

Advanced Kernel-Based Approaches for Robust Inference in Intractable Models

by

Elham Afzali

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics
University of Manitoba
Winnipeg

Copyright © 2025 by Elham Afzali

Abstract

Modern statistical and machine learning models often involve high-dimensional parameters, latent variables, or complex hierarchical structures, leading to likelihood functions that are both difficult to express in closed form and costly to differentiate. With data volumes growing and architectures such as deep neural networks becoming more complex, conventional likelihood-based methods struggle to provide reliable inference and model diagnostics, giving rise to intractable likelihoods and gradients.

In response to these challenges, this thesis develops and evaluates new Kernel Stein Discrepancy (KSD) methods for efficient and robust inference in intractable settings. First, it introduces Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD), a nonparametric goodness-of-fit test that requires no gradient information. GF-KCSD offers theoretical guarantees and strong empirical results, outperforming standard approaches in importance sampling scenarios.

Next, the thesis investigates GF-KCSD’s interpretability and error control under measurement errors, demonstrating its robustness and highlighting its ability to maintain Type-I error control and stable power even when data are noisy.

Finally, while KSD-based methods (or score-based methods in general) continue to gain traction, they often encounter mode blindness. To resolve this, the thesis proposes Mode-Sensitive KSD-Bayes (MS-KSD-Bayes), a weighted extension of KSD that balances multimodal posterior distributions within a generalized Bayesian framework. Across the Galaxy velocities and lung-cancer gene expression datasets,

MS-KSD-Bayes shows improved mode sensitivity and comparable or tighter posteriors relative to KSD-Bayes and gradient-based methods, within the light-tailed, moderate-dimensional regimes we study.

Collectively, these contributions form a gradient-free, mode-sensitive suite of KSD methodologies that significantly expand feasible statistical modeling for intractable likelihoods, ultimately enriching the toolkit for modern data analysis and inference.

Keywords: Intractable Likelihoods; Intractable Gradients; Kernel Stein Discrepancy; Mode Blindness; Measurement Errors; Multimodal Posteriors; Bayesian Inference; Goodness-of-Fit Testing.

Acknowledgments

First and foremost, I wish to express my profound gratitude to my supervisors, Dr. Liqun Wang and Dr. Saman Muthukumarana. Your insightful guidance, constant encouragement, and trust in my abilities were invaluable throughout my doctoral studies. Your mentorship was instrumental in shaping this research and has been pivotal in my academic development.

I extend my sincere thanks to my advisory committee members, Dr. Max Turgeon and Dr. Carson Leung, whose thoughtful feedback and dedicated support significantly enhanced the quality of this dissertation. I am also grateful to my external examiner, Dr. Jingjing Wu, for her time and constructive feedback.

My appreciation also extends to the faculty, staff, and my colleagues in the Department of Statistics at the University of Manitoba for fostering a supportive and intellectually stimulating academic environment.

I gratefully acknowledge the financial support I received during my studies from the Natural Sciences and Engineering Research Council of Canada (NSERC), the University of Manitoba (GETS), the Faculty of Science (SEGS), and the Department of Statistics (GESS). Their support made this research possible.

I want to give my deepest thanks to my family. This achievement would not have been possible without their love and support.

Dedication

*To my husband, **Vahid**,
for your unwavering love, patience, and belief in me.*

*To my parents, **Homa** and **Ali**,
for instilling in me a love of learning and for your lifelong support.*

*To my sisters, **Hoda** and **Elnaz**,
for your endless encouragement and unconditional love.*

Contents

Abstract	i
Acknowledgments	iii
Dedication	iv
List of Tables	x
List of Figures	xi
Contribution of Authors	xiv
1 Introduction	1
1.1 Core Challenges in Intractable and Gradient-Free Settings	3
1.2 Motivating Examples	4
1.3 Objectives and Research Questions	5
1.4 Main Contributions	5
1.5 Thesis Organization	7
Bibliography	8

2	Gradient-Free Kernel Conditional Stein Discrepancy Goodness of Fit Testing	11
2.1	Introduction	12
2.2	Background Methods	15
2.2.1	Kernel Stein Discrepancy (KSD)	16
2.2.2	Gradient Free Kernel Conditional Stein Discrepancy (GF-KCSD)	19
2.2.3	Explicit Form of GF-KCSD	25
2.2.4	Convergence Detection and Control	26
2.3	Experiments	29
2.3.1	Conditional Gaussian Model	31
2.3.2	Conditional Gaussian Mixture Model	32
2.4	Application	34
2.4.1	Gradient-Free Conditional Stein Importance Sampling	34
2.5	Conclusion	37
	Bibliography	39
3	Navigating Interpretability and Alpha Control in GF-KCSD Testing with Measurement Error: A Kernel Approach	45
3.1	Introduction	46
3.2	Foundational Concepts and Theoretical Background	50
3.2.1	The Stein Operator	50

3.2.2	Conceptual Backgrounds of the GF-KCSD	53
3.3	Methodology	56
3.3.1	Bootstrap Resampling	57
3.3.2	Measurement Errors	58
3.4	Experiment design	63
3.4.1	Synthetic Data Experiments	63
3.4.2	Real-world Data Experiments (Brain Magnetic Resonance Imaging (MRI))	70
3.5	Conclusion	75
	Bibliography	76
4	Correcting Mode Proportion Bias in Generalized Bayesian Inference via a Weighted Kernel Stein Discrepancy	80
4.1	Introduction	81
4.2	Background	83
4.2.1	Kernelized Stein Discrepancy (KSD)	84
4.2.2	Generalized Bayesian Inference with KSD	87
4.3	Methodology	90
4.3.1	Diffusion Stein operator	90
4.3.2	Mode-sensitive diffusion term	92
4.3.3	Vector-valued RKHS and matrix-valued kernels	93

4.3.4	Mode-Sensitive Kernel Stein Discrepancy (MS-KSD)	94
4.3.5	MS-KSD-Bayes posterior	95
4.3.6	Generalized Bayesian inference for exponential families . . .	96
4.4	Experiment Design	98
4.4.1	Galaxy Velocity Data Experiments	100
4.4.2	Lung Cancer Gene Expression Experiments	102
4.4.3	Gaussian Location Experiment: Unimodal Setting with Con- tamination	104
4.5	Conclusion and Future Work	106
	Bibliography	108
5	Discussion and Conclusion	114
5.1	Key Findings and Contributions	115
5.2	Future Work and Directions	116
5.3	Concluding Remarks	118
A		120
A.1	Proofs of Theoretical Results in Chapter 2	121
A.1.1	Proof of Theorem 1	121
A.1.2	Proof of Proposition 2	123
A.2	Proofs of Theoretical Results for Chapter 4	124
A.2.1	Proof of Proposition 4	124

A.2.2	Proof of Proposition 5	125
A.2.3	Proof of Proposition 6	126
Bibliography		129

List of Tables

3.1	Comparison of GF-KCSD and KCSD Test Results for Brain MRI Data.	73
-----	---	----

List of Figures

2.1	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. .	31
2.2	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.	33

2.3	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 $\rho(y x) \sim \mathcal{N}(y; x + \nu, \eta)$	38
3.1	This figure presents a comparative analysis of Type-I error rates and rejection rates across four distinct tests, all conducted at a significance level of $\alpha = 0.05$. The first row depicts scenarios devoid of measurement errors, while subsequent rows introduce measurement errors. From the second to the fourth row, these errors follow a normal distribution with mean 0 and standard deviations of 0.5, 1.0, and 2.0, respectively. The final row involves measurement errors modeled by a $\chi^2(5)$. (a): In these contexts, when H_0 is true, the GF-KCSD method demonstrates robustness in maintaining control over Type-I errors, effectively managing fluctuations attributable to sampling noise. (b): Under the alternative hypothesis (H_1) the efficacy of the GF-KCSD in discerning global differences is suboptimal. (c): Independent of the presence or absence of measurement errors, GF-KCSD consistently shows a highly effective in identifying local differences.	68
3.2	Comparison of Target and Predicted Age Distributions. The predicted age distribution is generated using a Mixture Density Network (MDN) with 20 gaussian components trained on preprocessed brain MRI data.	73

4.1	Posterior density estimates for the Galaxy dataset. Top: data histogram. Middle: KSD–Bayes (insensitive to mode proportions). Bottom: MS–KSD–Bayes (captures the multimodal structure with more faithful proportions in this example).	101
4.2	Generalized posterior density estimates for the lung cancer gene expression dataset (Setting: $p = 10$, $L = 9$, prior scale = 81, prior decay = 1.2). From right to left: (1) Histogram of the observed gene expression data; (2) Posterior density estimated via MS–KSD–Bayes; (3) Posterior density estimated via KSD–Bayes. Unlike KSD–Bayes, MS–KSD–Bayes captures the bimodal structure more clearly in this example.	104
4.3	Posterior distributions for a Gaussian location model under varying contamination proportions (ϵ , top row) and outlier displacements (y , bottom row). Columns: (left) standard Bayes with a conjugate prior, (center) KSD–Bayes, (right) MS–KSD–Bayes. As ϵ or y increase, the first two methods shift away from $\theta_\star = 1$, while MS–KSD–Bayes remains comparatively concentrated near $\theta_\star$	105

Contribution of Authors

This thesis follows a "Sandwich" structure. Chapters 2 through 4 are adapted from three distinct journal articles, all of which have been published or submitted for review. Following are the titles of the prepared research articles, the manuscript review status at the time of the thesis submission, the authors' contributions, and the corresponding chapters in the thesis:

1. Chapter 2 is based on the article titled “*Gradient-Free Kernel Conditional Stein Discrepancy goodness of fit testing*”, published in *Machine Learning With Applications* (2023).

DOI: <https://www.sciencedirect.com/science/article/pii/S2666827023000166>

Author contributions:

- **Elham Afzali** (Thesis Author): Conceptualization, Visualization, Methodology, Software, Draft the Manuscript.
- **Saman Muthukumarana** (Advisor): Supervision, Reviewing, Validation, Editing the manuscript.

2. Chapter 3 is based on the manuscript titled “*Navigating interpretability and*

alpha control in GF-KCSD testing with measurement error: A Kernel approach”, published in *Machine Learning With Applications* (2024).

DOI: <https://www.sciencedirect.com/science/article/pii/S2666827024000574>

Author contributions:

- **Elham Afzali** (Thesis Author): Conceptualization, Visualization, Methodology, Software, Draft the Manuscript, Data Curation.
- **Liqun Wang** (Advisor): Methodology, Supervision, Reviewing, Validation, Editing the manuscript.
- **Saman Muthukumarana** (Advisor): Supervision, Reviewing, Validation, Editing the manuscript.

3. Chapter 4 is based on the article titled “*Correcting Mode Proportion Bias in Generalized Bayesian Inference via a Weighted Kernel Stein Discrepancy*”, currently under review at *Bayesian Analysis* (Under review, 2025).

DOI: <https://arxiv.org/abs/2503.02108>

Author contributions:

- **Elham Afzali** (Thesis Author): Conceptualization, Visualization, Methodology, Software, Draft the Manuscript, Data Curation.
- **Liqun Wang** (Advisor): Supervision, Reviewing, Validation, Editing the manuscript.
- **Saman Muthukumarana** (Advisor): Supervision, Reviewing, Validation, Editing the manuscript.

Chapter 1

Introduction

In recent years, rapid technological progress and expanded data-gathering efforts have produced an unprecedented surge of large-scale, high-dimensional datasets in fields such as genomics, neuroscience, economics, and the social sciences. These extensive data offer enormous promise for predictive modeling—leveraging historical patterns to anticipate future outcomes, thereby enhancing decision-making across a range of domains. Yet, fully capitalizing on their complexity brings forth substantial computational and methodological hurdles.

A core difficulty arises when these modern models rely on intractable likelihoods for which neither the likelihood nor its normalizing constant can be conveniently expressed or efficiently computed (Andrieu and Roberts, 2009; Filippone and Girolami, 2014). Such situations often occur in hierarchical Bayesian structures, latent variable formulations, or deep generative models, rendering conventional inference methods computationally expensive or even impossible. For instance, in a Bayesian setting with parameters $\theta \in \mathbb{R}^d$, the posterior distribution

$$\pi(\theta \mid y) = \frac{p(y \mid \theta) \pi_0(\theta)}{\int p(y \mid \theta') \pi_0(\theta') d\theta'}$$

can become computationally prohibitive whenever $p(y \mid \theta)$ lacks a closed form or requires large-scale numerical integration.

In addition, many widely used inference routines—ranging from Markov chain Monte Carlo (MCMC) (Neal et al., 2011) to stochastic variational inference (Hoffman et al., 2013)—depend on $\nabla_{\theta} \log p(y \mid \theta)$ to efficiently explore the parameter space. Where $\nabla_{\theta} \log p(y \mid \theta)$ is unavailable or too expensive to compute, these score-based techniques may fail to provide reliable or timely inferences, motivating the pursuit of gradient-free solutions.

Recently, kernel-based Stein methods (Stein, 1972; Oates et al., 2017) have shown particular promise for addressing such obstacles, enabling goodness-of-fit assessments and related tasks without requiring a fully normalized likelihood. The Kernel Stein Discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016) in particular leverages a Stein operator that bypasses direct likelihood normalization. However, its classical forms—including Kernel Conditional Stein Discrepancy (KCSD) (Jitkrittum et al., 2020)—still assume that gradients of the log-likelihood are tractable. This thesis addresses that gap by developing gradient-free kernel-based frameworks specifically tailored to conditional modeling, while also tackling measurement error and multimodal targets—two practical challenges frequently encountered in modern data analysis.

1.1 Core Challenges in Intractable and Gradient-Free Settings

Computational Bottlenecks. In high-dimensional Bayesian problems, evaluating $\log p(y \mid \theta)$ or its gradient at every iteration can be a major bottleneck. High-dimensional parameter spaces, multiple hierarchical layers, and the presence of latent variables commonly lead to explosion in runtime and sensitivity to tuning, making standard MCMC or variational techniques (Andrieu and Roberts, 2009; Filippone and Girolami, 2014) computationally overwhelming.

Model Diagnosis and Goodness-of-Fit. Classical tests of model appropriateness (for example, likelihood ratio tests or distribution-based tests) often presuppose a tractable likelihood function or cumulative distribution function. Under intractable likelihoods, these approaches do not apply directly. Although the Stein discrepancy resolves the need for a fully normalized target, it typically requires $\nabla \log p$. Hence, achieving a completely gradient-free and likelihood-free discrepancy remains an open challenge in the domain of model assessment.

Conditional Models. A significant subset of machine learning and statistical tasks involves modeling $p(y \mid x)$. Examples include regression with potentially complex noise structures, multi-task learning, and conditional generative modeling. Current Stein-based methods like KSD or KCSD (Liu et al., 2016; Chwialkowski et al., 2016; Jitkrittum et al., 2020) must be adapted so that they can fully capture these conditional dependencies without requiring the log-derivative $\nabla_y \log p(y \mid x)$ to be tractable.

Score-Based Approaches and Mode Blindness. Even if one manages to avoid explicit gradients, standard KSD-type or variational methods can fail to preserve distinct modes of a posterior distribution—a phenomenon sometimes called mode blindness (Yang et al., 2018; Barp et al., 2019). This oversight is especially problematic for mixture models or other inherently multimodal settings, where ignoring minor modes skews parameter estimates and can lead to misguided inferences.

1.2 Motivating Examples

These challenges are not merely theoretical; they appear frequently in many applied domains. For instance, hierarchical Bayesian models in fields like genomics often involve thousands of parameters and latent variables, making gradient-based MCMC computationally infeasible (Oates et al., 2017; Filippone and Girolami, 2014). Similarly, deep generative models such as Variational Autoencoders (VAEs) use implicit distributions defined by complex transformations, where the log-likelihood gradient is not available in a straightforward closed form (Kingma et al., 2013). Moreover, in conditional inference tasks common to personalized medicine, modeling an outcome y given high-dimensional features x can make the required conditional gradient, $\nabla_y \log p(y \mid x)$, prohibitively expensive to compute (Jitkrittum et al., 2020). Finally, in mixture models used to identify distinct sub-populations, such as different cancer subtypes from gene expression data, score-based inference methods can suffer from mode blindness. This causes them to incorrectly estimate the proportions of each subtype, potentially masking the presence of a smaller but clinically important group. These examples highlight the clear motivation for the robust, gradient-free, and mode-sensitive methods developed in this thesis.

1.3 Objectives and Research Questions

Given the above challenges, this thesis is driven by the following key questions:

1. How can we formulate a Stein-based discrepancy measure for conditional distributions without relying on $\nabla \log p(y | x)$?
2. What theoretical guarantees ensure that such a gradient-free approach accurately identifies convergence or divergence in high-dimensional contexts?
3. How can we incorporate measurement error while preserving Type-I error control and interpretability in goodness-of-fit tests?
4. What strategies can overcome mode blindness, ensuring that multi-modal posterior structures remain well-captured in gradient-free Stein inference?

Addressing these questions paves the way for a broader application of Stein-based tools where explicit gradient evaluations and tractable normalizing constants are simply unavailable.

1.4 Main Contributions

This thesis presents several advancements aimed at tackling the described challenges:

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

We propose a novel kernel-based discrepancy measure that extends the conventional Kernel Conditional Stein Discrepancy (Jitkrittum et al., 2020) to a

fully gradient-free framework. By introducing an auxiliary distribution whose log-derivatives are tractable, GF-KCSD can detect divergence from a conditional target without needing $\nabla \log p(y | x)$. The method inherits theoretical properties from classical Stein approaches and retains strong power for identifying model misspecifications.

2. Robustness Under Measurement Error.

Real data often contain noisy or corrupted observations, posing additional risks to model validity. We extend GF-KCSD to account for measurement error by developing bootstrap procedures and kernel selection strategies that preserve Type-I error rates. Our analysis reveals how moderate error levels affect test power and provides recommendations for robust practice (Gorham and Mackey, 2015; Fisher et al., 2022).

3. Mode-Sensitive KSD-Bayes (MS-KSD-Bayes).

To address multi-modality in posterior distributions, we develop a diffusion-weighted, score-only Stein loss with a θ -independent, tail-vanishing weight. By carefully reweighting contributions across modes, the method restores sensitivity to mixing proportions so the generalized posterior retains distinct modes rather than collapsing them (Yang et al., 2018; Barp et al., 2019). This method preserves the Stein identity under standard boundary conditions while remaining likelihood-free and compatible with unnormalized models. Under mild regularity, we give a kernel representation and an $O(n^2)$ V-statistic estimator. Empirically—on multimodal and contaminated-unimodal benchmarks—MS-KSD-Bayes recovers well-separated mode proportions and matches KSD-Bayes in unimodal/robust regimes.

4. Applications and Empirical Validation.

We comprehensively test these new methods in both controlled (synthetic) experiments and real-world settings, including Brain MRI dataset, galaxy velocity data and gene expression examples. Our empirical studies confirm that GF-KCSD successfully handles intractable gradients in conditional models, while MS-KSD-Bayes offers superior performance in detecting and quantifying posterior modes.

Together, these contributions expand the scope of Stein-based inference from likelihood-free to gradient-free methods that systematically handle multimodality, noisy observations, and conditional dependencies.

1.5 Thesis Organization

The remainder of this dissertation is structured as follows:

- **Chapter 2: Gradient-Free Kernel Conditional Stein Discrepancy.**

We introduce GF-KCSD, elaborating on its theoretical foundations and presenting initial experiments that illustrate its capacity to detect divergence even when $\nabla \log p(y \mid x)$ is unavailable.

- **Chapter 3: Measurement Error and Robust GF-KCSD Testing.**

Focusing on noisy or corrupted data scenarios, we analyze how GF-KCSD behaves under varying degrees of measurement error. We describe bootstrap-based variance controls and kernel tuning procedures that safeguard Type-I error control and interpretability.

- **Chapter 4: Correcting Mode Proportion Bias via Mode-Sensitive KSD-Bayes.**

Here, we tackle the issue of multi-modal posterior structures. We detail the diffusion-weighted Stein framework (MS-KSD-Bayes) give its RKHS/kernel representation and an exponential-family conjugacy result, and demonstrate empirical gains on synthetic and real data.

- **Chapter 5: Discussion and Future Directions.**

We conclude by synthesizing the main theoretical results and empirical findings, reflecting on their implications for large-scale Bayesian modeling and proposing future directions for integrating our methods with deep learning frameworks and advanced measurement-error modeling.

Through these chapters, we highlight how gradient-free, mode-sensitive kernel-based methods can substantially enhance inference in otherwise intractable models, facilitating robust goodness-of-fit tests, improved sampling protocols, and reliable uncertainty quantification in contemporary data analysis.

