, g\,.\,n\,dy \le m(r)\times c\times r^{D-1} \overset{r\to\infty}{\to} 0.\end{aligned}$$

Now, after this proof, we can rewrite Eq. (8) as follows:

$$\begin{aligned}D^2_{p,\rho}(\pi) &= \left\|\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\pi_{xy}}K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.)\right\|^2_{\mathcal{G}_K} \\ &= \left\|\mathbb{E}_{(\mathbf{x})\sim\pi_x}K_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim\pi_{.|x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.)\right\|^2_{\mathcal{G}_K} \\ &= \left\|\mathbb{E}_{(\mathbf{x})\sim\pi_x}K_{\mathbf{x}}\mathbf{r}_{p,\pi}(.,\mathbf{x})\right\|^2_{\mathcal{G}_K}\end{aligned}$$

Based on the Theorem 2 in Chwialkowski et al. (2016), for almost all π_x, the Kernel Stein Discrepancy between two probability density functions $p|x$ and $\pi|x$ is 0 if and only if they overlap. Thus, given $x \sim \pi_x$, $\left\|\mathbf{r}_{p,\pi}(.,\mathbf{x})\right\|^2_{\mathcal{G}^{D_y}} = 0$ if and only if $p_{.|x} = \pi_{.|x}$. The rest can be proved as the previous part. □

A.2. Proof of Proposition 2

Proof. To prove this proposition, we need to rewrite the $D^2_{p,\rho}(\pi)$ as follows:

$$\begin{aligned}D^2_{p,\rho}(\pi) &= \left\|\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\pi_{xy}}K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.)\right\|^2_{\mathcal{G}_K} \\ &= \left\langle\mathbb{E}_{\mathbf{xy}}K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.), \mathbb{E}_{\mathbf{x}'\mathbf{y}'}K_{\mathbf{x}'}\zeta_{\rho(.|\mathbf{x}')}\left(\mathbf{y}',.\right)\right\rangle_{\mathcal{G}_K} \\ &\overset{(*)}{=} \mathbb{E}_{\mathbf{xy}}\mathbb{E}_{\mathbf{x}'\mathbf{y}'}\left\langle K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.), K_{\mathbf{x}'}\zeta_{\rho(.|\mathbf{x}')}\left(\mathbf{y}',.\right)\right\rangle_{\mathcal{G}_K} \\ &\overset{(**)}{=} \mathbb{E}_{\mathbf{xy}}\mathbb{E}_{\mathbf{x}'\mathbf{y}'}\left\langle K^*_{\mathbf{x}'}K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.), \zeta_{\rho(.|\mathbf{x}')}\left(\mathbf{y}',.\right)\right\rangle_{\mathcal{G}_l^{D_y}} \\ &= \mathbb{E}_{\mathbf{xy}}\mathbb{E}_{\mathbf{x}'\mathbf{y}'}k\left(\mathbf{x},\mathbf{x}'\right)\frac{\rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{y}'|\mathbf{x}')}{p(\mathbf{y}|\mathbf{x})p(\mathbf{y}'|\mathbf{x}')}b_\rho\left((\mathbf{x},\mathbf{y}),\left(\mathbf{x}',\mathbf{y}'\right)\right),\end{aligned}$$

From the requirement 4 in Theorem 1, and due to the Bochner integrability of $K_{\mathbf{x}}\zeta_{\rho(.|\mathbf{x})}(\mathbf{y},.)$, at $(*)$, the inner product and the expectation are interchangeable.

In the second step $(**)$, the adjoint of the operator $K_{\mathbf{x}'}$, denoted as $K^*_{\mathbf{x}'}$, is used. The adjoint is defined as the operator that satisfies $\langle K_{\mathbf{x}'}f, g\rangle = \langle f, K^*_{\mathbf{x}'}g\rangle$ for all functions f, g in the domain of $K_{\mathbf{x}'}$. Taking advantage of the reproducing property of the kernel function $K_{\mathbf{x}}$ is used to rewrite the inner product in terms of the kernel function $k(\mathbf{x},\mathbf{x}')$. Thus we can write $K^*_{\mathbf{x}'}K_{\mathbf{x}} = K\left(\mathbf{x},\mathbf{x}'\right) = k\left(\mathbf{x},\mathbf{x}'\right)I$ and I is an identity operator. Thus $b_\rho\left((\mathbf{x},\mathbf{y}),\left(\mathbf{x}',\mathbf{y}'\right)\right)$ can be written as:

$$\begin{aligned}b_\rho\left((\mathbf{x},\mathbf{y}),\left(\mathbf{x}',\mathbf{y}'\right)\right) &= l\left(\mathbf{y},\mathbf{y}'\right)\mathbf{s}^T_\rho(\mathbf{y}\mid\mathbf{x})\mathbf{s}_\rho\left(\mathbf{y}'\mid\mathbf{x}'\right) + \sum_{i=1}^{D_y}\frac{\partial^2}{\partial y_i\partial y'_i}l\left(\mathbf{y},\mathbf{y}'\right) \\ &\quad + \mathbf{s}^T_\rho(\mathbf{y}\mid\mathbf{x})\nabla_{\mathbf{y}'}l\left(\mathbf{y},\mathbf{y}'\right) + \mathbf{s}^T_\rho\left(\mathbf{y}'\mid\mathbf{x}'\right)\nabla_{\mathbf{y}}l\left(\mathbf{y},\mathbf{y}'\right),\end{aligned}$$

□

References

Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*(2), 697–725.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., & Mackey, L. (2019). Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, *32*.

Berlinet, A., & Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Carmeli, C., De Vito, E., & Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, *4*(04), 377–408.

Carmeli, C., De Vito, E., Toigo, A., & Umanitá, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, *8*(01), 19–61.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., et al. (2019). Stein point Markov chain Monte Carlo. In *International conference on machine learning* (pp. 1011–1021). PMLR.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., & Oates, C. (2018). Stein points. In *International conference on machine learning* (pp. 844–853). PMLR.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., & Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, *28*.

Chwialkowski, K., Strathmann, H., & Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning* (pp. 2606–2615). PMLR.

Filippone, M., & Girolami, M. (2014). Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11), 2214–2226.

Fisher, M., Nolan, T., Graham, M., Prangle, D., & Oates, C. (2021). Measure transport with kernel Stein discrepancy. In *International conference on artificial intelligence and statistics* (pp. 1054–1062). PMLR.

Fisher, M. A., Oates, C., et al. (2022). Gradient-free kernel stein discrepancy. arXiv preprint arXiv:2207.02636.

Gorham, J., Duncan, A. B., Vollmer, S. J., & Mackey, L. (2019). Measuring sample quality with diffusions. *Annals of Applied Probability*, *29*(5), 2884–2928.

Gorham, J., & Mackey, L. (2015). Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, *28*.

Gorham, J., & Mackey, L. (2017). Measuring sample quality with kernels. In *International conference on machine learning* (pp. 1292–1301). PMLR.

Han, J., & Liu, Q. (2018). Stein variational gradient descent without gradient. In *International conference on machine learning* (pp. 1900–1908). PMLR.

Hodgkinson, L., Salomone, R., & Roosta, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. arXiv preprint arXiv:2001.09266.

Huggins, J., & Mackey, L. (2018). Random feature Stein discrepancies. *Advances in Neural Information Processing Systems*, *31*.

Hyvärinen, A., & Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*(4).

Jitkrittum, W., Kanagawa, H., & Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on uncertainty in artificial intelligence* (pp. 221–230). PMLR.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. In *Inst. ital. attuari, giorn. Vol. 4* (pp. 83–91).

Liu, Q., & Lee, J. (2017). Black-box importance sampling. In *Artificial intelligence and statistics* (pp. 952–961). PMLR.

Liu, Q., Lee, J., & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* (pp. 276–284). PMLR.

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in Neural Information Processing Systems*, *29*.

Matsubara, T., Knoblauch, J., Briol, F.-X., Oates, C., et al. (2021). Robust generalised Bayesian inference for intractable likelihoods. arXiv preprint arXiv:2104.07359.

Oates, C. J., Girolami, M., & Chopin, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *79*(3), 695–718.

O'donoghue, B., Chu, E., Parikh, N., & Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, *169*(3), 1042–1068.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S., Mackey, L., et al. (2022). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *84*, http://dx.doi.org/10.1111/rssb.12503.

Salakhutdinov, R. (2015). Learning deep generative models. *Annual Review of Statistics and Its Application*, *2*, 361–385.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281.

Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, *12*(7).

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth berkeley symposium on mathematical statistics and probability, Volume 2: probability theory. Vol. 6* (pp. 583–603). University of California Press.

Stein, C., Diaconis, P., Holmes, S., & Reinert, G. (2004). Use of exchangeable pairs in the analysis of simulations. In *Lecture notes-monograph series*, (pp. 1–26). JSTOR.

Szabó, Z., & Sriperumbudur, B. K. (2017). Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, *18*, 1–29.

Yang, J., Liu, Q., Rao, V., & Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International conference on machine learning* (pp. 5561–5570). PMLR.