Bibliography

- Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725. 1, 3
- Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32. 4, 6

- Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR. 2, 3
- Filippone, M. and Girolami, M. (2014). Pseudo-marginal bayesian inference for gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226. 1, 3, 4
- Fisher, M. A., Oates, C., et al. (2022). Gradient-free kernel stein discrepancy. *arXiv preprint arXiv:2207.02636*. 6
- Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein’s method. *Advances in Neural Information Processing Systems*, 28. 6
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347. 2
- Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR. 2, 3, 4, 5
- Kingma, D. P., Welling, M., et al. (2013). Auto-encoding variational bayes. 4
- Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR. 2, 3
- Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2. 2

- Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718. 2, 4
- Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, pages 583–602. 2
- Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR. 4, 6

Chapter 2

Gradient-Free Kernel Conditional Stein Discrepancy Goodness of Fit Testing

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 requires the computation of model derivatives and is therefore impractical for complex statistical models. Our method, known as Gradient-Free Kernel Conditional Stein Discrepancy, does not require the calculation of derivatives, making 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.

2.1 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 and Smirnov (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 and 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 and Mackey’s approach (2015). To deal with this challenge, Liu et al. (2016) and Chwialkowski et al. (2016) introduced a likelihood-free method that is applicable to high-dimensional distributions. Their work was 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. Liu et al. (2016) and Chwialkowski 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 et al. (2018) extended the kernel Stein discrepancy test in such a way that can be used for the discrete domain. Jitkrittum et al. (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 (Liu and Lee, 2017; Chen et al., 2018, 2019; Barp et al., 2019; Matsubara et al., 2021; Hodgkinson et al., 2020; Riabiz et al., 2022; Fisher et al., 2021). 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 and Roberts, 2009; Filippone and Girolami, 2014).

To overcome this problem, Liu and Wang (2016) and Han and Liu (2018) proposed the Stein variational gradient descent by employing Gradient-Free Stein operators. Fisher 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 to be effec-

tive, 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 2.2.2) generalizes the kernel conditional Stein discrepancy (KCSD) (Jitkrittum et al., 2020) and the Gradient-Free Fernel Stein Discrepancy (GF-KSD) (Fisher et al., 2022).

Paper outline Section 3.2 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 2.3 we provide experiments using GF-KCSD. Section 2.4 contains the application of this method in importance sampling and the conclusion of the paper is discussed in Section 2.5.

2.2 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 Liu et al. (2016) and Chwialkowski et al. (2016) which is the base well-known goodness-of-fit test we will discuss in Section 2.2.1. In Section 2.2.2, 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.

2.2.1 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 et al., 2017), which corresponds to the gradient of the log-likelihood with regard to the input (Hyvärinen and 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 $\mathbf{g} \in \mathcal{G}^D$, $\mathbf{g} : \mathcal{X} \rightarrow \mathbb{R}^D$, the Langevin Stein operator of differentiable functions is defined as:

$$\mathcal{T}_p \mathbf{g}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})^T \mathbf{g}(\mathbf{x}) + \nabla_{\mathbf{x}}^T \mathbf{g}(\mathbf{x}), \quad (2.1)$$

where $\mathcal{T}_p \mathbf{g}(\mathbf{x}) \in \mathbb{R}$. In this case, these test 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 in Gorham and Mackey (2015, Proposition 1). In particular, we have

$$\mathbb{E}_{\mathbf{x} \sim p}[\mathcal{T}_p \mathbf{g}(\mathbf{x})] = 0. \quad (2.2)$$

Equation (2.2), 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) and Liu et al. (2016), is as follows:

$$\mathcal{S}_p(\pi, \mathcal{T}_p, \mathcal{G}) = \sup_{\|\mathbf{g}\|_{\mathcal{G}^D} \leq 1} |\mathbb{E}_{\mathbf{x} \sim \pi} \mathcal{T}_p \mathbf{g}(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p} \mathcal{T}_p \mathbf{g}(\mathbf{x})|. \quad (2.3)$$

The Stein discrepancy thus emphasizes the score difference between p and π . To make it easier to compute and analyze, Chwialkowski et al. (2016); Liu et al. (2016); Oates et al. (2017); Gorham and Mackey (2017) have applied the use of Reproducing Kernel Hilbert Spaces (RKHS) (Berlinet and 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 $k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ 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 et al., 2011), as long as $\mathbb{E}_{\mathbf{x} \sim \pi} \|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log \pi(\mathbf{x})\|_2^2 < \infty$ based on Chwialkowski et al. (2015, Theorem 2.2), $\mathcal{S}_p(\pi, \mathcal{T}_p, \mathcal{G}) = 0$ 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 p . Define $b_p(\mathbf{x}, \mathbf{x}')$ as follows:

$$\begin{aligned} b_p(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') \mathbf{s}_p^T(\mathbf{x}) \mathbf{s}_p(\mathbf{x}') + \sum_{i=1}^D \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_i \partial x'_i} \\ &\quad + \mathbf{s}_p^T(\mathbf{x}) \nabla_{x'} k(\mathbf{x}, \mathbf{x}') + \nabla_x k(\mathbf{x}, \mathbf{x}')^T \mathbf{s}_p(\mathbf{x}'), \end{aligned} \quad (2.4)$$

where the $\sum_{i=1}^D \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial x_i \partial x'_i} = \text{trace}(\nabla_{x,x'} k(\mathbf{x}, \mathbf{x}'))$. 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 Equations (2.5) and (2.6), respectively.

U-statistics:

$$\hat{S}_p^2(\pi) = \frac{1}{n(n-1)} \sum_{i \neq j} b_p(x_i, x_j). \quad (2.5)$$

V-statistics:

$$\hat{S}_p^2(\pi) = \frac{1}{n^2} \sum_{i,j=1}^n b_p(x_i, x_j), \quad (2.6)$$

where n is the sample size. Chwialkowski et al. (2016) 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.

2.2.2 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 following 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 point 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 $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{D_x} \times \mathbb{R}^{D_y}$. Let $\mathcal{P}(\mathbb{R}^D)$ be a set of probability distributions on $\mathbb{R}^D$. For $q, q_0 \in \mathcal{P}(\mathbb{R}^d)$, we write $q \ll q_0$ to mean that for every measurable set $A \subseteq \mathbb{R}^d$, $q_0(A) = 0$ implies $q(A) = 0$ (i.e., q is absolutely continuous w.r.t. q_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 $\mathcal{Z}_n = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \stackrel{i.i.d.}{\sim} \pi_{xy}$ 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:

$$H_0 : p_{\mathbf{xy}} \stackrel{\pi_x}{=} \pi_{\mathbf{xy}} \quad vs \quad H_1 : p_{\mathbf{xy}} \stackrel{\pi_x}{\neq} \pi_{\mathbf{xy}}. \quad (2.7)$$

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

$$\forall \mathbf{x} \in S, \quad p(\cdot | \mathbf{x}) \neq \pi(\cdot | \mathbf{x}).$$

The goodness-of-fit test in Equation (2.7) 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 similarities or dissimilarities between them. Reproducing kernels also provide 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 advanced details, please refer to Carmeli et al. (2010, 2006); Szabó and Sriperumbudur (2017), and Sriperumbudur et al. (2011). Let $\mathcal{X}$ be a nonempty set, $\mathcal{W}$ be a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, 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 $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{W})$ 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 $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$, $\{\mathbf{w}_i\}_{i=1}^n \subset \mathcal{W}$ 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 $\mathbf{w} \in \mathcal{W}$, define the linear operator function $K_x \mathbf{w} = K(\cdot, \mathbf{x})$ 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) $\mathcal{G}_K$ such that $K_{\mathbf{x}} \in \mathcal{L}(\mathcal{W}; \mathcal{G}_K)$, 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 $K_{\mathbf{x}}$ is defined as $K_{\mathbf{x}}^* : \mathcal{G}_K \rightarrow \mathcal{W}$.

Taking into account that $\mathbb{C}(\mathcal{X}; \mathcal{W}) : \mathcal{X} \rightarrow \mathcal{W}$ is a vector space of continuous functions, and that $\mathcal{C}_0(\mathcal{X}; \mathcal{W})$ is a subspace of the $\mathbb{C}(\mathcal{X}; \mathcal{W})$ whose elements vanish with the uniform norm at infinity. This means that the elements of $\mathcal{C}_0(\mathcal{X}; \mathcal{W})$ become arbitrarily close to 0 as the input values get farther and farther away from the origin, *i.e.* $\|g(\mathbf{x})\|_{\mathcal{W}} \rightarrow 0$ as $\|\mathbf{x}\| \rightarrow \infty$. We can consider a $\mathcal{W}$ -reproducing kernel K to be a $\mathcal{C}_0$ if $\mathcal{G}_K \subset \mathcal{C}_0(\mathcal{X}; \mathcal{W})$. 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 $\mathcal{C}_0$ -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 $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be connected with the RKHS $\mathcal{G}_l$. For two vectors $\mathbf{a} = (a_1, \dots, a_{D_y})$, $\mathbf{m} = (m_1, \dots, m_{D_y})$ 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 $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{G}_l^{D_y}$ that is $\mathcal{W} = \mathcal{G}_l^{D_y}$. Assume a real-valued kernel connected with RKHS $\mathcal{G}_k$ as $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

Definition 1. *Let the gradient-free Stein operator be the one defined in Fisher et al. (2022, Definition 1), 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 $g : \mathbb{R}^D \rightarrow \mathbb{R}^D$ in a way that its first derivatives are bounded is defined as:*

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

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 (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_0 -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 the Definition 1, the discrepancy between two conditional distributions p and π is defined as:*

$$D_{p,\rho}^2(\pi) = \left\| \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \pi_{xy}} K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot) - \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim p_{xy}} K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot) \right\|_{\mathcal{G}_K}^2, \quad (2.8)$$

where $\zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot) = \frac{\rho}{p} [l(\mathbf{y}, \cdot) \nabla_{\mathbf{y}} \log \rho(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{y}} l(\mathbf{y}, \cdot)] \in \mathcal{G}_l^{D_y}$. $D_{p,\rho}^2(\pi)$ is referred to as Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). The second expectation term in $D_{p,\rho}^2(\pi)$ (under the distribution P) is zero by considering the Stein operator as a non-standard instance of the density method in Stein et al. (2004). In this case, the first expectation term in $D_{p,\rho}^2(\pi)$, which is the expectation with respect to the distribution $\pi \in \mathcal{P}(\mathbb{R}^D)$ is zero if and only if $p \stackrel{\pi_x}{=} \pi$; 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 $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{G}_l^{D_y})$ (where $\mathcal{L}$ is a Banach space) be two positive definite C_0 -universal kernels. Consider the conditional witness function defined as $\mathbf{r}_{p,\pi}(\omega|\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim \pi(\cdot|\mathbf{x})} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \omega) \in \mathbb{R}^{D_y}$ and for all $\mathbf{x}$, $\mathbf{r}_{p,\pi}(\cdot|\mathbf{x}) \in \mathcal{G}_l^{D_y}$, 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. $\int_{\mathcal{X}} \|\mathbf{r}_{p,\pi}(\cdot|\mathbf{x})\|_{\mathcal{G}_l^{D_y}}^2 \pi_x(\mathbf{x}) d\mathbf{x} < \infty;$

4. $\int \|K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}|\cdot)\|_{\mathcal{G}_K} d\pi_{xy} < \infty.$

Then $D_{p,\rho}^2(\pi) = 0$ if and only if $p \stackrel{\pi_x}{=} \pi$, for almost all π_x , $\mathbf{x} \in \mathcal{X}$, $p(\cdot|\mathbf{x}) = \pi(\cdot|\mathbf{x})$.

A.1.1 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 Multiquadratic kernel (IMQ), Gaussian kernels, Laplace kernel, for example, are 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:

$$l(\mathbf{y}, \mathbf{y}') = (\sigma_y^2 + \|\mathbf{y} - \mathbf{y}'\|^2)^{-\beta}, \quad \beta \in (0, 1), \sigma_y \in (0, \infty). \quad (2.9)$$

2.2.3 Explicit Form of GF-KCSD

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

Proposition 2. *For a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, 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:*

$$D_{p,\rho}^2(\pi) = \int \int k(\mathbf{x}, \mathbf{x}') \frac{\rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{y}'|\mathbf{x}')}{p(\mathbf{y}|\mathbf{x})p(\mathbf{y}'|\mathbf{x}')} b_\rho((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) d\pi_{\mathbf{x}\mathbf{y}} d\pi_{\mathbf{x}'\mathbf{y}'}, \quad (2.10)$$

where

$$\begin{aligned} b_\rho((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) &= l(\mathbf{y}, \mathbf{y}') \mathbf{s}_\rho^T(\mathbf{y}|\mathbf{x}) \mathbf{s}_\rho(\mathbf{y}'|\mathbf{x}') + \sum_{i=1}^{D_y} \frac{\partial^2 l(\mathbf{y}, \mathbf{y}')}{\partial y_i \partial y'_i} \\ &\quad + \mathbf{s}_\rho^T(\mathbf{y}|\mathbf{x}) \nabla_{\mathbf{y}'} l(\mathbf{y}, \mathbf{y}') + \nabla_{\mathbf{y}} l(\mathbf{y}, \mathbf{y}')^T \mathbf{s}_\rho(\mathbf{y}'|\mathbf{x}'). \end{aligned} \quad (2.11)$$

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 $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \stackrel{i.i.d.}{\sim} \pi_{xy}$ an unbiased, consistent estimator for Equation (2.10)

is given by

$$\widehat{D}_{p,\rho}^2(\pi) = \frac{1}{n(n-1)} \sum_{i \neq j} B_\rho((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)). \quad (2.12)$$

The proof is in A.1.2. From Equation (2.11) 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.

2.2.4 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 and 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 $h : \mathbb{R}^D \rightarrow \mathbb{R}^D$ is defined as $L(h) = \sup_{y \neq y'} \frac{\|h(y) - h(y')\|}{\|y - y'\|}$, 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 $f : \mathbb{R}^D \rightarrow \mathbb{R}$ as:

$$W_1(\pi, p; f) = \sup_{L(h) \leq 1} \left| \int h f d\pi - \int h f dp \right|, \quad (2.13)$$

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 be conditional distributions in $\mathcal{P}(\mathbb{R}^D)$ and ρ be absolutely continuous with respect to p , the score function of ρ be 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 $\int \|K_x \nabla \log \rho\|^2 d\pi_n < \infty$, $\int \left(\frac{\rho}{p}\right)^2 d\pi_n \in (0, \infty)$, $\int \|K_x \nabla \log \rho\|^2 \left(\frac{\rho}{p}\right) d\pi_n < \infty$, $\int \|y\| \frac{\rho(y)}{p(y)} d\pi_n(y) < \infty$. Consider l and k as kernels such that all their derivatives exist and are bounded. Then

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

According to Equation (2.13), 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 $\rho \in \mathcal{P}(\mathbb{R}^D)$, $p \in \mathcal{P}(\mathbb{R}^D)$, and $\inf_{y \in \mathbb{R}^D} \frac{\rho(y|x)}{p(y|x)} > 0$ such that p is continuous. Consider a sequence $\pi_n \subset \mathcal{P}(\mathbb{R}^D)$ for which the preconditions of Theorem 2 are satisfied. Then we have:*

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

The convergence detection necessitates extra conditions on ρ . Consider the weight matrix $M(y) = \left(\frac{\rho(y)}{p(y)}\right)I$ which is a position-dependent matrix and Lipschitz, ex-

plained in (Gorham et al., 2019, proposition 8). 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 2.3, we shall see that proper choices for kernel parameters and ρ lead to a higher test power. Note that in Sections 2.3 and 2.4 we consider all the kernels to be inverse multi-quadric (IMQ) kernels.

2.3 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 (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 need 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 $\beta = 0.5$. However, 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 2.1, some sequences 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) lie in the set of 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 2.3.1 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 2.3.2 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.

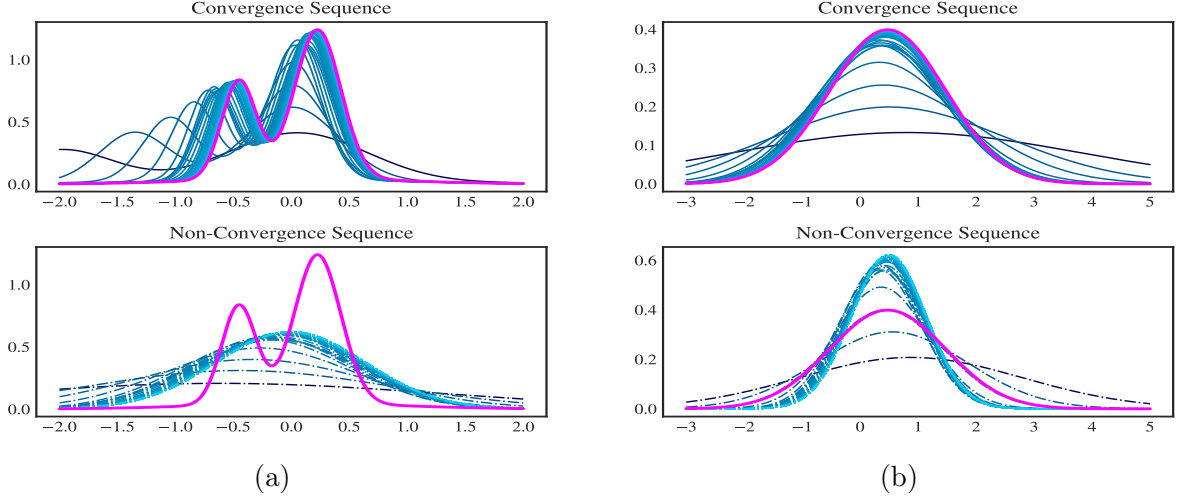


Figure 2.1: Experimental slant sequences π_n with fixed kernel parameters ($\sigma = 1$, $\beta = 0.5$). The first row shows the sequences (straight line) converge to the target distribution p (magenta), while the second row depicts the sequences (dash-dot) that converge to different Gaussian distribution. (a) The target distribution is conditional Gaussian mixture model. (b) The target distribution is conditional Gaussian model.

2.3.1 Conditional Gaussian Model

Problem setting: We take into account a basic univariate problem where the data generating distribution is $\pi_x(x) = \text{Uniform}(-2, 2)$, 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 in 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 $\rho_{light} \sim \mathcal{N}(0.5, 0.64)$.

2.3.2 Conditional Gaussian Mixture Model

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

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

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

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 2.2. It is evident

that two selected ρ s (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 $n > 50$.

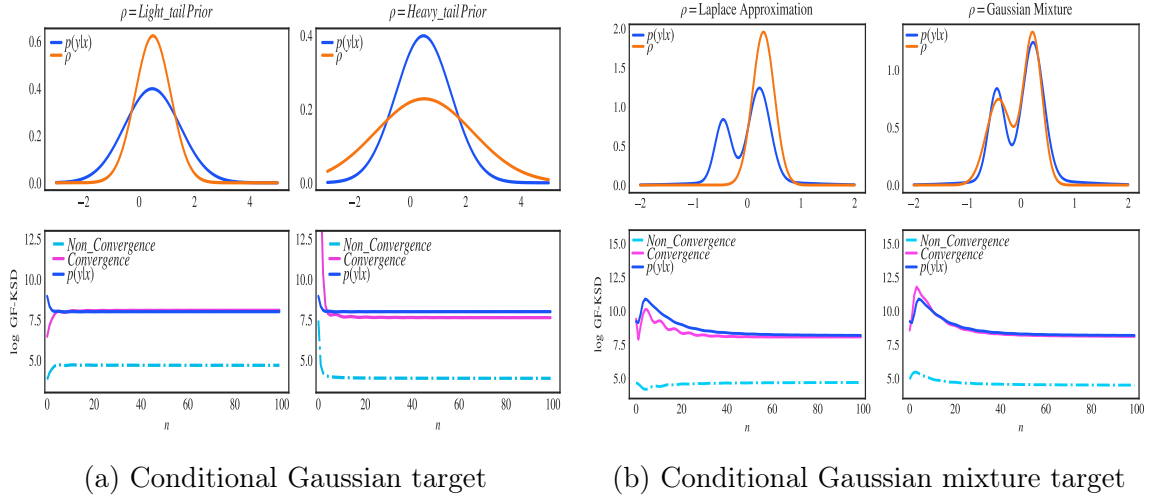


Figure 2.2: Experimental evaluation of the gradient-free kernel conditional Stein discrepancy on the sequence π_n . The first row displays the distribution of the target model p and its corresponding ρ . (a) displays the assessed GF-KCSD for the conditional Gaussian target. (b) depicts the evaluated 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.

2.4 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 2.4.1). This extension expands the scope of existing algorithms to statistical models where certain derivatives of p are either prohibitively expensive or intractable.

2.4.1 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 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 (Liu and Lee, 2017; Hodgkinson et al., 2020). 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 (Fisher et al., 2022, Theorem 3) 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 $\int \exp(\lambda \|K_x \nabla \log \rho\|^2) d\rho < \infty$ for some positive λ . For independent samples (y_n) from ρ , the optimal weights are as follows:*

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

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

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 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 et al., 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 2.3 displays the result of each scenarios. These results indicate that, "in correcting sample bias", gradient-free conditional Stein sampling performs similarly to—or even better than—conditional Stein importance sampling.

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 similarly 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.

2.5 Conclusion

We propose Gradient-Free Conditional Stein Discrepancy which is a measure of the difference between two distributions, and it can be used in statistical testing and

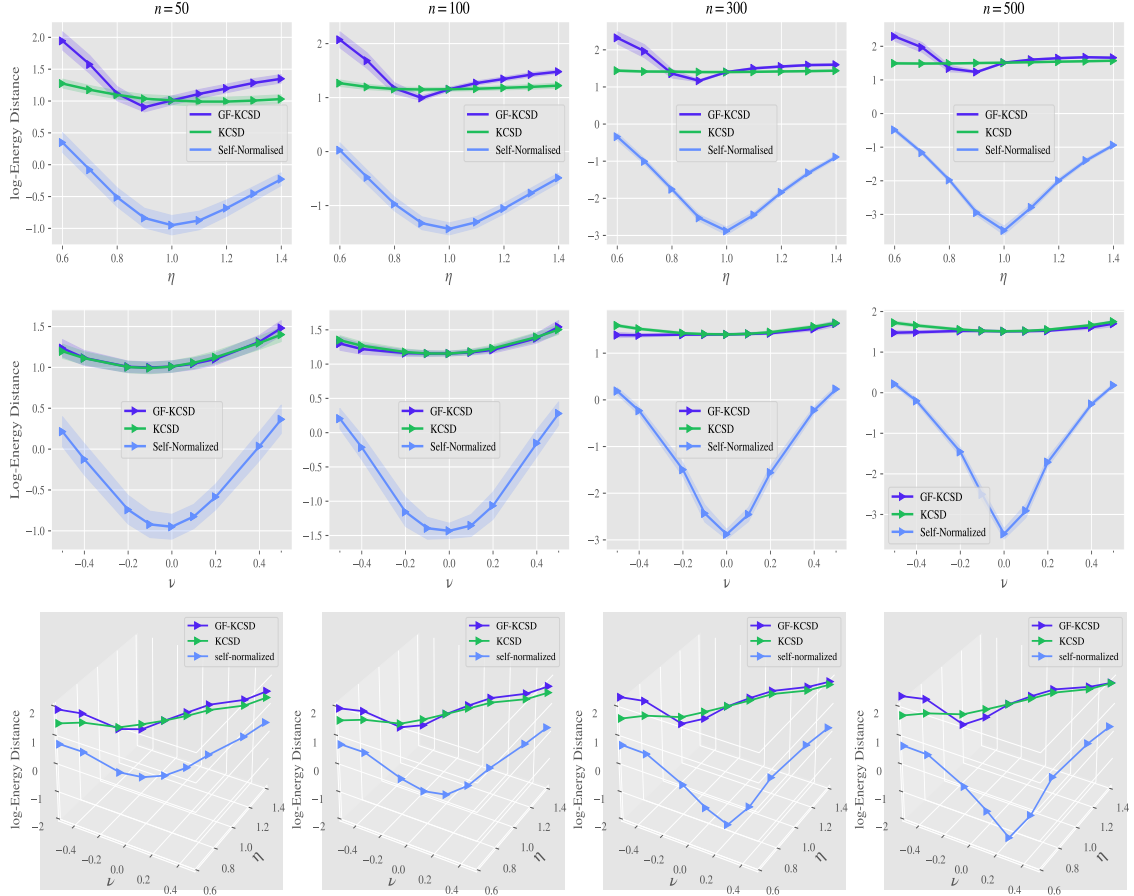


Figure 2.3: Evaluation of gradient-free conditional importance sampling in different sample size and scenarios: First Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x, \eta)$, Second Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x + \nu, 1)$, Third Row: under the scenario $\rho(y|x) \sim \mathcal{N}(y; x + \nu, \eta)$.

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.

Bibliography

- Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725. 13
- Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32. 13
- Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media. 17
- Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408. 20

- Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61. 20, 25
- Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (2019). Stein point markov chain monte carlo. In *International Conference on Machine Learning*, pages 1011–1021. PMLR. 13
- Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. (2018). Stein points. In *International Conference on Machine Learning*, pages 844–853. PMLR. 13
- Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR. 12, 13, 15, 17, 18
- Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28. 17
- Filippone, M. and Girolami, M. (2014). Pseudo-marginal bayesian inference for gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226. 13
- Fisher, M., Nolan, T., Graham, M., Prangle, D., and Oates, C. (2021). Measure transport with kernel stein discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR. 13

- Fisher, M. A., Oates, C., et al. (2022). Gradient-free kernel stein discrepancy. *arXiv preprint arXiv:2207.02636*. 13, 14, 15, 19, 22, 23, 34, 35
- Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928. 29
- Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein’s method. *Advances in Neural Information Processing Systems*, 28. 12, 17
- Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR. 17, 26, 28
- Han, J. and Liu, Q. (2018). Stein variational gradient descent without gradient. In *International Conference on Machine Learning*, pages 1900–1908. PMLR. 13
- Hodgkinson, L., Salomone, R., and Roosta, F. (2020). The reproducing stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*. 13, 34
- Huggins, J. and Mackey, L. (2018). Random feature stein discrepancies. *Advances in Neural Information Processing Systems*, 31. 26
- Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4). 16
- Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR. 13, 14, 15, 19

- Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press. 12
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91. 12
- Liu, Q. and Lee, J. (2017). Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR. 13, 34
- Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR. 12, 13, 15, 17
- Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29. 13
- Matsubara, T., Knoblauch, J., Briol, F.-X., Oates, C., et al. (2021). Robust generalised bayesian inference for intractable likelihoods. *arXiv preprint arXiv:2104.07359*. 13
- Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718. 12, 16, 17
- O’donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068. 36

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32. 29
- Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S., Mackey, L., and Oates, C. (2022). Optimal thinning of mcmc output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84. 13
- Salakhutdinov, R. (2015). Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385. 12
- Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons. 18
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281. 12
- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7). 17, 20, 25, 29
- 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*, volume 6, pages 583–603. University of California Press. 16, 23
- Stein, C., Diaconis, P., Holmes, S., and Reinert, G. (2004). Use of exchangeable

pairs in the analysis of simulations. *Lecture Notes-Monograph Series*, pages 1–26.

23

Szabó, Z. and Sriperumbudur, B. K. (2017). Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.*, 18:233–1. 20

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR. 13

Chapter 3

Navigating Interpretability and Alpha Control in GF-KCSD Testing with Measurement Error: A Kernel Approach

The Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD), presented in prior work (Afzali and Muthukumarana, 2023), represents a significant advancement in goodness-of-fit testing for conditional distributions. This method offers a robust alternative to previous gradient-based techniques, specially when the gradient calculation is intractable or computationally expensive. In this study, we explore previously unexamined aspects of GF-KCSD, with a particular focus on critical values and test power—essential components for effective hypothesis testing. We also present novel investigation on the impact of measurement errors on the performance of GF-KCSD in comparison to established benchmarks, enhancing our understanding of its resilience to these errors. Through controlled experiments using synthetic data, we demonstrate GF-KCSD’s superior ability to control Type-I error rates and

maintain high statistical power, even in the presence of measurement inaccuracies. Our empirical evaluation extends to real-world datasets, including brain MRI data. The findings confirm that GF-KCSD performs comparably to KCSD in hypothesis testing effectiveness while requiring significantly less computational time. This demonstrates GF-KCSD’s capability as an efficient tool for analyzing complex data, enhancing its value for scenarios that demand rapid and robust statistical analysis.

3.1 Introduction

In the dynamic field of statistical and machine learning models, the rapid development of complex algorithms has outpaced the development of robust model validation and diagnostic methods. It is essential to bridge this gap to understand the strengths and limitations of complex models, enhance their interpretability, and identify areas for improvement. By assessing how well a model fits observed data, these techniques enable model validation, interpretability, and finding ways to make it better. Acknowledging that "all models are wrong, but some are useful" (Box, 1976; Box and Draper, 1987), the challenge lies not in validating the truth of a model but rather in quantifying how much it deviates from the observed data (Gelman and Shalizi, 2013). This continuous cycle of model criticism, revision, and reevaluation is a part of the data analysis process as illustrated by Blei’s Box’s loop (Blei, 2014).

A cornerstone in statistical model criticism is the goodness-of-fit test, with traditional tests such as the χ^2 test (Pearson, 1900), Kolmogorov–Smirnov test (Smirnov, 1939), and Anderson–Darling test (Anderson and Darling, 1952) being essential

tools for practitioners to ensure their models match the real-world data. However, these traditional tests often demand a fully specified model distribution, a limitation in modern applications where distributions are frequently known only up to an intractable normalization constant. This challenge arises in scenarios such as statistical models for network data, large-scale graphical models, and deep generative models.

To address this issue, subsequent studies have developed a novel goodness-of-fit assessment strategy. This strategy applies Stein’s technique (Stein, 1972), which works directly with model distributions that are non-normalized. The foundational groundwork laid by Gretton et al. (2012) in integral probability metrics showcased the efficacy of utilizing kernel embeddings to compare distributions. Further developments, building on this by Gorham and Mackey (2015) proposed a discrepancy measure based on Stein operators and functions in Sobolev space. In order to improve goodness-of-fit testing methods for complex and non-normalized model distributions researchers Chwialkowski et al. (2016); Liu et al. (2016) used the unit ball of a reproducing kernel Hilbert space (RKHS) as the designated test function class. After that, Jitkrittum et al. (2020) made further strides by developing the kernel conditional Stein discrepancy (KCSD) test tailored for conditional goodness-of-fit assessment. Notably, it is imperative to acknowledge that these methodologies necessitate gradient computations, a computational aspect that may introduce instability, as highlighted by Fisher et al. (2022). They presented the gradient-free kernel Stein discrepancy (GF-KSD), a novel method for assessing goodness-of-fit that does not rely on gradients. It uses kernel-based algorithms to calculate the difference between model distributions and observed data, making it appropriate

for complicated models and high-dimensional data when gradients are either unavailable or computationally costly. However, one shortcoming of the GF-KSD technique is that it focuses on analyzing marginal distributions rather than conditional distributions. Building on this foundation, Afzali and Muthukumarana (2023) further expanded the application of gradient-free methodologies to the realm of conditional distribution testing through the development of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). This advancement integrates several cutting-edge concepts, including kernel embedding of distributions, Stein operators, and the application of gradient-free discrepancies within a conditional framework. Such integration not only addresses the challenges of gradient instability but also enhances the adaptability and reliability of goodness-of-fit assessment in the context of complex modern data sets. The rest of this study rigorously investigates empirical performance of GF-KCSD.

The effect of measurement error on non-parametric test statistics critically undermines the accuracy and credibility of statistical interpretations. These methods, which avoid the assumption of a specific data distribution, are notably susceptible to errors in measurement. Such inaccuracies can cause improper data analysis, which is fundamental to analyses that rely on precise data representation. As a result, measurement error can substantially skew results, typically diminishing the perceived magnitude of effects and decreasing the ability to discern significant differences. These are important consequences that necessitate the implementation of robust approaches to mitigate these effects and ensure more reliable statistical analyses. This highlights the critical need to account for measurement error when designing and interpreting research employing non-parametric statistics (Carroll

et al., 2006; Delaigle and Van Keilegom, 2021; Fuller and Fuller, 1987).

In this work, we aim to (i) explore the critical values of the GF-KCSD test statistic, (ii) evaluate the statistical power of GF-KCSD against relevant benchmarks, and (iii) examine the effects of measurement error on the performance of the GF-KCSD and established benchmarks. Understanding the distribution and appropriate critical values will enable principled hypothesis testing using GF-KCSD. Comparisons with KCSD and other divergence tests will provide insight into the advantages of GF-KCSD. Furthermore, assessing the impact of measurement errors is critical for assuring the robustness of statistical inference, especially in actual applications with noisy data.

Paper Outline: This paper begins with Section 3.2, which lays the framework by explaining the key concepts and theoretical backgrounds required to understand the study’s aims. In Section 3.3, we go into the methodology used to evaluate the performance of the GF-KCSD approach, including a thorough discussion of the features of measurement errors and their expected impact on GF-KCSD effectiveness. Section 3.4 provides experimental examples that highlights the practicality and effectiveness of the GF-KCSD approach for assessing statistical models. This section is separated into two major subsections: one for synthetic datasets and one for real-world brain MRI database. Finally, Section 3.5 summarizes our findings and offers a discussion of the implications for future research initiatives.

3.2 Foundational Concepts and Theoretical Background

Having established the challenges of interpreting test statistics in complex models, we now turn our attention to an overview of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD). We will present the theoretical background and fundamental principles, laying the groundwork for our investigation of GF-KCSD’s critical values, type-I error and test power.

To grasp how GF-KCSD works, we first need to cover some background information. We begin with an exploration of Stein’s method and its extensions, discussing the Stein operator and Stein identity, which are crucial concepts that allow us to apply GF-KCSD to estimating conditional probability densities. Moreover, the role of Reproducing Kernel Hilbert Space (RKHS) in kernel methods will be explored to highlight its value in embedding distributions and enabling gradient-free discrepancy calculations. Our objective is to create a strong basis for the debate in the next sections of this paper.

3.2.1 The Stein Operator

The Stein operator, originally introduced by Stein (Stein, 1972), is a pivotal element in statistical analysis, particularly within the scope of the Kernel Stein Discrepancy (KSD) method and its extensions. This operator serves as a linear differential operator that maps functions from a certain function space (often a reproducing kernel Hilbert space) to the space of functions whose expectations under the target distribution are zero. Let $\mathcal{H}$ denote a Hilbert space of functions. The Stein operator

$\mathcal{T}_p$ is mathematically represented as $\mathcal{T}_p : \mathcal{H} \rightarrow \mathcal{H}_0$, where $\mathcal{H}_0$ is the subspace of functions in $\mathcal{H}$ such that $\mathbb{E}_p[\mathcal{T}_p \mathbf{g}(x)] = 0$. This operator transforms a function $\mathbf{g}(x) \in \mathcal{H}$ into

$$\mathcal{T}_p \mathbf{g}(x) = \sum_{i=1}^n \left(g_i(x) \cdot \frac{\partial}{\partial x_i} \log p(x) + \frac{\partial}{\partial x_i} g_i(x) \right),$$

where $\frac{\partial}{\partial x_i} \log p(x)$ is the gradient of the log-density of the target distribution, and $\frac{\partial}{\partial x_i} g_i(x)$ is the gradient of the test function $\mathbf{g}(x)$ with respect to x . The Stein operator thus combines the original function $\mathbf{g}(x)$ with information from the target distribution. This allows it to capture how $\mathbf{g}(x)$ changes with respect to both its original values and the density of the target distribution, resulting in $\mathcal{T}_p \mathbf{g}(x) \in \mathcal{H}_0$.

A fundamental property of the Stein operator is Stein’s identity, which states that the expectation of $\mathcal{T}_p \mathbf{g}(x)$ is zero under the target distribution. This identity is critical within the framework of kernel Stein discrepancy approaches, acting as a key property that supports their efficacy. It guarantees that the connection between the test function and the target distribution is correctly represented, giving a trustworthy basis for model evaluation:

$$\mathbb{E}_{x \sim p}[\mathcal{T}_p \mathbf{g}(x)] = 0. \tag{3.1}$$

Equation (3.1), not only supports Stein’s identity, but it also serves as the theoretical foundation for the use of the Stein discrepancy techniques in statistical analysis.

Stein’s identity is fundamentally linked to the framework of the Reproducing Kernel Hilbert Space (RKHS)—a Hilbert Space where functions are endowed with the capacity for inner product operations, ensuring closure. Within this context, we

consider the class of test functions to reside in an RKHS, specifically, $g_i \in \mathcal{H}$ for all i . Here, $\mathcal{H}$ represents the RKHS wherein each scalar-valued function, g_i , is an element, as discussed in seminal works (Chwialkowski et al., 2016; Liu et al., 2016).

In the well-defined framework of a Reproducing Kernel Hilbert Space (RKHS), defined over a set $\mathcal{X}$ with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, we delve into the analysis of function properties through RKHS's inherent operations:

1. **Inner Product:** In the context of a reproducing kernel Hilbert space (RKHS), the inner product between any two functions f and g within this space, denoted as $f, g \in \mathcal{H}$, is defined by the expected value of their product under a target probability distribution $p(x)$:

$$\langle f, g \rangle_{\mathcal{H}} = \mathbb{E}_p[f(x) \cdot g(x)]. \quad (3.2)$$

This definition integrates statistical expectations with the geometric principles of RKHS, thereby enhancing the way interactions between functions are analyzed within this mathematical space.

2. **Norm:** The norm of a function $f \in \mathcal{H}$, reflects its magnitude and behavior within the RKHS. It is calculated as the square root of the function's self-inner product:

$$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}} = \sqrt{\mathbb{E}_p[f(x)^2]}, \quad (3.3)$$

where $\mathbb{E}_p$ denotes the expectation with respect to the target distribution $p(x)$.

This norm serves as a critical measure, providing insights into the function's variance and its distributional characteristics across the space.

This exposition not only highlights the role of RKHS in the fusion of geometric and probabilistic principles in function analysis but also showcases its critical application in statistical inference and the theoretical underpinnings of Stein’s method.

The mathematical foundation provided by the Stein Operator and the reproducing kernel Hilbert space (RKHS) ensures a rigorous and systematic approach to the analysis of functions within the context of kernel Stein discrepancy methods. Given an observed dataset $(x_i)_{i=1}^n$ with an unknown distribution p_0 , and a proposed model with distribution p , we utilize a class of test functions $\mathbf{g} \in \mathcal{G}$ in RKHS to measure the Kernel Stein Discrepancy (KSD). Here, $\mathcal{G}$ represents the class of test functions within the RKHS. The KSD is defined as:

$$\begin{aligned} S_p(p_0, \mathcal{T}_p, \mathcal{G}) &= \sup_{\|\mathbf{g}\|_{\mathcal{G}} \leq 1} \left\| \mathbb{E}_{x \sim p_0}[\mathcal{T}_p \mathbf{g}(x)] - \overset{0}{\mathbb{E}_{x \sim p}[\mathcal{T}_p \mathbf{g}(x)]} \right\| \\ &= \sup_{\|\mathbf{g}\|_{\mathcal{G}} \leq 1} \|\mathbb{E}_{x \sim p_0}[\mathcal{T}_p \mathbf{g}(x)]\|. \end{aligned} \quad (3.4)$$

The KSD is zero if and only if the two distributions, p and p_0 , are identical in the sense that there is no statistical difference between them detectable by any test function in the chosen RKHS. This method allows us to systematically assess how well the proposed model p captures the true characteristics of the dataset $\{x_i\}_{i=1}^n$, which follows the distribution p_0 .

3.2.2 Conceptual Backgrounds of the GF-KCSD

The GF-KCSD framework emerges as an extension of the kernel Stein discrepancy paradigm, tailored to address the complexities of conditional distributions. It does not depend on gradients of the proposed distribution like some other KSD methods,

so it's not bogged down by typical computational limitations. GF-KCSD leverages the gradient-free conditional Stein operator, as introduced by Afzali and Muthukumarana (2023), which enables a thorough examination of conditional distributions that are both precise and computationally effective. This operator is defined as follows, capturing the essence of its gradient-free approach:

$$\zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}') = \omega(\mathbf{y}|\mathbf{x})\omega(\mathbf{y}'|\mathbf{x}') [\mathbf{g} \cdot \nabla_{\mathbf{y}} \log q(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{y}} \cdot \mathbf{g}], \quad (3.5)$$

where $\omega(\mathbf{y}|\mathbf{x}) = \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})}$ is an importance weight to correct the bias introduced by the surrogate, q denotes the surrogate distribution, p signifies the target distribution, $\mathbf{g}$ is the test function within the RKHS. This operator adeptly integrates the derivative of the log-density of q with respect to y , bypassing direct derivatives of p , thus embodying the gradient-free characteristic. Here, the function $\mathbf{g} \in \mathcal{G}^d$, where $\mathcal{G}_l^{d_y} = \prod_{j=1}^{d_y} \mathcal{G}_l$, and $\mathcal{G}_l$ represents the RKHS associated with a positive definite kernel $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y}$. This operator assumes that q is absolutely continuous with respect to p , signifying that events with probability zero under p also have probability zero under q . This scenario suggests that q provides more detailed information compared to p , with the density ratio $\frac{q}{p}$ being bounded and ∇q being Lipschitz, meaning it varies smoothly across the entire space and remains within a certain range.

Consider two random vectors, $\mathbf{x}$ and $\mathbf{y}$, in the space $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{D_x} \times \mathbb{R}^{D_y}$. We have a joint sample of independent and identically distributed pairs denoted as $(x_i, y_i)_{i=1}^n \sim p_{0_{xy}}$. The joint density function is defined by Bayes rule as $p_{0_{xy}}(\mathbf{x}, \mathbf{y}) = p_0(\mathbf{y}|\mathbf{x})p_{0_x}(\mathbf{x})$ with $p_0(\mathbf{y}|\mathbf{x})$ being unknown and only accessible through the joint sample. The goal is to test the proposed conditional density function $p(\mathbf{y}|\mathbf{x})$ as the true conditional distribution $p_0(\mathbf{y}|\mathbf{x})$. This assessment is undertaken through

a hypothesis test aimed at gauging the goodness-of-fit of the proposed model with respect to an unknown marginal distribution, denoted as p_{0_x} . The hypothesis test is formulated as follows:

$$\mathbf{H}_0 : p(\mathbf{y}|\mathbf{x}) \stackrel{p_{0_x}}{=} p_0(\mathbf{y}|\mathbf{x}), \quad \mathbf{H}_1 : p(\mathbf{y}|\mathbf{x}) \stackrel{p_{0_x}}{\neq} p_0(\mathbf{y}|\mathbf{x}), \quad (3.6)$$

where $p \stackrel{p_{0_x}}{=} p_0$ is specified that $p(\mathbf{y}|\mathbf{x}) = p_0(\mathbf{y}|\mathbf{x})$, for p_{0_x} -almost all supporting points $\mathbf{x}$ and for every $\mathbf{y} \in \mathcal{Y}$. Incorporating the conditional component involves multiplying the Stein operator by the kernel function $K(\mathbf{x}, \mathbf{x}')$. As a result, the Kernelized Conditional Stein Operator, denoted as $G_q(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}, \mathbf{x}') \times \zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}')$, exhibits injectiveness owing to the C_0 -universality of $K(\mathbf{x}, \mathbf{x}')$. This injectiveness serves as a direct confirmation of the Stein identity property, showcasing the versatility and reliability of the Kernelized Stein Operator within the context of C_0 -universal kernels. The key to the GF-KCSD is the G_q such that the expectation under the distribution p vanish, i.e., for any function $g \in \mathcal{G}_l$:

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{xy}} G_q(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{xy}} [K(\mathbf{x}, \mathbf{x}') \times \zeta_{q(\mathbf{y}|\mathbf{x})}(\mathbf{y}, \mathbf{y}')] = 0 \quad (3.7)$$

The Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) between conditional probability distributions P_0 and P is a discrepancy measure defined as the supremum (maximum value) of the difference between the expected value of the Kernelized Stein Operator under the distributions P_0 and P , with respect to the norm in the RKHS $\mathcal{G}_l$:

$$\mathcal{S}_{p,q}(p_0) = \sup_{\|g\|_{\mathcal{G}_l} \leq 1} \left\| \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{0xy}} [G_q(\mathbf{x}, \mathbf{y})] - \overset{0}{\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{xy}}} [G_q(\mathbf{x}, \mathbf{y})] \right\|_{\mathcal{G}_l}. \quad (3.8)$$

Having a joint sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n \stackrel{i.i.d.}{\sim} p_{0_{xy}}$, the estimator for the GF-KCSD is shown as:

$$\widehat{\mathcal{S}}_{p,q}^2(p_0) = \frac{1}{n(n-1)} \sum_{i \neq j} G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)). \quad (3.9)$$

The explicit form of the GF-KCSD estimator for a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ is presented as:

$$\widehat{\mathcal{S}}_{p,q}(p_0) = E_{xy} E_{x'y'} K(\mathbf{x}, \mathbf{x}') \zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y')). \quad (3.10)$$

Furthermore, the explicit form of $\zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y'))$ is as follows:

$$\begin{aligned} \zeta_{q(\mathbf{y}|\mathbf{x})}((x, y), (x', y')) &= \omega(\mathbf{y}|\mathbf{x}) \omega(\mathbf{y}'|\mathbf{x}') \{ l(\mathbf{y}, \mathbf{y}') \mathbf{s}_q^T(\mathbf{y}|\mathbf{x}) \mathbf{s}_q(\mathbf{y}'|\mathbf{x}') \\ &\quad + \sum_{i=1}^{D_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(\mathbf{y}, \mathbf{y}') + \mathbf{s}_q^T(\mathbf{y}|\mathbf{x}) \nabla_{\mathbf{y}'} l(\mathbf{y}, \mathbf{y}') \\ &\quad + \mathbf{s}_q^T(\mathbf{y}'|\mathbf{x}') \nabla_{\mathbf{y}} l(\mathbf{y}, \mathbf{y}') \} , \end{aligned} \quad (3.11)$$

where $\mathbf{s}_q(\mathbf{y}|\mathbf{x})$ is the gradient of the log-density of $q(\mathbf{y}|\mathbf{x})$. We provided a concise overview of this concept in this section. For more in-depth information and proofs about the GF-KCSD, please refer to Afzali and Muthukumarana (2023, Section 2.2).

3.3 Methodology

In this section, we describe the methodology used to investigate the GF-KCSD's sampling distribution, critical values, and performance within our experimental framework. We utilize a modification of the bootstrap method (Efron, 1992) to evaluate the sample distribution and identify critical values for hypothesis testing.,

which provides insights into the robustness of the GF-KCSD estimator. Simultaneously, we address the complications presented by measurement error by incorporating it into our analysis, thus accounting for the inherent uncertainties in empirical data and enhancing the accuracy and reliability of our findings. This methodological framework lays the groundwork for our subsequent experiment, where we explore the performance of GF-KCSD under various conditions.

3.3.1 Bootstrap Resampling

Our investigation into the GF-KCSD for goodness-of-fit testing embarks on a critical comparison between a specific non-normalized conditional model distribution $p(y|x)$ and a collection of independently and identically distributed (i.i.d.) joint samples $(x_i, y_i)_{i=1}^n$, drawn from an unspecified data-generating distribution $p_0(y|x)$. Given the complexity and non-parametric nature of the GF-KCSD, particularly the absence of a closed-form distribution under H_0 , we deploy the bootstrap method. This essential resampling technique estimates an estimator’s properties by sampling with replacement from the original dataset making it indispensable strategy for approximating the test threshold when traditional computational approaches are impractical (Arcones and Gine, 1992; Huskova and Janssen, 1993). This aligns our methodology with the forefront of statistical innovation as exemplified in Kernelized Stein Discrepancy (KSD) tests (Liu et al., 2016; Yang et al., 2018; Jitkrittum et al., 2020).

By setting the rejection threshold to $\eta_{1-\alpha}$, corresponding to the $(1 - \alpha)$ -quantile of the asymptotic null distribution, our test rigorously rejects the null hypothesis H_0 if the computed statistic $n\widehat{\mathcal{S}}_{p,q}(p_0)$ surpasses this threshold. This technique

Algorithm 1 Estimation of Test Threshold for Hypothesis Testing

- 1: **Input:** Dataset $(x_i, y_i)_{i=1}^n \sim p_{0_{xy}}$, Number of bootstrap samples $\tilde{n}$, Significance level α
 - 2: **Objective:** To test $H_0 : p \stackrel{p_{0_x}}{=} p_0$ against $H_1 : p \stackrel{p_{0_x}}{\neq} p_0$
 - 3: **Compute test statistic** $\hat{\mathcal{S}}_{p,q}^2(p_0) = \frac{1}{n(n-1)} \sum_{i \neq j} G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j))$
 - 4: **for** $b = 1, \dots, \tilde{n}$ **do**
 - 5: Sample with replacement $\{w_1, \dots, w_n\} \sim \text{Multinomial}(n; \frac{1}{n}, \dots, \frac{1}{n})$
 - 6: Adjust weights: $\tilde{w}_i = \frac{w_i - 1}{n}$
 - 7: Compute bootstrap statistic:
 $\hat{\mathcal{S}}_{(p,q)b}^*(p_0) = \sum_{i=1}^n \sum_{j \neq i} \tilde{w}_i \tilde{w}_j G_q((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j))$
 - 8: **end for**
 - 9: Determine the $(1 - \alpha)$ -quantile $\eta_{1-\alpha}$ of the bootstrap statistics $\{\hat{\mathcal{S}}_{(p,q)b}^*(p_0)\}_{b=1}^{\tilde{n}}$
 - 10: **Output:** Test threshold $\eta_{1-\alpha}$ as the estimate for rejecting H_0
-

not only tackles computational issues, but it also represents the evolving landscape of non-parametric hypothesis testing. The Algorithm 1 summarizes our approach and outlines the procedures required for the estimation of the test threshold. This algorithm does more than automate computations; it represents a methodological evolution, ensuring that each step—from sampling to the final quantile determination—enhances the accuracy and interpretability of our goodness-of-fit testing.

3.3.2 Measurement Errors

In many applied fields of study, our objective is to model the relationship between a response variable and one or more explanatory variables. However, in some cases, the covariates of interest are neither observable nor measurable. In such instances, we employ indirect measures of the true covariates, known as proxy variables. These proxy variables inevitably measure the actual covariates with some degree of error, leading to the designation of such models as measurement error models. Essentially, measurement error models are useful when the explanatory variables that

genuinely govern the response behavior cannot be directly observed. The classical measurement error model is articulated as:

$$w_i = x_i + u_i, \tag{3.12}$$

where w_i is the observed proxy variable, x_i is the true unobserved covariate, and u_i is the random measurement error, assumed to be independent of x_i and characterized by a mean of zero ($\mathbb{E}[u_i] = 0$) and a variance of σ_u^2 ($\text{Var}[u_i] = \sigma_u^2$).

Two key assumptions underlie measurement error models: (i) unbiasedness, ensuring $\mathbb{E}[w_i|x_i] = x_i$, such that observed values reflect true values on average; and (ii) independence, requiring $\text{Cov}[x_i, u_i] = 0$, preventing systematic biases related to the magnitude of x_i . Violation of these assumptions can lead to biased and inefficient estimates.

This section delves deeply into classical measurement error models, with an emphasis on their consequences in statistical frameworks such as the Linear Gaussian Model and Heteroscedastic Gaussian Model, elucidating their theoretical underpinnings, estimation approaches, and implications when faced with measurement errors.

Linear Gaussian Model with Classical Measurement Error

In the analysis of simple linear regression models, it is crucial to account for measurement errors, especially when the true covariates cannot be directly observed. This section explores a simple linear regression model incorporating measurement error, highlighting its impact on the observed relationship between the response variable and covariates.

Consider the true relationship for the i^{th} observation in a simple linear regression model expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

Here, β_0 and β_1 are the regression coefficients, and ϵ_i represents the random error term, which is assumed to be independent of x_i and normally distributed with mean zero and variance σ_ϵ^2 , i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. However, instead of observing the true covariate x_i , we observe a proxy variable w_i , which is related to x_i through the classical measurement error model $w_i = x_i + u_i$, where u_i is the measurement error, assumed to be independent of x_i and ϵ_i , and normally distributed with mean zero and variance σ_u^2 , i.e., $u_i \sim \mathcal{N}(0, \sigma_u^2)$. Substituting $x_i = w_i - u_i$ into the linear Gaussian model, we obtain:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(w_i - u_i) + \epsilon_i \\ &= \beta_0 + \beta_1 w_i - \beta_1 u_i + \epsilon_i. \end{aligned}$$

Defining a new combined error term $\eta_i = \epsilon_i - \beta_1 u_i$, the observed model can be written as:

$$y_i = \beta_0 + \beta_1 w_i + \eta_i.$$

The distribution of the combined error term η_i is also normal, as it is a linear combination of two independent normal variables:

$$\eta_i \sim \mathcal{N}(0, \sigma_\epsilon^2 + \beta_1^2 \sigma_u^2). \quad (3.13)$$

This formulation highlights the impact of measurement error in the explanatory variable x_i on the linear Gaussian model, causing the inflation of error variance and making standard methods like ordinary least squares (OLS) biased and inconsistent. To accurately estimate the regression coefficients β_0 and β_1 , as well as the error

variances σ_ϵ^2 and σ_u^2 , alternative techniques such as errors-in-variables models or instrumental variables are required. These methods are crucial for adjusting the biases introduced by measurement error and ensuring reliable estimation of the true relationship between variables.

Heteroscedastic Gaussian Model with Classical Measurement Errors:

Heteroscedastic models play a fundamental role in statistical analyses, particularly when the assumption of homoscedasticity, or constant variance of error terms ϵ_i across observations, is violated. These models adeptly address the variability in uncertainty associated with predictions, a common occurrence in real-world data. The heteroscedastic Gaussian model can be simply expressed as:

$$y_i = f(x_i, \beta) + \epsilon_i, \quad (3.14)$$

where for the i^{th} observation, y_i denotes the response variable, x_i represents the set of predictor variables, $f(x_i, \beta)$ is a function defining the relationship between predictors and response reliant on parameters β , and ϵ_i is the random error term. Unlike in homoscedastic models, ϵ_i follows a normal distribution with a mean of zero and a variance of σ_i^2 , where σ_i^2 is a function of predictor variables or other known quantities, expressed as $\sigma_i^2 = g(x_i; \theta)$, making the variance heteroscedastic.

Incorporating classical measurement error into a heteroscedastic Gaussian model requires adjusting both the observed relationship and the variance to account for the error introduced by using measured predictors w_i instead of the true covariates x_i . Assuming that the measurement error u_i is normally distributed with mean zero and variance σ_u^2 , substituting $x_i = w_i - u_i$ into the heteroscedastic Gaussian model,

we obtain:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(w_i - u_i) + \epsilon_i \\ &= \beta_0 + \beta_1 w_i - \beta_1 u_i + \epsilon_i, \end{aligned}$$

Defining a new combined error term $\eta_i = \epsilon_i - \beta_1 u_i$, the observed model can be written as:

$$y_i = \beta_0 + \beta_1 w_i + \eta_i.$$

The distribution of the combined error term η_i is more complex than in the linear Gaussian case, as its variance now depends on both the true covariate x_i and the measurement error variance σ_u^2 :

$$\text{Var}[\eta_i] = \sigma^2(x_i; \theta) + \beta_1^2 \sigma_u^2.$$

This formulation highlights the additional challenges posed by measurement error in the heteroscedastic Gaussian model. Not only does the measurement error lead to an inflation of the error variance, but it also complicates the estimation of the variance function $\sigma^2(x_i; \theta)$ since the true covariate x_i is unobserved.

To obtain consistent estimates of the regression coefficients β_0 and β_1 , the parameter vector θ governing the variance function, and the measurement error variance σ_u^2 , specialized estimation techniques are required. These techniques must account for both the heteroscedasticity in the error term and the presence of measurement error in the explanatory variable.

Potential approaches may include extensions of errors-in-variables models, generalized method of moments (GMM) estimators, or other methods specifically designed to handle heteroscedasticity and measurement error simultaneously. The choice of

estimation method will depend on the specific assumptions and characteristics of the heteroscedastic Gaussian model under consideration.

3.4 Experiment design

This section presents an empirical study that investigates the interpretability of GF-KCSD test statistics in conditional models under a variety of settings. Our experiments are separated into two parts: analyses using synthetic data and evaluations using real-world data, each designed to evaluate different aspects of the GF-KCSD method.

3.4.1 Synthetic Data Experiments

In the first part, we focus on synthetic data experiments using the experimental setup provided by Jitkrittum et al. (2020), who proposed the Kernel Conditional Stein Discrepancy (KCSD) approach, serving as a benchmark for our analysis. Our work is unique in that it takes into account both symmetric and asymmetric distributions of measurement errors, acknowledging their considerable influence on statistical method performance—a factor often omitted in previous studies including the foundational KCSD paper. We evaluate the statistical power of GF-KCSD by comparing it to existing gradient-based techniques, focusing on how various distributions of measurement errors affect method performance. Our analysis has two main objectives: first, to assess GF-KCSD’s effectiveness relative to other methods; and second, to demonstrate how measurement errors influence the efficacy of GF-KCSD and these comparative techniques. This approach allows us to provide

a comprehensive evaluation of the performance across methods. We focus on the performance of the subsequent methods.

GF-KCSD: Introduced by Afzali and Muthukumarana (2023), GF-KCSD assesses the fit of conditional distributions without computing gradients, using Gaussian kernels for both x and y . The kernel function is $k(w, w') = \exp\left(-\frac{|w-w'|^2}{2\sigma_w^2}\right)$, with bandwidth $\sigma_w = \text{median}\{|w_i - w_j|\}_{i,j=1}^n$ which is determined by the median heuristic.

KCSD: The Kernel Conditional Stein Discrepancy (KCSD) established by Jitkrittum et al. (2020), evaluates conditional density models using a kernel-based approach, facilitating a reliable assessment of model fit.

MMD: This method, adapting the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) for conditional goodness-of-fit testing, which was extended by Jitkrittum et al. (2020) to evaluate conditional distributions. It partitions the dataset into two subsets, generating new samples from the model $p(\cdot|x^{(2)})$.

Zheng: An enhancement of Zheng’s method (2012) by Jitkrittum et al. (2020) utilizes Gaussian kernels for smoothing, with bandwidth parameter heuristic $h_j = \hat{s}_j n^{-1/(4+d_x)}$.

We explore the performance of these methods on three conditional modeling problems: Linear Gaussian Model (LGM), Heteroscedastic Gaussian Model (HGM), and Quadratic Gaussian Model (QGM). In all cases, we assume the conditional densities $p(y|x)$ and $p_0(y|x)$ are non-normalized. The LGM serves as a baseline, accurately reflecting the true data-generating process, while the HGM and QGM introduce heteroscedastic behavior and a subtle deviation from linearity, respectively, challenging

the goodness-of-fit tests. Details of each model are provided below.

Quadratic Gaussian Model (QGM): The Quadratic Gaussian Model (QGM) is essential for assessing the performance of goodness-of-fit testing methods, especially when the suggested conditional model’s assumption do not align perfectly with the underlying true distribution. We denote data points by $(x, y) \in \mathbb{R} \times \mathbb{R}$. The proposed model hypothesizes a linear relationship, $p(y|x) \sim \mathcal{N}(x + 1, 1)$, whereas the true distribution, $p_0(y|x)$, has a quadratic relationship, $\mathcal{N}(0.1x^2 + x + 1, 1)$, and a marginal distribution $p_{0_x}(x) \sim \text{Uniform}(-2, 2)$.

This setting demonstrates a clear comparison between the basic linear assumption of $p(y|x)$ and the quadratic dynamic of $p_0(y|x)$, highlighting a scenario where the simplicity of the model’s assumption might not fully encapsulate the complexity of the data’s inherent structure. The inclusion of a slight quadratic factor, weighted at 0.1, introduces a nuanced deviation challenging for goodness-of-fit tests to detect. This reflects real-world situations where theoretical models might oversimplify the actual complexity of the data they aim to represent. This scenario suggests that the alternative hypothesis H_1 is valid, indicating a noticeable discrepancy between the modeled and true distributions. This situation emphasizes the importance of scrutinizing model fit, especially in instances where conventional assumptions may inadequately capture the complexity of the underlying distributions.

Heteroscedastic Gaussian Model (HGM): The Heteroscedastic Gaussian Model (HGM) involves data points $(\mathbf{x}, y) \in \mathbb{R}^3 \times \mathbb{R}$, where the non-normalized proposed conditional model is specified as $p(y|\mathbf{x}) \sim \mathcal{N}\left(\sum_{i=1}^3 x_i, 1 + 10e^{-|\mathbf{x}-\mathbf{c}|^2}\right)$. This model exhibits heteroscedastic behavior, with the conditional variance being a function of

$\mathbf{x}$, centered at the point $\mathbf{c} = \frac{2}{3}\mathbf{1}$, where $\mathbf{1}$ is a vector of ones.

Conversely, the true target conditional distribution, $p_0(y|\mathbf{x}) \sim \mathcal{N}(\sum_{i=1}^3 x_i, 1)$, is a homoscedastic Gaussian model with a constant unit variance. This sets a distinct contrast between the proposed model’s heteroscedastic nature and the homoscedastic true distribution, highlighting the challenge for goodness-of-fit tests to discern the heteroscedastic anomaly localized within specific covariate regions.

The marginal distribution for $\mathbf{x}$ is assumed to be a multivariate Gaussian distribution with a mean vector of zeros and an identity covariance matrix, denoted as $p_{0_x}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This scenario challenges the capacity of goodness-of-fit tests to detect subtle deviations from a constant variance assumption, thereby illuminating the nuanced dynamics of real-world data modeling where such uniformity in variance may not always be present.

Linear Gaussian Model (LGM): In this study, we evaluate the Linear Gaussian Model (LGM), where the data points are denoted as $(\mathbf{x}, y) \in \mathbb{R}^5 \times \mathbb{R}$, with $\mathbf{x}$ being a 5-dimensional vector of covariates, and y a scalar response variable. The conditional distribution $p(y|\mathbf{x})$ is assumed to follow a normal distribution with a mean $\sum_{i=1}^5 ix_i$ and a constant variance of 1, i.e., $p(y|\mathbf{x}) = \mathcal{N}(\sum_{i=1}^5 ix_i, 1)$. The predictor variables $\mathbf{x}$ are drawn from a multivariate normal distribution with mean zero and an identity covariance matrix, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This model specification ensures that the true conditional distribution $p_0(y|\mathbf{x})$ is perfectly aligned with our proposed model $p(y|\mathbf{x})$, satisfying the null hypothesis $H_0 : p(y|\mathbf{x}) = p_0(y|\mathbf{x})$. This perfect alignment is critical in our evaluation of the Type I error rates of various goodness-

of-fit tests. Through this alignment, the model serves as an ideal framework to assess whether statistical tests appropriately maintain their nominal error rates under conditions of correct model specification.

Models with Measurement Errors:

In the context of these three scenarios, we add complexity by including measurement errors in each case. Each data observation w_i is modified by an error term u_i . To realistically reflect the variety of errors found in real-world data, we use two different distributions for these errors. This strategy allows us to more accurately portray the irregularities that are common in real-world data. Here’s how we implement this:

- **Symmetric Errors:** For errors following a normal distribution, we define $w_i = x_i + u_i$, with $u_i \sim \mathcal{N}(0, \sigma_u^2)$ where $\sigma_u^2 = [0.5^2, 1^2, 2^2]$. This setup enables an exploration of the models’ robustness against random measurement noise at different intensity levels.
- **Asymmetric Errors:** To examine the impact of errors that skew in one direction, showing possible systematic biases, we model $w_i = x_i + u_i$, where $u_i \sim \chi^2(5)$, follows a Chi-square distribution to simulate asymmetric measurement errors.

This structured approach allows us to thoroughly assess how different types of measurement errors affect the accuracy and reliability of our statistical models.

Central to our evaluation is the comparison of GF-KCSD’s test power against several established conditional distribution tests: KCSD, FSCD, MMD, and Zheng, across

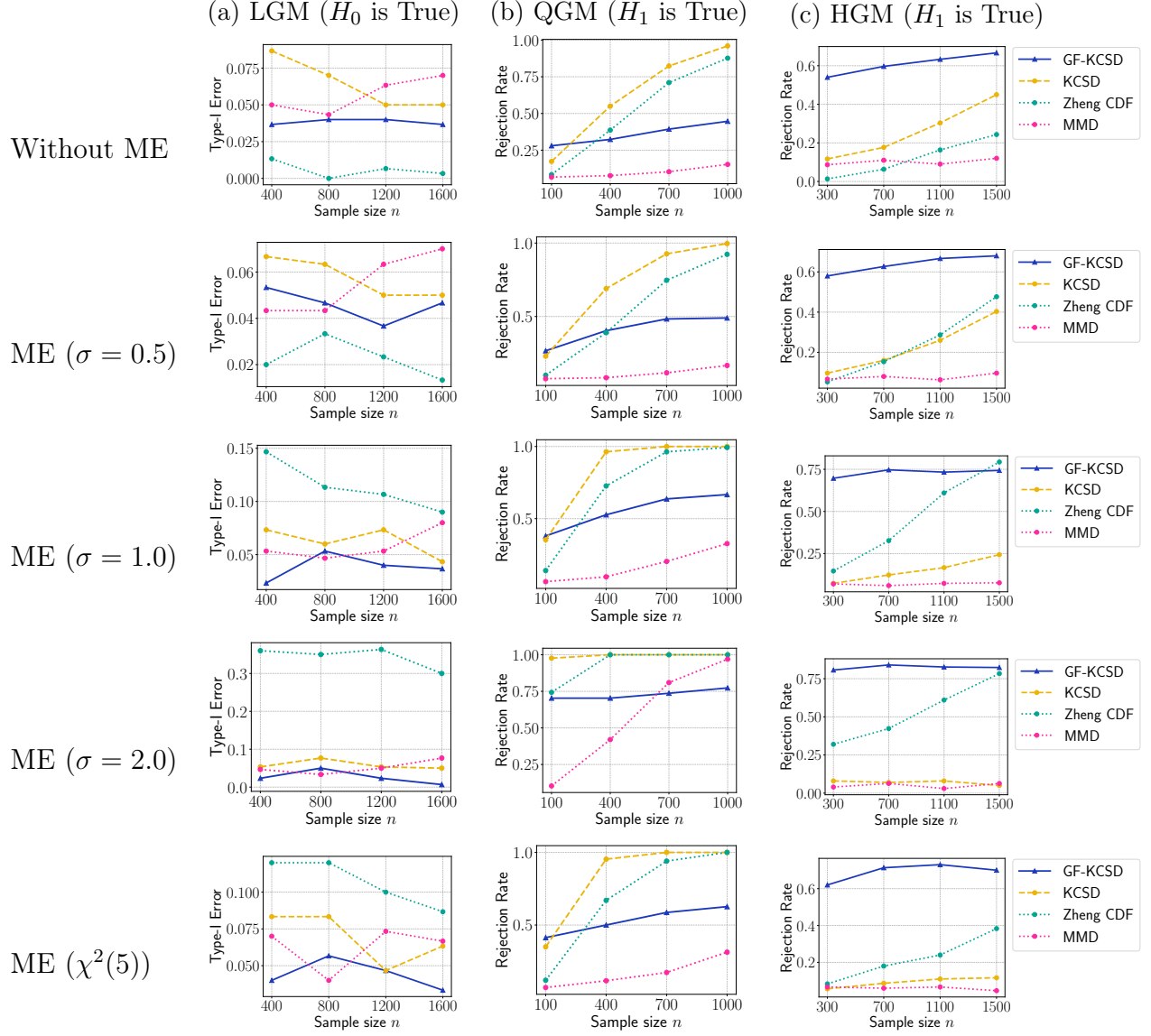


Figure 3.1: This figure presents a comparative analysis of Type-I error rates and rejection rates across four distinct tests, all conducted at a significance level of $\alpha = 0.05$. The first row depicts scenarios devoid of measurement errors, while subsequent rows introduce measurement errors. From the second to the fourth row, these errors follow a normal distribution with mean 0 and standard deviations of 0.5, 1.0, and 2.0, respectively. The final row involves measurement errors modeled by a $\chi^2(5)$. **(a):** In these contexts, when H_0 is true, the GF-KCSD method demonstrates robustness in maintaining control over Type-I errors, effectively managing fluctuations attributable to sampling noise. **(b):** Under the alternative hypothesis (H_1) the efficacy of the GF-KCSD in discerning global differences is suboptimal. **(c):** Independent of the presence or absence of measurement errors, GF-KCSD consistently shows a highly effective in identifying local differences.

300 trials with a bootstrap sample size of 500 at a significance level of $\alpha = 0.05$, utilizing Gaussian kernels. To ensure a fair comparison across similar situations in the experiments and all models, the hyperparameters remain consistent throughout all simulations. This consistency allowed for reliable and comparable results across different simulations and models. It is crucial for accurately assessing the performance of the GF-KCSD method relative to other established techniques. This analysis focuses on the method’s efficacy in hypothesis testing, specifically in distinguishing the modeled distribution $p(y|x)$ from the true distribution $p_0(y|x)$, with rejection rates serving as a barometer for both Type I error control and the test’s power. In this study, we utilized an empirical Gaussian approximation as a surrogate model for the target model, fitting Gaussian parameters (mean and variance) directly from the data.

Figure 3.1 illustrates the Type-I error rates and the power of four established tests across three distinct scenarios. The leftmost column of the figure shows that under the true null hypothesis (H_0) scenario, as the sample size increases, the false rejection rates for all tests remain within the specified significance level ($\alpha = 0.05$), accounting for predetermined sampling noise. Notably, the GF-KCSD test maintains error rates below this threshold even in small sample sizes, distinguishing its performance from the other tests when sample size is limited. Additionally, by incorporating measurement errors into the LGM setup, the performance of GF-KCSD remained stable, maintaining control at or below 0.05. In contrast, the other methods exhibited fluctuations and nuances in their performance.

In the scenario involving QGM, presented in the middle column of Figure 3.1, an increase in test power with larger sample sizes is observed for all tests, despite the

minimal weight assigned to the quadratic component in p_0 . Evidently, the GF-KCSD test demonstrates superior performance compared to the MMD test.

In the HGM context, depicted in the third column of Figure 3.1, where the disparities between p and p_0 are localized within the X domain, the GF-KCSD test is demonstrated to significantly outperform all other tests in identifying differences in X , particularly the KCSD test. Unlike the KCSD, which shows decreased performance with an increase in the standard deviation of measurement errors or under asymmetric error conditions, the GF-KCSD is adept at handling measurement errors and remains robust against them, maintaining its efficacy even when the measurement errors are significant or asymmetric. Meanwhile, the performance of both the KCSD and MMD tests decreases, especially under conditions of asymmetric measurement errors and when the standard deviation of measurement errors increases in symmetric scenarios.

3.4.2 Real-world Data Experiments (Brain Magnetic Resonance Imaging (MRI))

In this section, we conduct a comparative evaluation of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) and the Kernel Conditional Stein Discrepancy (KCSD) methodologies for hypothesis testing on a real-world dataset. The fundamental goal is to determine whether GF-KCSD performs similarly to the established KCSD test in rejecting or failing to reject the null hypothesis, which posits that the proposed model accurately captures the distribution of the observed data. Our analysis critically examines the consistency of outcomes between GF-KCSD and KCSD, assessing whether both methods lead to similar conclusions about the

model’s adequacy.

The dataset under investigation comprises preprocessed T1-weighted Magnetic Resonance Imaging (MRI) data from the study by Beheshti et al. (2021). Specifically, the analysis focuses on gray matter (GM) images, where voxel-level brain features are extracted, resulting in 3,747 voxels per volume. The dataset includes 876 Healthy Controls (HC), 70 individuals with Mild Cognitive Impairment (MCI), and 30 patients diagnosed with Alzheimer’s Disease (AD), totaling 976 samples. The goal is to predict the age of the samples using these voxel-level features as covariates in the proposed model.

To propose the $p(y|\mathbf{x})$ model for hypothesis test, we implemented a Mixture Density Network (MDN) as described by Bishop (2006), trained on the preprocessed brain MRI dataset. The MDN aims to predict the conditional probability distribution of participants’ ages, denoted y , based on their MRI-derived features, $\mathbf{x}$. Each dataset pair $(\mathbf{x}, y)$ belongs to $(\mathbb{R}^{3748} \times \mathbb{R})$, where 3748 features are extracted from MRI data, along with the category of the disease group. The architecture of the network comprises multiple layers, each dedicated to generating parameters for a Gaussian mixture model that characterizes the conditional distribution $p(y|\mathbf{x})$. The probability density function for a Gaussian Mixture Model, given the condition x , can be expressed as:

$$p(y|\mathbf{x}) = \sum_{i=1}^K \frac{\exp(a_i(\mathbf{x}))}{\sum_{j=1}^K \exp(a_j(\mathbf{x}))} \cdot \mathcal{N}(y|b_i(\mathbf{x}), \exp(c_i(\mathbf{x}))),$$

where K is the number of Gaussian components, determining the model’s complexity and flexibility. The parameters $a_i(\mathbf{x})$ are the raw network outputs for the mixing

coefficients, influencing the weight of each Gaussian component. The mixing coefficients $\pi_i(\mathbf{x}) = \frac{\exp(a_i(\mathbf{x}))}{\sum_{j=1}^K \exp(a_j(\mathbf{x}))}$ ensure they sum to 1, representing the probability of each Gaussian component. The parameters $b_i(\mathbf{x})$ are the means, and $c_i(\mathbf{x})$ are the log-variances, with the exponential function ensuring positive variances. The Gaussian distribution $\mathcal{N}(y|b_i(\mathbf{x}), \exp(c_i(\mathbf{x})))$ models the conditional density of y given $\mathbf{x}$. This approach allows the MDN to capture complex, multimodal relationships between $\mathbf{x}$ and y .

In this experiment, the network is configured with $K = 20$ Gaussian components and employs a ReLU-based framework to estimate the variances, means, and mixing coefficients. The network is trained to minimize the negative log-likelihood of the observed age relative to the model’s predictions, using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-4} and a batch size of 32. We used grid search to determine the optimal number of Gaussian components (K) and batch size, ensuring that the chosen parameters were well-suited to the data, thereby enhancing the test’s effectiveness. Figure 3.2 illustrates the comparison of the distribution of the real age and the sampled ages from the trained MDN model.

Next, we regard the trained MDN model as the proposed model $p(y|\mathbf{x})$. Our objective is to compare the performance of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD) and the Kernel Conditional Stein Discrepancy (KCSD) in determining if p represents the true distribution of the dataset. Given that the MDN model has been trained on this dataset, we expect both GF-KCSD and KCSD to fail to reject the null hypothesis, indicating high model performance.

We calculate test statistics using GF-KCSD and KCSD to measure the discrepancy

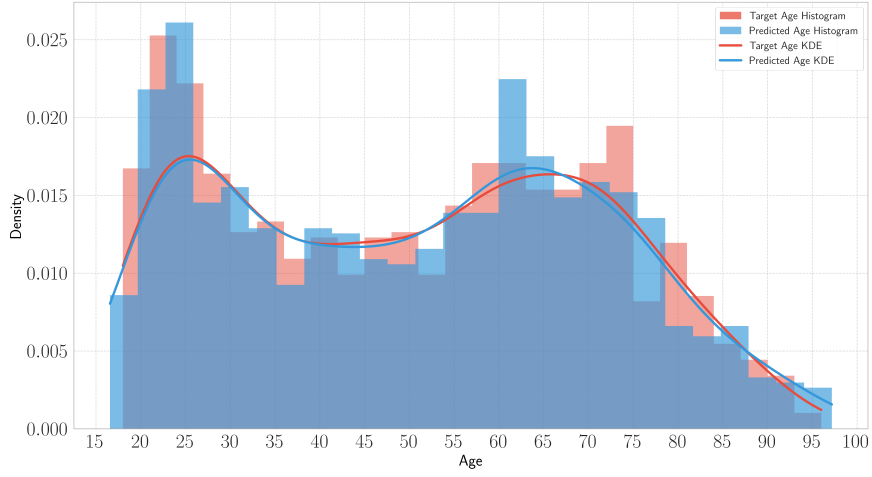


Figure 3.2: Comparison of Target and Predicted Age Distributions. The predicted age distribution is generated using a Mixture Density Network (MDN) with 20 gaussian components trained on preprocessed brain MRI data.

between the model-predicted and empirically observed distributions. For this purpose, Gaussian kernels with $\sigma = 1$ are employed for both the kernel for x and the kernel for y in the computation of these statistics, using a bootstrap sample size of 500. A significant discrepancy suggests that the model may not adequately represent the data. We compute p-values for each method using bootstrap resampling techniques to quantify the evidence against the null hypothesis

Table 3.1: Comparison of GF-KCSD and KCSD Test Results for Brain MRI Data.

Metric	GF-KCSD	KCSD
α	0.05	0.05
p-Value	0.81	0.98
Test Statistic	-9.59×10^{-8}	-5.0×10^{-4}
H0 Rejected	False	False
Simulations	500	500
Time (s)	0.57	1.99

The results are presented in Table 3.1. Both the GF-KCSD and KCSD methods

failed to reject the null hypothesis. The p-values associated with each method are 0.81 for GF-KCSD and 0.98 for KCSD, indicating weak evidence against the null hypothesis. The corresponding test statistics, representing the magnitude of deviation from the null hypothesis, are -9.59×10^{-8} and -5×10^{-4} for GF-KCSD and KCSD, respectively. These results highlight subtle distinctions in performance between the two methodologies.

To assess computational efficiency, we measured the mean execution time for both tests. Each test was executed 10 times, and the total time taken was recorded. The mean execution time per test was then calculated by dividing the total time by the number of executions. This process provides an estimate of the computational cost of performing the tests on the given dataset. The results indicated an average duration of 0.57 seconds for GF-KCSD and 1.99 seconds for KCSD per execution.

Notably, while both methods perform comparably in terms of p-values and test statistics, GF-KCSD requires significantly less computational time. This efficiency arises despite GF-KCSD involving an additional step to compute the surrogate model q . GF-KCSD computes the score function by deriving it from a simpler distribution compared to the more complex approach taken by KCSD. Although using the variational inference estimation of the proposed model p in GF-KCSD might lead to some loss of information, it achieves comparable results to the gradient-based KCSD in less time. The proposed model p could exhibit complex characteristics like multimodality or possess various minima; however, by utilizing an auxiliary distribution in GF-KCSD, which typically represents a unimodal distribution without such complexities, the derivation process becomes significantly more straightforward and time-efficient.

3.5 Conclusion

In this paper, we have explored the capabilities of the (GF-KCSD, a method we previously introduced). Our investigation delved into synthetic analyses that focused on challenging scenarios like models with measurement errors, as well as complex real-world datasets, with special emphasis on demanding situations such as brain MRI data analysis. The GF-KCSD has proven effective in managing Type-I errors and maintaining statistical power in the face of measurement errors, highlighting its robustness and computational efficiency in a variety of testing conditions. Looking ahead, several promising avenues for future research emerge. The GF-KCSD method could be adapted to discrete domains by modifying the Stein operator, as explored by Yang et al. (2018). Additionally, drawing inspiration from Matsubara et al. (2022), the capacity of gradient-free methods to conduct Bayesian inference without complex gradient calculations presents a compelling opportunity to develop a generalized framework for Bayesian inference in models with intractable gradients. This framework aims to leverage the robust and efficient properties of GF-KCSD across a broader spectrum of Bayesian applications, particularly in scenarios where gradient computation is challenging or unfeasible. Beyond these advancements, tuning hyperparameters can significantly enhance model performance. Such developments encourage further exploration into kernel-based methods for statistical inference. We leave these and other potential developments for future exploration.

Bibliography

- Afzali, E. and Muthukumarana, S. (2023). Gradient-free kernel conditional stein discrepancy goodness of fit testing. *Machine Learning with Applications*, page 100463. 45, 48, 54, 56, 64
- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212. 46
- Arcones, M. A. and Gine, E. (1992). On the bootstrap of u and v statistics. *The Annals of Statistics*, pages 655–674. 57
- Beheshti, I., Ganaie, M., Paliwal, V., Rastogi, A., Razzak, I., and Tanveer, M. (2021). Predicting brain age using machine learning algorithms: A comprehensive evaluation. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1432–1440. 71
- Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer google schola*, 2:645–678. 71
- Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232. 46
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799. 46
- Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons. 46

- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC. 48
- Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR. 47, 52
- Delaigle, A. and Van Keilegom, I. (2021). Deconvolution with unknown error distribution. In *Handbook of Measurement Error Models*, pages 245–270. Chapman and Hall/CRC. 49
- Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer. 56
- Fisher, M. A., Oates, C., et al. (2022). Gradient-free kernel stein discrepancy. *arXiv preprint arXiv:2207.02636*. 47
- Fuller, W. A. and Fuller, W. A. (1987). *Measurement error models*. WILEY. 49
- Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38. 46
- Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein’s method. *Advances in Neural Information Processing Systems*, 28. 47
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).

- A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773. 47, 64
- Huskova, M. and Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate u-statistics. *The Annals of Statistics*, pages 1811–1823. 57
- Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR. 47, 57, 63, 64
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 72
- Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR. 47, 52, 57
- Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022. 75
- Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175. 46
- Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14. 46

- Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, pages 583–602. 47, 50
- Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR. 57, 75
- Zheng, X. (2012). Testing parametric conditional distributions using the nonparametric smoothing method. *Metrika*, 75:455–469. 64

Chapter 4

Correcting Mode Proportion Bias in Generalized Bayesian Inference via a Weighted Kernel Stein Discrepancy

Generalized Bayesian Inference (GBI) provides a flexible framework for updating prior distributions using various loss functions instead of the traditional likelihoods, thereby enhancing robustness to model misspecification. However, GBI often encounters models with intractable likelihoods. Kernelized Stein Discrepancy (KSD), as in recent work, addresses this challenge by relying only on the gradient of the log-density. Despite this innovation, KSD–Bayes has a critical pathology: insensitivity to the relative weights of well-separated modes in multimodal posteriors. We address this by introducing a diffusion-weighted KSD with a θ -independent, tail-vanishing weight, preserving the Stein identity and likelihood-free computation while aiming to improve sensitivity to mode proportions. On our benchmarks, MS–KSD–Bayes shows improved mode sensitivity relative to KSD–Bayes and, in unimodal regimes with contamination, often produces tighter posteriors. Our construction inherits the standard KSD tail assumptions—most notably, the distantly

dissipative assumption—which effectively rules out heavy-tailed models. We do not evaluate large, high-dimensional, or heavy-tailed applications.

4.1 Introduction

Bayesian inference provides a principled framework for updating probabilistic beliefs about model parameters given data via the likelihood. Classical Bayesian methods, however, typically assume tractable likelihoods and well-specified models—assumptions that are often violated in complex, real-world settings. In such cases, when likelihoods are intractable or misspecified, Bayesian inference can yield misleading conclusions and unreliable predictions.

To address misspecification, the Generalized Bayesian Inference (GBI) framework (Bissiri et al., 2016; Jewson et al., 2018) redefines posterior construction by replacing the likelihood with a data-dependent loss. These loss-based approaches modify how parameter values are evaluated while leaving the model intact, which is valuable when the exact form of misspecification is unknown. A practical challenge remains when the likelihood is *intractable*: traditional MCMC algorithms require explicit likelihood evaluations, so posterior computation and sampling are difficult—or impossible—without specialized approximations.

Many practical models have intractable likelihoods of the form $p_\theta(x) = q_\theta(x)/Z_\theta$, with tractable unnormalized density q_θ but intractable normalizing constant $Z_\theta = \int q_\theta(x) dx$. The Bayesian posterior is then *doubly intractable* (Murray et al., 2012):

$$\pi_n(\theta) = \pi_0(\theta) \frac{1}{Z_X} \prod_{i=1}^n \frac{q_\theta(x_i)}{Z_\theta},$$

where $\pi_0(\theta)$ is the prior and Z_X is the marginal likelihood. Such models arise in, e.g., spatial interaction models (Besag, 1974; Diggle, 1990), exponential random graph models (Park and Haran, 2018), lattice/spatial models (Moore et al., 2020), and gene expression applications (Jiang et al., 2022).

A prominent score-based approach in GBI is *KSD-Bayes* (Matsubara et al., 2022), which uses the Kernel Stein Discrepancy (KSD) (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017) as the loss. Because KSD depends only on the score (the gradient of the log density) of an unnormalized model, it bypasses unknown normalizing constants and has been successful across challenging settings, including high-dimensional exponential-family models and precision estimation (Liu et al., 2019).

Despite these advances, Matsubara et al. (2022) and subsequent work show that KSD-based methods face a fundamental limitation in multimodal targets: when the posterior has multiple well-separated modes, KSD can be *mode-proportion blind*, i.e., relatively insensitive to differences in the mixing weights of the modes (Wenliang and Kanagawa, 2020; Liu et al., 2023). This pathology is distinct from the weak-convergence failures of certain kernels and persists even when using convergence-determining kernels (Gorham and Mackey, 2017). In practice, mode-proportion blindness can bias inference toward near-uniform mixtures, undermining posterior accuracy in mixture models (Bénard et al., 2024), gene expression with multimodal behavior (Jiang et al., 2022), and other applications where capturing mode proportions is essential.

We address this limitation by introducing a diffusion-weighted KSD for generalized

Bayes. Our construction uses a tail-vanishing weight anchored to a fixed pilot density, which is inserted into a diffusion-form Stein operator; this preserves the Stein identity under standard boundary conditions while improving sensitivity to mixing proportions and retaining tractability for unnormalized models. In contrast to prior weighted/diffusion-Stein or diffusion score-matching work—which target robustness or testing/thinning—we use a kernelized Stein loss to form a generalized posterior specifically aimed at mitigating mode-proportion blindness (Gorham et al., 2019; Barp et al., 2019; Altamirano et al., 2023a,b; Laplante et al., 2025; Liu et al., 2023; B  nard et al., 2024).

Outline. Section 4.2 reviews Generalized Bayesian Inference and Kernel Stein Discrepancy. Section 4.3 develops MS-KSD-Bayes and explains how the diffusion weighting addresses mode-proportion blindness while remaining likelihood-free. Section 4.4 presents empirical studies on synthetic and real data. Section 4.5 concludes and outlines future work.

4.2 Background

In this section we review *Kernelized Stein Discrepancy* (KSD), *Generalized Bayesian Inference* (GBI), and the *KSD-Bayes* framework as foundations for our proposed Mode-Sensitive KSD-Bayes approach.

Foundational Assumptions. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an *open* set and $\{x_i\}_{i=1}^n \stackrel{\text{iid}}{\sim} \mathbb{Q}$ with empirical measure $\mathbb{Q}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Let $\Theta \subseteq \mathbb{R}^p$ be a parameter set with nonempty interior (assume Θ is compact or the population risk has compact sublevel sets).

For each $\theta \in \Theta$, $\mathbb{P}_\theta$ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}$ and admits a density $p_\theta \in C^1(\mathcal{X})$ with $p_\theta(x) > 0$ for all $x \in \mathcal{X}$; hence the score $s_{p_\theta}(x) = \nabla_x \log p_\theta(x)$ is well defined. Unnormalized models $p_\theta = q_\theta/Z_\theta$ are allowed since $\nabla_x \log p_\theta = \nabla_x \log q_\theta$. We assume standard integrability ensuring Stein discrepancies and generalized posteriors are well defined, e.g. $\mathbb{E}_{\mathbb{P}_\theta} \|s_{p_\theta}(X)\| < \infty$.

4.2.1 Kernelized Stein Discrepancy (KSD)

Kernelized Stein Discrepancy (KSD) is a divergence measure introduced to assess discrepancies between two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$. KSD is particularly useful when the target distribution $\mathbb{P}$ has an intractable normalizing constant (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017) and has been extended to discrete distributions (Yang et al., 2018), conditional distributions (Jitkrittum et al., 2020), and intractable gradient target distributions (Afzali and Muthukumarana, 2023; Fisher and Oates, 2024; Afzali et al., 2024).

At its core, KSD leverages a Stein operator $\mathcal{A}_p$ associated with a density p on $\mathcal{X}$. Given a vector field $g : \mathcal{X} \rightarrow \mathbb{R}^d$ with components $g(x) = (g_1(x), \dots, g_d(x))^\top$, define

$$(\mathcal{A}_p g)(x) = s_p(x)^\top g(x) + \nabla \cdot g(x), \quad s_p(x) = \nabla_x \log p(x),$$

where $\nabla \cdot g(x) = \sum_{i=1}^d \partial g_i(x)/\partial x_i = \text{tr}(\nabla g(x))$ and $\nabla g(x) = (\partial g_i/\partial x_j)_{i,j=1}^d$ is the Jacobian of g .

Let $\mathcal{G}(k)$ denote the scalar RKHS induced by a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Define the product RKHS

$$\mathcal{G}(k)^d := \left\{ g = (g_1, \dots, g_d)^\top : g_i \in \mathcal{G}(k) \text{ for } i = 1, \dots, d \right\},$$

equipped with inner product and norm

$$\langle g, h \rangle_{\mathcal{G}(k)^d} = \sum_{i=1}^d \langle g_i, h_i \rangle_{\mathcal{G}(k)}, \quad \|g\|_{\mathcal{G}(k)^d} = \left(\sum_{i=1}^d \|g_i\|_{\mathcal{G}(k)}^2 \right)^{1/2}.$$

The *unit ball* of $\mathcal{G}(k)^d$ is

$$\mathbb{B}_1(\mathcal{G}(k)^d) = \left\{ g = (g_1, \dots, g_d)^\top : \sum_{i=1}^d \|g_i\|_{\mathcal{G}(k)}^2 \leq 1 \right\}.$$

When $k \in C^2(\mathcal{X} \times \mathcal{X})$ with bounded first and second partial derivatives, each component g_i is C^1 and obeys the reproducing property for derivatives,

$$\partial_{x_j} g_i(x) = \langle g_i, \partial_{x_j} k(x, \cdot) \rangle_{\mathcal{G}(k)},$$

so $\nabla g(x)$ and $\nabla \cdot g(x)$ are well defined for all $g \in \mathbb{B}_1(\mathcal{G}(k)^d)$. (Equivalently, this setup coincides with the vector-valued RKHS induced by the matrix-valued kernel $K(x, x') = k(x, x')I_d$.)

Under mild boundary conditions (integration by parts), the Stein identity holds:

$$\mathbb{E}_{X \sim p}[(\mathcal{A}_p g)(X)] = 0 \quad \text{for all } g \text{ in the unit ball of the product RKHS } \mathcal{G}(k)^d.$$

The Kernelized Stein Discrepancy (KSD) between $\mathbb{P}_\theta$ and $\mathbb{Q}$ distributions is then defined by

$$\text{KSD}(\mathbb{P}_\theta \| \mathbb{Q}) = \sup_{\|g\|_{\mathcal{G}(k)^d} \leq 1} \left| \mathbb{E}_{X \sim \mathbb{P}_\theta}[(\mathcal{A}_{p_\theta} g)(X)] - \mathbb{E}_{X \sim \mathbb{Q}}[(\mathcal{A}_{p_\theta} g)(X)] \right|. \quad (4.1)$$

By leveraging the Stein identity, this simplifies to:

$$\text{KSD}(\mathbb{P}_\theta \| \mathbb{Q}) = \sup_{\|g\|_{\mathcal{G}(k)^d} \leq 1} \left| \mathbb{E}_{X \sim \mathbb{Q}}[(\mathcal{A}_{p_\theta} g)(X)] \right|. \quad (4.2)$$

In practice, for a sample $\{x_i\}_{i=1}^n \sim \mathbb{Q}$, by the RKHS reproducing property ¹(point evaluation and, when $k \in C^1$, differentiation reduce to RKHS inner products with $k(x, \cdot)$ and its derivatives), the squared KSD admits the closed form

$$\begin{aligned} \text{KSD}^2(\mathbb{P}_\theta \parallel \mathbb{Q}) &= \mathbb{E}_{X, X' \sim \mathbb{Q}}[k_{p_\theta}(X, X')], \\ \widehat{\text{KSD}}^2(\mathbb{P}_\theta \parallel \mathbb{Q}_n) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{p_\theta}(x_i, x_j), \end{aligned} \quad (4.3)$$

where $k_{p_\theta}(x, x')$ is the Stein kernel, defined in terms of C^2 base kernel k and the score function $s_{p_\theta}(x)$ as:

$$\begin{aligned} k_{p_\theta}(x, x') &= \text{trace} \left(\nabla_x \nabla_{x'}^\top k(x, x') \right) \\ &\quad + s_{p_\theta}(x)^\top \nabla_{x'} k(x, x') \\ &\quad + s_{p_\theta}(x')^\top \nabla_x k(x, x') \\ &\quad + (s_{p_\theta}(x)^\top s_{p_\theta}(x')) k(x, x'). \end{aligned}$$

KSD is particularly advantageous because it depends only on the score $s_{p_\theta}(x) = \nabla_x \log p_\theta(x)$, and therefore does not require the normalizing constant (Gorham and Mackey, 2017). With the inverse multiquadric (IMQ) kernel $k(x, x') = (c + \|x - x'\|^2)^{-\beta}$ ($c > 0$, $\beta \in (0, 1)$), convergence guarantees are available under standard boundary conditions when the target is distantly dissipative—that is, when $-\log p_\theta(x)$ grows sufficiently fast as $\|x\| \rightarrow \infty$ (Gorham and Mackey, 2017). This implies tails that are light enough for IMQ-KSD to control convergence in the settings where its theory applies.

¹Reproducing property: $f(x) = \langle f, k(x, \cdot) \rangle$; if $k \in C^1$, $\partial_{x_j} f(x) = \langle f, \partial_{x_j} k(x, \cdot) \rangle$. Applies componentwise for $g \in \mathcal{G}(k)^d$.

These guarantees rely on distantly dissipative assumption and therefore exclude heavy-tailed targets; in practice they can also lead to sensitivity to kernel choice/bandwidth or slower or unstable behavior in challenging regimes. MS-KSD-Bayes inherits the same tail assumptions, so our experiments focus on light-tailed settings, and we do not claim results for heavy-tailed or very high-dimensional models. See Gorham and Mackey (2017); Chwialkowski et al. (2016); Liu et al. (2016) for details.

4.2.2 Generalized Bayesian Inference with KSD

Traditional Bayesian inference updates prior beliefs π_0 about model parameters $\theta \in \Theta$ based on observed data via the likelihood function. While this framework is optimal under correct specification, real-world applications often involve misspecification due to simplifying assumptions or incomplete knowledge of the data-generating process, potentially leading to unreliable inferences.

To address this, alternative Bayesian approaches have been developed to enhance robustness, including robust priors (Berger et al., 1994), power posteriors (Holmes and Walker, 2017), coarsened posteriors (Miller and Dunson, 2019), and scoring rule-based inference (Giummolè et al., 2019). Among these, *Generalized Bayesian Inference* (GBI) (Bissiri et al., 2016) provides a flexible framework by replacing the likelihood with a data-dependent loss function $L_n(\theta)$.

For data $\{x_i\}_{i=1}^n$ drawn from $\mathbb{Q}$ and a prior distribution $\pi_0(\theta)$ over Θ , the generalized posterior is

$$\pi_n^\ell(\theta) \propto \pi_0(\theta) \exp\left\{-\alpha n L_n(\theta)\right\}, \quad (4.4)$$

where $\alpha > 0$ is a temperature (scaling) parameter that controls how strongly the loss influences the update. Because the factor αn multiplies the sample-average loss, α acts like a learning-rate or precision weight: larger α increases the influence of the loss and typically yields a more concentrated posterior, whereas smaller α downweights the loss, giving the prior relatively more influence and often improving robustness. When $\alpha = 1$ and $L_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log p_\theta(x_i)$, standard Bayesian updating is recovered. Values $\alpha < 1$ temper the loss and can add robustness; values $\alpha > 1$ increase concentration and are usually unnecessary unless the loss is under-scaled. In practice, α may be fixed in advance or calibrated by predictive criteria or cross-validation to trade off robustness and efficiency. A generally accepted, fully principled rule for selecting α in misspecified and non-likelihood settings remains an open problem.

Choosing a robust loss $L_n(\theta)$ mitigates misspecification without requiring explicit knowledge of the likelihood’s flaws. The properties and limitations of the chosen divergence determine the theoretical behavior of the generalized posterior (Knoblauch et al., 2019). However, implementing divergence-based losses becomes challenging when the likelihood is *intractable*, which is common in modern statistical models involving complex stochastic processes or high-dimensional data.

KSD–Bayes. For intractable likelihoods within the GBI framework, Matsubara et al. (2022) propose *KSD–Bayes*, which uses the Kernelized Stein Discrepancy as a proxy loss,

$$L_n(\theta) = \text{KSD}^2(\mathbb{P}_\theta \| \mathbb{Q}_n), \quad \pi_n^{\text{KSD}}(\theta) \propto \pi_0(\theta) \exp\left\{-\alpha n \text{KSD}^2(\mathbb{P}_\theta \| \mathbb{Q}_n)\right\}.$$

KSD–Bayes circumvents explicit likelihood evaluations and normalizing constants

by relying only on score functions and kernel computations; it is computationally practical with standard MCMC over θ , scales to high-dimensional models, and (under suitable regularity) enjoys guarantees such as posterior consistency and local normal approximations.

Nonetheless, KSD–Bayes exhibits *mode-proportion blindness*: it can fail to capture the relative weights of well-separated modes (Gorham et al., 2019; Wenliang and Kanagawa, 2020; Liu et al., 2023). Intuitively, KSD-based—and more broadly score-based—losses may be comparatively insensitive to mixture weights when local score fields are similar across modes. As noted by Matsubara et al. (2022, Sec. 3.5), the data score $\nabla_x \log p_\theta(x)$ can change little as mixing proportions vary, leading the KSD loss to prefer overly uniform mixtures and thus biasing inference. B  nard et al. (2024) formalize this pathology in Gaussian mixtures: for $p_\theta(x) = w_1 \mathcal{N}(x; \mu, \sigma^2) + (1 - w_1) \mathcal{N}(x; -\mu, \sigma^2)$, as the separation μ increases the inferred weight tends toward $w_1^* \approx 0.5$ regardless of the true w_1 . This limitation undermines reliability in multimodal posterior inference, where recovering mode contributions is essential.

These observations motivate our mode-sensitive extension. In the next section we introduce **MS–KSD–Bayes**, which retains the score-only, likelihood-free advantages of KSD–Bayes—computations depend on p_θ only through the data score $s_{p_\theta}(x) = \nabla_x \log p_\theta(x)$ —while mitigating mode-proportion blindness via a density-weighted Stein operator that restores sensitivity to mixing proportions.

4.3 Methodology

In this section, we introduce a diffusion-weighted Stein discrepancy to recover mode-proportion sensitivity, define the diffusion Stein operator with a tail-vanishing, θ -independent diffusion term, and construct the Mode-Sensitive KSD (MS-KSD) together with the corresponding generalized posterior.

4.3.1 Diffusion Stein operator

Let $\mathcal{X} \subset \mathbb{R}^d$ be open. For each $\theta \in \Theta$, let p_θ be a strictly positive density on $\mathcal{X}$ with data score $s_{p_\theta}(x) = \nabla_x \log p_\theta(x)$. Given a diffusion term (weight) $w : \mathcal{X} \rightarrow [0, \infty)$ and a test function $g : \mathcal{X} \rightarrow \mathbb{R}^d$, define the *diffusion Stein operator*

$$\begin{aligned} (\mathcal{A}_{p_\theta}^w g)(x) &= \frac{1}{p_\theta(x)} \nabla \cdot (w(x) g(x) p_\theta(x)) \\ &= w(x) \{ s_{p_\theta}(x)^\top g(x) + \nabla \cdot g(x) \} + \nabla w(x)^\top g(x). \end{aligned} \quad (4.5)$$

Assumption 1 (Regularity and boundary). *We adopt the regularity and boundary conditions for diffusion-Stein operators from Gorham et al. (2019, Sec. 2; App. A). In particular:*

1. $p_\theta \in C^1(\mathcal{X})$, p_θ is a Lebesgue density on $\mathcal{X}$ with $p_\theta(x) > 0$, and $\mathbb{E}_{p_\theta} \|s_{p_\theta}(X)\| < \infty$.
2. $w \in C^1(\mathcal{X})$, $0 \leq w(x) \leq C_w < \infty$ for all $x \in \mathcal{X}$, and ∇w is bounded.
3. The base kernel k is C^2 with bounded first and second partial derivatives, and test functions g lie in the unit ball of the product RKHS $\mathcal{G}(k)^d$.

4. *The boundary term vanishes:*

$$\int_{\partial B_R} n(x)^\top w(x) g(x) p_\theta(x) dS(x) \rightarrow 0 \quad \text{as} \quad R \rightarrow \infty.$$

where $B_R = \{x : \|x\| \leq R\}$, $n(x)$ is the outward unit normal on ∂B_R (i.e., $x/\|x\|$), and dS is the surface-area element on ∂B_R .

Proposition 4 (Stein identity). *Under Assumption 1, for every test function $g \in \mathcal{G}(k)^d$ with $\|g\|_{\mathcal{G}(k)^d} \leq 1$,*

$$\mathbb{E}_{X \sim p_\theta} [(\mathcal{A}_{p_\theta}^w g)(X)] = 0. \quad (4.6)$$

This is the reversible specialization ($a(x) = w(x)I_d$) of the diffusion–Stein identity in Gorham et al. (2019, Sec. 2; App. A), under the boundary conditions stated in Assumption 1.

Remark (tail behavior and operator assumptions): Our use of the diffusion–form Stein operator follows the usual KSD regularity: a C^2 base kernel with bounded partials and target tails strong enough (e.g., distantly dissipative) to justify integration by parts and boundary decay (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017). The diffusion weight w is bounded and tail-vanishing. MS–KSD–Bayes does not relax the distantly dissipativity requirement; when targets are heavy-tailed or otherwise violate these conditions, prior work reports degraded testing power or unstable behavior, and our method may similarly be affected. We therefore view MS–KSD–Bayes as most appropriate under the same tail regimes where classical KSD is known to perform well.

4.3.2 Mode-sensitive diffusion term

Definition 2 (Mode-sensitive diffusion term). *A diffusion term is a measurable map $w : \mathcal{X} \rightarrow [0, \infty)$ used inside the divergence form of Equation (4.5). For mode sensitivity, we restrict attention to diffusion terms w that are θ -independent, bounded and tail-vanishing.*

Assumption 2 (Design conditions for w). *Throughout MS-KSD we impose the following standing conditions on $w : \mathcal{X} \rightarrow [0, \infty)$:*

1. w does not depend on θ .
2. $0 \leq w(x) \leq C_w < \infty$ for all $x \in \mathcal{X}$ and $w(x) \rightarrow 0$ as $\|x\| \rightarrow \infty$.

The diffusion Stein identity (Proposition 4) holds for any admissible w ; thus, mode sensitivity is governed by *where* the operator places weight. Item (1) (parameter independence) prevents feedback between θ and the discrepancy, preserving likelihood-free tractability and aligning with the KSD-Bayes setting. Item (2) (bounded, tail-vanishing) suppresses long-range interactions in the contracted Stein kernel, attenuating cross-component terms for well-separated modes while preserving within-mode contributions. This restores sensitivity to mixing proportions and mitigates the mode-proportion blindness documented for unweighted Langevin operators (Wenliang and Kanagawa, 2020; Liu et al., 2023; Bénard et al., 2024).

In our empirical studies we fix an *anchor density* $p_\star$ that does not depend on θ (either a pilot model or a data-driven KDE), and take

$$w(x) = \min \left\{ C_w, \left(\frac{p_\star(x)}{p_\star(x) + \tau} \right)^\gamma \right\}, \quad C_w > 0, \tau > 0, \gamma \in (0, 1]. \quad (4.7)$$

This choice is θ -independent, bounded by C_w , and satisfies $w(x) \rightarrow 0$ as $\|x\| \rightarrow \infty$ whenever $p_\star(x) \rightarrow 0$ in the tails. When $p_\star \in C^1$ with mild decay, ∇w inherits bounded growth via the chain rule, so Assumption 2 and Assumption 1 apply. Throughout the analysis we treat $p_\star$ (and thus w) as fixed; when $p_\star$ is estimated from data (e.g., KDE), all statements are understood conditional on the realized w . Dataset-specific choices of $p_\star$, kernel, bandwidth, and (C_w, τ, γ) are reported in Section 4.4.

From this point forward, we adopt the vector-valued RKHS setting induced by a matrix-valued (operator-valued) kernel K , so test functions are $g : \mathcal{X} \rightarrow \mathbb{R}^d$ with $g \in \mathcal{G}(K)$. This framework subsumes the earlier product-RKHS case via the choice $K(x, x') = k_0(x, x')I_d$.

4.3.3 Vector-valued RKHS and matrix-valued kernels

Definition 3 (Matrix-valued kernel and test class). *Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{d \times d}$ be a symmetric, positive-definite operator-valued kernel, i.e., $K(x, x') = K(x', x)^\top$ and $\sum_{i,j=1}^n v_i^\top K(x_i, x_j) v_j \geq 0$ for any $x_1, \dots, x_n \in \mathcal{X}$ and $v_1, \dots, v_n \in \mathbb{R}^d$. The associated vector-valued RKHS is denoted $\mathcal{G}(K)$; our test functions g will be taken from the unit ball of $\mathcal{G}(K)$.*

Assumption 3 (Matrix kernel and operator contraction). *We require the following to ensure the diffusion Stein operator acts coordinatewise on K and the induced V -statistic is finite:*

1. K is symmetric and operator-valued positive-definite as in Definition 3.

2. For each $(x, x') \in \mathcal{X} \times \mathcal{X}$, the contraction

$$K_{p_\theta}^w(x, x') := (\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'}) K(x, x') \in \mathbb{R}^{d \times d}$$

exists and is continuous.

3. $\mathbb{E}_{X, X' \sim \mathbb{Q}}[\text{tr}(K_{p_\theta}^w(X, X'))] < \infty$ for the $\mathbb{Q}$ of interest.

A convenient construction (used in KSD–Bayes) is

$$K(x, x') = M(x) k_0(x, x') M(x')^\top, \quad (4.8)$$

with scalar base kernel k_0 of class C^2 and a matrix field $M \in C_b^1(\mathcal{X}; \mathbb{R}^{d \times d})$ (bounded norm and bounded first derivatives). Under Assumption 1, these smoothness and boundedness conditions are sufficient for the coordinatewise action of $\mathcal{A}_{p_\theta}^w$ on both arguments of K , for continuity of $K_{p_\theta}^w$, and for the trace integrability in item (3).

Taking $M \equiv I_d$ gives $K(x, x') = k_0(x, x') I_d$ and $\mathcal{G}(K) = \mathcal{G}(k_0)^d$, reducing to the product–RKHS setting used in classical KSD.

4.3.4 Mode–Sensitive Kernel Stein Discrepancy (MS–KSD)

Definition 4 (MS–KSD). Let $\mathcal{G}(K)$ be the vector-valued RKHS induced by the matrix-valued kernel K , and let $\mathbb{Q}$ be a probability measure on $\mathcal{X}$. For each $\theta \in \Theta$, the mode-sensitive KSD associated with (p_θ, w) is

$$\text{KSD}_w(\mathbb{P}_\theta \| \mathbb{Q}) = \sup_{\|g\|_{\mathcal{G}(K)} \leq 1} \left| \mathbb{E}_{X \sim \mathbb{Q}}[(\mathcal{A}_{p_\theta}^w g)(X)] \right|. \quad (4.9)$$

By Proposition 4, the expectation under p_θ vanishes, so the discrepancy depends on $\mathbb{Q}$ only.

Proposition 5 (Kernel representation of MS–KSD and V-statistic). *Under Assumptions 1 and 3, the coordinatewise contraction*

$$K_{p_\theta}^w(x, x') := (\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'}) K(x, x') \in \mathbb{R}^{d \times d}$$

is well defined and continuous. Moreover, the mode-sensitive diffusion KSD admits the RKHS representation

$$\text{KSD}_w^2(\mathbb{P}_\theta \| \mathbb{Q}) = \mathbb{E}_{X, X' \sim \mathbb{Q}} [\text{tr}(K_{p_\theta}^w(X, X'))]. \quad (4.10)$$

Given $X_{1:n} \stackrel{\text{iid}}{\sim} \mathbb{Q}$ with empirical measure $\mathbb{Q}_n$, the empirical MS–KSD is the V-statistic

$$\widehat{\text{KSD}}_w^2(\mathbb{P}_\theta \| \mathbb{Q}_n) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{tr}(K_{p_\theta}^w(X_i, X_j)). \quad (4.11)$$

4.3.5 MS–KSD–Bayes posterior

We incorporate MS–KSD into generalized Bayesian inference by defining, for data $\{X_i\}_{i=1}^n \sim \mathbb{Q}$ and prior $\pi_0(\theta)$ on Θ ,

$$\pi_n^w(\theta) \propto \pi_0(\theta) \exp \left\{ -\alpha n \text{KSD}_w^2(\mathbb{P}_\theta \| \mathbb{Q}_n) \right\}, \quad (4.12)$$

where $\alpha > 0$ and $\text{KSD}_w^2(\mathbb{P}_\theta \| \mathbb{Q}_n)$ is the empirical MS–KSD in Equation (4.11). Under Assumptions 1, 2, and 3, the loss is well defined and locally bounded in θ , so the right-hand side is integrable for standard priors. The update is likelihood-free because KSD_w depends on p_θ only through its score $s_{p_\theta} = \nabla \log p_\theta$. If $w \equiv c \in (0, \infty)$ is constant then $\mathcal{A}_{p_\theta}^w = c \mathcal{A}_{p_\theta}$ and $\text{KSD}_w^2 = c^2 \text{KSD}^2$, so Equation (4.12) reduces to

the KSD–Bayes posterior after rescaling $\alpha \mapsto c^2\alpha$. More generally, for any $c > 0$ the replacements $(\alpha, w) \mapsto (c^2\alpha, cw)$ leave Equation (4.12) unchanged; consequently the *shape* of w drives mode sensitivity, while its global scale can be absorbed into α .

4.3.6 Generalized Bayesian inference for exponential families

We illustrate how MS–KSD–Bayes applies to (natural) exponential families and yields a closed–form generalized posterior under Gaussian priors. Let $(\mathcal{X}, \mathcal{F})$ be measurable and $\Theta \subseteq \mathbb{R}^p$. A family $\{p_\theta : \theta \in \Theta\}$ is an exponential family if

$$p_\theta(x) = \exp\left(\xi(\theta)^\top t(x) - a(\theta) + b(x)\right), \quad (4.13)$$

with sufficient statistic $t : \mathcal{X} \rightarrow \mathbb{R}^p$, natural parameter $\xi(\theta) \in \mathbb{R}^p$, log-partition a , and carrier b . For MS–KSD, only the *data score* is needed,

$$s_{p_\theta}(x) = \nabla_x \log p_\theta(x) = J_t(x)^\top \xi(\theta) + \nabla_x b(x), \quad (4.14)$$

so $a(\theta)$ (and any normalizing constant) never appears. With the θ –independent, bounded, tail–vanishing diffusion w from Assumption 2 and the matrix–valued kernel K of Assumption 3, the empirical MS–KSD loss is

$$L_n(\theta) := \text{KSD}_w^2(\mathbb{P}_\theta \| \mathbb{Q}_n) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{tr}(K_{p_\theta}^w(X_i, X_j)), \quad K_{p_\theta}^w = (\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'}) K.$$

Moreover, $L_n(\theta)$ is a quadratic polynomial in $\xi(\theta)$:

$$L_n(\theta) = \xi(\theta)^\top \Gamma_n \xi(\theta) + \xi(\theta)^\top \tau_n + c_n, \quad (4.15)$$

where $\Gamma_n \in \mathbb{R}^{p \times p}$ is symmetric positive semidefinite, $\tau_n \in \mathbb{R}^p$, and $c_n \in \mathbb{R}$ depend on the data $\{X_i\}_{i=1}^n$, on w and K , but not on θ . (Explicit expressions obtained by expanding $(\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'})K$ using Equation (4.14) are collected in Appendix A.2.3.)

Proposition 6 (Gaussian conjugacy for MS–KSD–Bayes). *Assume a natural exponential family $(\xi(\theta) = \theta)$ and a Gaussian prior $\pi_0(\theta) \propto \exp(-\frac{1}{2}(\theta - \mu_0)^\top \Sigma_0^{-1}(\theta - \mu_0))$. The MS–KSD–Bayes posterior based on*

$$\pi_n^w(\theta) \propto \pi_0(\theta) \exp\left\{-\alpha n L_n(\theta)\right\}, \quad \alpha > 0, \quad (4.16)$$

is Gaussian:

$$\pi_n^w(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu_n)^\top \Sigma_n^{-1}(\theta - \mu_n)\right), \quad (4.17)$$

with precision and mean

$$\Sigma_n^{-1} = \Sigma_0^{-1} + 2\alpha n \Gamma_n, \quad \mu_n = \Sigma_n(\Sigma_0^{-1}\mu_0 - \alpha n \tau_n).$$

The constant c_n in Equation (4.15) affects only the normalizing constant and not (μ_n, Σ_n) .

Outside exponential families (and our KEF examples), we rely on the general form (4.12) and MCMC; we do not provide additional theory or large-scale evaluations for those cases in this chapter.

For **non-conjugate** models, posterior computation targets the form in Equation (4.12).

A naive evaluation of the empirical MS–KSD is $O(n^2)$ due to the double sum. Many terms, however, are independent of θ and can be cached once, e.g., $\{w(X_i), \nabla w(X_i)\}_{i=1}^n$ and the kernel blocks. In natural exponential families this caching enables the

quadratic decomposition $L_n(\theta) = \xi(\theta)^\top \Gamma_n \xi(\theta) + \xi(\theta)^\top \tau_n + c_n$, so subsequent evaluations reduce to cheap quadratic forms in $\xi(\theta)$. For large n , stochastic approximations to the U-statistic or V-statistic (mini-batching or low-rank kernel approximations) lower per-iteration cost from $O(n^2)$ to $O(B^2)$ with $B \ll n$, yielding unbiased estimates in the U-statistic case and low-bias surrogates for V-statistics (Huggins and Mackey, 2018; Gorham et al., 2020). This preserves the likelihood-free nature of MS-KSD-Bayes while improving scalability.

All proofs are deferred to the Appendix A.2.

4.4 Experiment Design

Following the approach in the KSD-Bayes benchmark study (Matsubara et al., 2022), we employ the *Kernel Exponential Family* (KEF) model to compare the performance of MS-KSD-Bayes against KSD-Bayes. This framework provides a flexible yet computationally efficient way to model complex distributions, particularly when normalizing constants are intractable. In what follows, we showcase MS-KSD-Bayes’ ability to infer correct mode proportions in two different multimodal settings. Section 4.4.1 examines the widely studied *Galaxy* dataset, a classic benchmark for comparing multimodal inference methods, while Section 4.4.2 focuses on a real-world *Lung Cancer Gene Expression* dataset, offering insight into more complex, biologically driven mode structures. These experiments collectively illustrate how MS-KSD-Bayes addresses a known limitation of KSD-Bayes by more accurately capturing well-separated modes in our settings.

All datasets considered here are light-tailed after preprocessing; we do not evaluate

heavy-tailed targets or very high-dimensional structured models in this chapter.

Throughout the experiments we instantiate the diffusion weight from Equation (4.7) by taking $p_\star(x) = \tilde{p}(x)$, a data-driven proxy for the (possibly unnormalized) likelihood, and setting $\tau > 0$. Unless stated otherwise we use $C_w = 1$ and $\gamma = 1$, so that

$$w(x) = \left(\frac{\tilde{p}(x)}{\tilde{p}(x) + \tau} \right)^\gamma, \quad \tau > 0, \gamma \in (0, 1].$$

This choice is θ -independent, bounded ($0 \leq w \leq 1$), and tail-vanishing ($w(x) \rightarrow 0$ whenever $\tilde{p}(x) \rightarrow 0$). The gradient has the closed form

$$\nabla w(x) = \gamma \tau \frac{\tilde{p}(x)^{\gamma-1}}{(\tilde{p}(x) + \tau)^{\gamma+1}} \nabla \tilde{p}(x),$$

which facilitates efficient evaluation inside the diffusion Stein operator. (Note that any global rescaling of w can be absorbed into the temperature α ; only the *shape* governed by (γ, τ) matters.)

Generalized Bayesian KEF Formulation. In this setting, the probability density function is written as

$$p_\theta(x) \propto q(x) \exp\left(\langle f, \kappa(\cdot, x) \rangle_{\mathcal{G}(\kappa)}\right), \quad (4.18)$$

where $q(x)$ is the reference density, chosen to be $\mathcal{N}(0, S^2)$ with S denoting the width (standard deviation) of the reference measure, $\kappa(x, y)$ is the reproducing kernel (here, a Gaussian RBF kernel), and f lies in the corresponding Reproducing Kernel Hilbert Space (RKHS) $\mathcal{G}(\kappa)$. This approach allows for a nonparametric

representation of probability distributions. To avoid computing the typically intractable normalizing constant, we approximate $f(x)$ by a finite-rank expansion $f(x) = \sum_{i=1}^p \theta_{(i)} \phi_{(i)}(x)$, where $\{\phi_{(i)}(x)\}_{i=1}^p$ are suitable basis functions, and $\theta_{(i)}$ are coefficients to be inferred. We adopt Hermite polynomials as basis functions $\phi_{(j+1)}(x) = \frac{x^j}{\sqrt{j!}} \exp(-x^2/2)$ which helps ensure numerical stability and an effective approximation within the RKHS. The parameter vector $\theta_{(i)}$ follows independent Gaussian priors $\theta_{(i)} \sim \mathcal{N}(0, L^2 i^{-\eta})$, where L is a global scale parameter governing the overall prior variance, and η adjusts how aggressively higher-order terms are shrunk.

For the Stein loss, we employ the inverse multiquadric (IMQ) kernel due to its favorable convergence properties in Stein-based methods. In the experiments described below, we focus on MS-KSD-Bayes' ability to capture correct mode proportions in multimodal settings, thus demonstrating how it mitigates the limitations of KSD-Bayes in scenarios with well-separated modes.

In the generalized posterior Equation (4.12), any global rescaling of w can be absorbed into α . Consequently, only the *shape* of w affects mode sensitivity, while its global scale is immaterial for comparisons. In all experiments we keep α fixed across methods.

4.4.1 Galaxy Velocity Data Experiments

To evaluate MS-KSD-Bayes in multimodal settings, we replicate the KSD-Bayes experiment using the *Galaxy* velocity dataset (Postman et al., 1986; Roeder, 1990), which comprises $n = 82$ velocity measurements (km/s) from six cones in the Corona

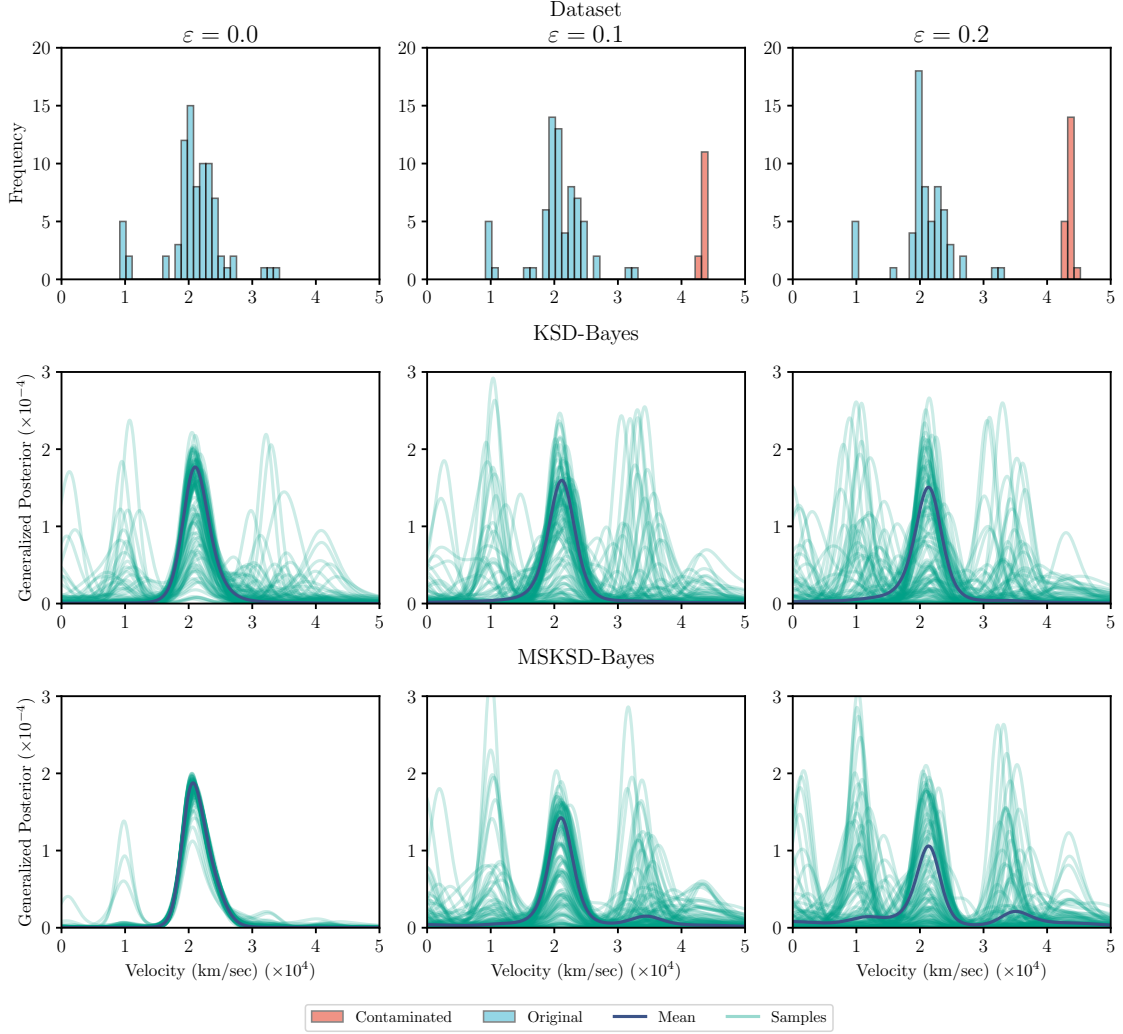


Figure 4.1: Posterior density estimates for the Galaxy dataset. Top: data histogram. Middle: KSD-Bayes (insensitive to mode proportions). Bottom: MS-KSD-Bayes (captures the multimodal structure with more faithful proportions in this example).

Borealis cluster survey. This dataset is well known for its multimodal structure, making it a natural benchmark for mode-sensitive inference. We adopt $p = 25$ Hermite polynomial basis functions within the KEF framework, set the reference density to $q(x) = \mathcal{N}(0, 3^2)$ (so $S = 3$), and use a Gaussian RBF kernel $\kappa(x, y) =$

$\exp(-\frac{(x-y)^2}{2})$. We fix $L = 10$ as the global prior scale and set $\eta = 1.1$, matching the original KSD–Bayes configuration. These choices focus the comparison on each method’s ability to recover distinct modes and their relative proportions.

Sensitivity to mode proportions: KSD–Bayes is known to be relatively insensitive to mode proportions when modes are well separated. To introduce a controlled secondary mode, we work with velocities rescaled by 10^4 km/s and contaminate the data by replacing a fraction ϵ of observations with draws from $\mathcal{N}(y, 0.1^2)$ at $y = 5$ (so $y = 5$ corresponds to 5×10^4 km/s in the original units). This tests whether the methods capture bimodality or remain biased toward a unimodal fit.

Figure 4.1 compares generalized posterior densities from KSD–Bayes and MS–KSD–Bayes across $\epsilon \in \{0.0, 0.1, 0.2\}$. For $\epsilon = 0$, the histogram shows multiple, not sharply separated peaks, and both methods yield unimodal posteriors (no spurious bimodality). At $\epsilon = 0.1$, a well-separated secondary peak emerges near the contaminated location; MS–KSD–Bayes more distinctly separates it from the primary mode, whereas KSD–Bayes remains essentially unimodal under the same settings. At $\epsilon = 0.2$, this contrast persists: MS–KSD–Bayes captures both peaks and more closely preserves their relative proportions, while KSD–Bayes often remains unimodal. These results highlight MS–KSD–Bayes’ heightened sensitivity to well-separated modes and improved recovery of mixing proportions.

4.4.2 Lung Cancer Gene Expression Experiments

Accurately capturing multiple modes in gene expression is crucial, as distinct tumor subtypes often display divergent expression patterns. We therefore test whether

MS-KSD-Bayes can detect and appropriately weight multiple modes in a genomic setting by focusing on a well-separated bimodal gene.

We evaluate MS-KSD-Bayes and KSD-Bayes on a lung cancer gene expression dataset from the `oompaData` library for R (Coombes, 2024; R Core Team, 2024; Abrams et al., 2018; Asiaee et al., 2020). The dataset (`lung.dataset`) contains log-transformed expression values for 150 genes measured in 444 tumor samples, with clinical annotations in `lung.clinical`. Expression values were processed as $\log_2(1+x)$. A subset of genes is used for illustration; full details are available under GEO accession [GSE68571](#).

To identify a probe set with a bimodal profile, we applied the `bimodalIndex` function from the `BimodalIndex` package (Coombes, 2019; Wang et al., 2009) to each gene in `lung.dataset`. This fits a two-component Gaussian mixture and computes a *Bimodality Index* (BI) that quantifies both separation and relative proportions. Ranking by BI, we selected AFFX-R2-EC-BIOD-5_AT as the most strongly bimodal (BI = 3.49). For this probe set, we use the same KEF-based formulation as in Section 4.4.1, with $p = 10$ Hermite basis functions, reference density $q(x) = \mathcal{N}(0, 4^2)$ (so $S = 4$), Gaussian RBF κ , and Gaussian priors with global scale $L = 9$ and decay exponent $\eta = 1.2$. All other settings mirror the Galaxy experiments to isolate the effect of the weighted Stein loss.

Figure 4.2 summarizes the results. The left panel shows the histogram of the selected gene, indicating a clear bimodal pattern. The middle panel reports the generalized posterior density under KSD-Bayes, which remains unimodal and fails to capture the second mode. The right panel shows MS-KSD-Bayes, which recovers two modes and more closely reflects their relative proportions. This visual com-

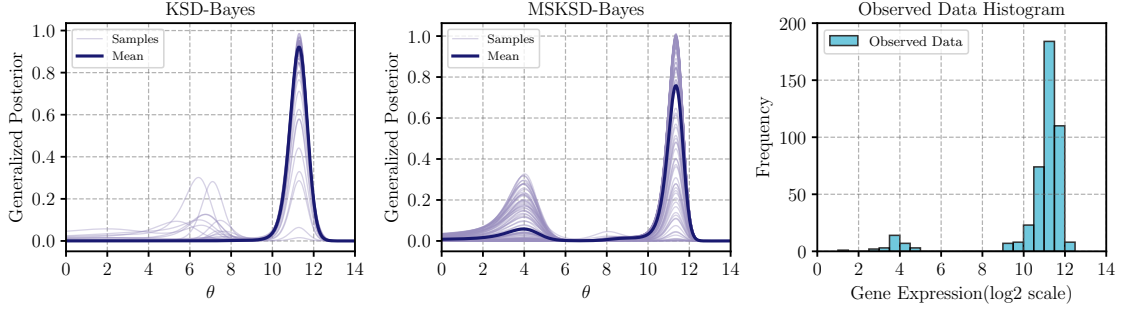


Figure 4.2: Generalized posterior density estimates for the lung cancer gene expression dataset (Setting: $p = 10$, $L = 9$, prior scale = 81, prior decay = 1.2). From right to left: (1) Histogram of the observed gene expression data; (2) Posterior density estimated via MS-KSD-Bayes; (3) Posterior density estimated via KSD-Bayes. Unlike KSD-Bayes, MS-KSD-Bayes captures the bimodal structure more clearly in this example.

parison illustrates the added mode sensitivity introduced by the diffusion-weighted Stein operator on this dataset.

By attenuating long-range cross-component interactions and reweighting by local density, MS-KSD-Bayes yields posterior inferences that better align with the multimodal structure of the expression data. In this setting, it identifies distinct biological subpopulations without introducing spurious modes, suggesting it is a promising practical alternative for Bayesian analysis of heterogeneous genomic cohorts.

4.4.3 Gaussian Location Experiment: Unimodal Setting with Contamination

Having demonstrated in Sections 4.4.1 and 4.4.2 that MS-KSD-Bayes effectively resolves mode-proportion blindness for well-separated modes, we now examine its performance in a unimodal Gaussian location setup with outlier contamination.

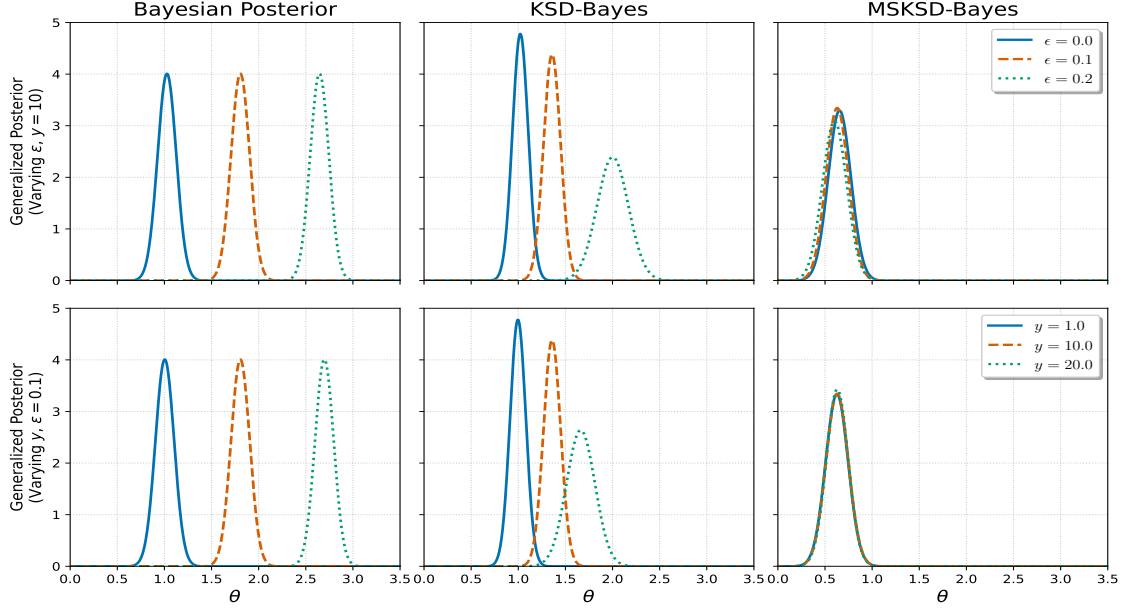


Figure 4.3: Posterior distributions for a Gaussian location model under varying contamination proportions (ϵ , top row) and outlier displacements (y , bottom row). Columns: (left) standard Bayes with a conjugate prior, (center) KSD-Bayes, (right) MS-KSD-Bayes. As ϵ or y increase, the first two methods shift away from $\theta_\star = 1$, while MS-KSD-Bayes remains comparatively concentrated near $\theta_\star$.

This experiment follows the design in (Matsubara et al., 2022) for KSD-Bayes, enabling a direct comparison. Specifically, we simulate data $X_i \in \mathbb{R}$ from the contaminated mixture

$$X_i \sim (1 - \epsilon) \mathcal{N}(\theta_\star, 1) + \epsilon \mathcal{N}(y, 1), \quad \theta_\star = 1,$$

with sample size $n = 100$ and a prior $\theta \sim \mathcal{N}(0, 1)$. The resulting dataset is unimodal around $x = 1$ but contains an increasing fraction of outliers as ϵ or the displacement y grows.

Figure 4.3 shows posteriors under three procedures: (i) standard Bayesian updating with a conjugate normal prior, (ii) the KSD-Bayes estimator, and (iii) MS-KSD-Bayes. The standard Bayesian posterior shifts substantially from $\theta_\star = 1$

as contamination strengthens, and KSD–Bayes offers only partial resistance. By contrast, MS–KSD–Bayes remains closely centered on $\theta_\star$ even for more extreme contamination, without discernible over-dispersion, suggesting that the weighting mechanism attenuates the influence of low-density observations while preserving posterior sharpness in this setting.

These results suggest that MS–KSD–Bayes addresses multimodal complexities and exhibits improved robustness in unimodal regimes with contamination, reducing bias and variance inflation relative to standard Bayes and non-weighted KSD–Bayes on our benchmarks.

4.5 Conclusion and Future Work

We introduced *MS–KSD–Bayes*, a likelihood-free generalized Bayesian update built from a diffusion-weighted Stein operator with a θ -independent, tail-vanishing weight w . The construction keeps the score-only advantage of KSD–Bayes while restoring sensitivity to mixing proportions. Across the Galaxy benchmark, a bimodal gene expression example, and a contaminated Gaussian location study, MS–KSD–Bayes more faithfully captured mode proportions for well-separated modes and showed improved robustness to outliers, while being comparable to KSD–Bayes in unimodal settings in our experiments. In all experiments α was fixed across methods; any global rescaling of w can be absorbed into α , so only the *shape* of w drives the effect. When w is estimated from data (e.g., via a pilot density $p_\star$), all statements are understood conditional on the realized w .

Limitations. KSD-based posteriors can behave poorly in some regimes—e.g., slow or unstable convergence, sensitivity to the choice of kernel and bandwidth, and under-weighting of tails have been reported in the literature (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017). In particular, the distantly dissipativity assumption underlying IMQ-KSD excludes heavy-tailed models, and MS-KSD-Bayes inherits this limitation. Our experiments do not probe heavy-tailed targets or very high-dimensional, structured applications. We therefore view the present results as evidence of feasibility under light-tail settings, and we leave heavy-tail robustness and large-scale validation for future work. Beyond exponential families and KEF examples, the behavior of MS-KSD-Bayes in large, structured, or very high-dimensional models remains to be systematically evaluated.

Future work. Several directions follow naturally from our empirical findings and construction.

(i) *Identification and asymptotics.* We did not pursue formal guarantees in this paper; deriving conditions under which the diffusion-weighted loss identifies mixture proportions and establishing uniform LLNs/BvM-type local approximations (conditional on w) are natural next steps.

(ii) *Design of w .* Systematic procedures to choose the anchor $p_\star$ and the shape parameters (C_w, τ, γ) , including data-driven or hierarchical choices, would improve reliability. An ablation on anchor misspecification and sensitivity analyses for w would make the method easier to use.

(iii) *Computation at scale.* The empirical loss is a V-statistic with $O(n^2)$ cost. Mini-batch U -statistic surrogates and low-rank kernel approximations tailored to the weighted operator could reduce cost while preserving accuracy.

(iv) *Kernels and geometry.* Exploring alternative base/matrix kernels K adapted to the diffusion operator and to high dimensions may further improve mixture sensitivity and stability.

(v) *Broader models and tasks.* Applying MS-KSD-Bayes to discrete and graphical models with intractable normalizers, and to additional multimodal scientific datasets, would test robustness beyond the cases studied here.

(vi) *Calibration and evaluation.* Studying temperature selection for generalized Bayes, as well as predictive and posterior calibration diagnostics specific to diffusion-weighted losses, would help practitioners tune the method in a principled way.

Overall, MS-KSD-Bayes offers a potentially practical route to mode-aware, likelihood-free Bayesian updating. The above extensions could clarify when the approach is likely to work and how to deploy it reliably at scale.

Bibliography

Abrams, Z. B., Zucker, M., Wang, M., Asiaee Taheri, A., Abruzzo, L. V., and Coombes, K. R. (2018). Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC genomics*, 19:1–14. 103

Afzali, E. and Muthukumarana, S. (2023). Gradient-free kernel conditional stein discrepancy goodness of fit testing. *Machine Learning with Applications*, 12:100463.

- Afzali, E., Muthukumarana, S., and Wang, L. (2024). Navigating interpretability and alpha control in gf-kcsd testing with measurement error: A kernel approach. *Machine Learning with Applications*, 17:100581. 84
- Altamirano, M., Briol, F.-X., and Knoblauch, J. (2023a). Robust and conjugate gaussian process regression. *arXiv preprint arXiv:2311.00463*. 83
- Altamirano, M., Briol, F.-X., and Knoblauch, J. (2023b). Robust and scalable bayesian online changepoint detection. In *International Conference on Machine Learning*, pages 642–663. PMLR. 83
- Asiaee, A., Abrams, Z. B., Nakayiza, S., Sampath, D., and Coombes, K. R. (2020). Explaining gene expression using twenty-one micrnas. *Journal of Computational Biology*, 27(7):1157–1170. 103
- Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32. 83
- Bénard, C., Staber, B., and Da Veiga, S. (2024). Kernel stein discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization. *Advances in Neural Information Processing Systems*, 36. 82, 83, 89, 92
- Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). An overview of robust bayesian analysis. *Test*, 3(1):5–124. 87
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems.

Journal of the Royal Statistical Society: Series B (Methodological), 36(2):192–225.

82

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130. 81, 87

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR. 82, 84, 87, 91, 107

Coombes, K. R. (2019). *BimodalIndex: The Bimodality Index*. R package version 1.1.9. 103

Coombes, K. R. (2024). *oompaData: Data to Illustrate OOMPA Algorithms*. R package version 3.1.4. 103

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3):349–362. 82

Fisher, M. and Oates, C. J. (2024). Gradient-free kernel stein discrepancy. *Advances in Neural Information Processing Systems*, 36. 84

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2019). Objective bayesian inference with proper scoring rules. *Test*, 28(3):728–755. 87

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928. 83, 89, 90, 91

- Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR. 82, 84, 86, 87, 91, 107
- Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942. 98
- Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503. 87
- Huggins, J. and Mackey, L. (2018). Random feature stein discrepancies. *Advances in neural information processing systems*, 31. 98
- Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principles of bayesian inference using general divergence criteria. *Entropy*, 20(6):442. 81
- Jiang, X., Xiao, G., and Li, Q. (2022). A bayesian modified ising model for identifying spatially variable genes from spatial transcriptomics data. *Statistics in medicine*, 41(23):4647–4665. 82
- Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR. 84
- Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*. 88

- Laplante, W., Altamirano, M., Duncan, A., Knoblauch, J., and Briol, F.-X. (2025). Robust and conjugate spatio-temporal gaussian processes. *arXiv preprint arXiv:2502.02450*. 83
- Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR. 82, 84, 87, 91, 107
- Liu, S., Kanamori, T., Jitkrittum, W., and Chen, Y. (2019). Fisher efficient inference of intractable models. *Advances in Neural Information Processing Systems*, 32. 82
- Liu, X., Duncan, A. B., and Gandy, A. (2023). Using perturbation to improve goodness-of-fit tests based on kernelized stein discrepancy. In *International Conference on Machine Learning*, pages 21527–21547. PMLR. 82, 83, 89, 92
- Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022. 82, 88, 89, 98, 105
- Miller, J. W. and Dunson, D. B. (2019). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*. 87
- Moore, M., Nicholls, G., Pettitt, A., and Mengersen, K. (2020). Scalable bayesian inference for the inverse temperature of a hidden potts model. 82
- Murray, I., Ghahramani, Z., and MacKay, D. (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*. 81

- Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390. 82
- Postman, M., Huchra, J. P., and Geller, M. J. (1986). Probes of large-scale structure in the corona borealis region. *Astronomical Journal (ISSN 0004-6256)*, vol. 92, Dec. 1986, p. 1238-1247., 92:1238–1247. 100
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 103
- Roeder, K. (1990). Density estimation with confidence sets exemplified by super-clusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624. 100
- Wang, J., Wen, S., Symmans, W. F., Pusztai, L., and Coombes, K. R. (2009). The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer informatics*, 7:CIN–S2846. 103
- Wenliang, L. K. and Kanagawa, H. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*. 82, 89, 92
- Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR. 84

Chapter 5

Discussion and Conclusion

The rapid growth of complex, high-dimensional data in modern science and industry, particularly in fields like neuroscience and genomics, has encouraged the development of sophisticated statistical and machine learning models. While these models offer powerful predictive capabilities, they often come at the cost of analytical tractability. Specifically, many cutting-edge models are defined by likelihood functions and their gradients that are either computationally expensive or entirely intractable. This fundamental challenge hinders the application of classical inferential and diagnostic tools, creating a critical gap between model development and robust model validation. This thesis addresses this gap by developing, validating, and extending a suite of advanced kernel-based approaches for robust inference. The research focuses on two key areas: settings where gradient information is unavailable, and Bayesian inference problems involving complex multimodal posteriors, particularly the challenge of capturing the relative weights of well-separated modes. In this concluding chapter, the key findings of the thesis are synthesized, their significance is highlighted, and several avenues for future research are outlined.

5.1 Key Findings and Contributions

This thesis has introduced a cohesive set of methodologies that progress from foundational goodness-of-fit testing to sophisticated techniques for robust Bayesian inference, each addressing a critical gap in the analysis of intractable models.

A New Frontier for Goodness-of-Fit Testing: The journey began with the development of the Gradient-Free Kernel Conditional Stein Discrepancy (GF-KCSD), as presented in Chapter 2. Traditional KSD-based tests, while powerful, depend on the gradient of the model’s log-likelihood, which can be a prohibitive bottleneck. GF-KCSD successfully overcomes this limitation by using a surrogate density and an importance-weighting scheme, resulting in a test that is entirely independent of the target model’s gradients. This innovation not only expands the applicability of kernel-based testing to a much broader class of models but was also shown to be a competitive and powerful tool in its own right, retaining strong statistical properties and performing comparably to its gradient-based counterparts in empirical studies.

Evaluating Robustness and Practicality: Building upon this foundational contribution, the research presented in Chapter 3 rigorously examined the method’s practical utility under non-ideal, real-world conditions. A statistical method’s true value is often revealed by its performance with noisy data, a common complication that is often overlooked. Our investigation into the effects of measurement error demonstrated the remarkable robustness of GF-KCSD. Through controlled experiments with both symmetric and asymmetric error distributions, GF-KCSD consistently maintained correct Type-I error control and stable statistical power, even

as data contamination increased. This resilience was further confirmed through the analysis of a real-world brain MRI dataset, where GF-KCSD provided a reliable and computationally efficient assessment of a complex Mixture Density Network model, solidifying its standing as a trustworthy tool for applied researchers.

Overcoming mode-proportion blindness in generalized Bayes. Chapter 4 tackles the well-documented tendency of score-based generalized Bayes to ignore the relative weights of well-separated modes. We introduce *MS-KSD-Bayes*, which augments the Stein loss with a θ -independent, tail-vanishing diffusion weight while retaining the score-only, likelihood-free computation of KSD-Bayes. Concretely, in our experiments we instantiate the weight by a plug-in estimate $\tilde{p}(x)$ of the (possibly unnormalized) data density, which is bounded and vanishes in the tails, thereby attenuating long-range cross-mode interactions while preserving within-mode signal. The resulting empirical loss is a diffusion-weighted KSD V -statistic that remains likelihood-free (all computations depend on p_θ only through the data score $s_{p_\theta} = \nabla_x \log p_\theta$), and any global rescaling of w_γ is absorbed by the temperature α , so only the *shape* of w_γ matters. Empirically, MS-KSD-Bayes restores sensitivity to mixing weights on multimodal benchmarks (Galaxy velocities; a bimodal gene) and maintains robustness in unimodal settings with outlier contamination, yielding sharper posteriors than standard Bayes and KSD-Bayes under the same tuning.

5.2 Future Work and Directions

The methodologies developed in this thesis provide a strong foundation for robust inference in intractable models and naturally motivate several promising directions

for future inquiry.

Computational Scalability: A key area for future development is the enhancement of computational scalability. The kernel-based nature of the proposed methods, while avoiding gradient computations, results in a computational complexity of $O(n^2)$. To make these tools practical for very large-scale datasets, a primary focus could be the exploration of low-rank kernel approximations, such as the Nyström method or Random Fourier Features, which can reduce complexity to near-linear in the sample size. A parallel effort could involve developing specialized MCMC or variational inference algorithms that are explicitly optimized for the MS-KSD-Bayes objective function, potentially leading to significant gains in efficiency.

Selection of the Auxiliary Density and Hyperparameters: Another promising research direction lies in developing more automated and adaptive methodologies for hyperparameter selection. The performance of the methods presented here is contingent on choices such as the kernel bandwidth and, for MS-KSD-Bayes, the weighting parameter γ . While standard heuristics were effective in this work, future research could move beyond these fixed approaches. For instance, one could develop data-driven methods, such as creating a hierarchical model where the weighting parameters are learned from the data. Such an advancement would enhance the robustness and ease of use of these methods, making them more accessible to a wider range of practitioners.

Extension to Broader Data Types and Models: Furthermore, a natural and significant extension of this research involves adapting these frameworks to

broader classes of data and models. The current work has focused on continuous domains, yet many important statistical problems involve discrete data. Extending the methodologies to handle such cases, perhaps by developing a "Gradient-Free Discrete KCSD" based on a discrete Stein operator, would be highly valuable for applications in fields like bioinformatics and social network analysis. Similarly, applying these robust inference techniques to other complex settings, such as spatio-temporal models where likelihoods are notoriously intractable, could greatly broaden the applicability and impact of this research.

Deepening theoretical understanding. For *MS-KSD-Bayes*, further theory would sharpen our guarantees. First, deriving finite-sample concentration rates (non-asymptotic risk bounds) for the generalized posterior would strengthen results beyond asymptotics. Second, analyzing challenging regimes—nearly overlapping or degenerate modes, heavy tails, and misspecification of the anchor/weight—would delineate when the diffusion-weighted loss correctly identifies mixture proportions and when it may fail. Third, studying how the weight-shape parameters (γ, C_w, τ) and kernel choices affect contraction and calibration would yield practical tuning guidance. Progress on these fronts would solidify kernel-based, likelihood-free inference as a reliable tool for modern data analysis.

5.3 Concluding Remarks

The landscape of data analysis is increasingly dominated by models whose complexity outpaces the reach of traditional statistical tools. This thesis has confronted this challenge directly by developing a suite of advanced kernel-based methods de-

signed for robust and tractable inference. The work presented here has successfully delivered on its core objectives, from the creation of a novel gradient-free test for conditional models to the development of a sophisticated Bayesian framework that corrects a fundamental flaw in score-based inference. Collectively, these contributions provide a powerful and theoretically-grounded set of tools that enrich the modern statistician’s toolkit. By enabling reliable inference where it was once impractical, this research helps bridge the gap between complex modeling and rigorous validation, ultimately paving the way for more trustworthy data-driven discovery.

Appendix A

A.1 Proofs of Theoretical Results in Chapter 2

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

A.1.1 Proof of Theorem 1

Proof. First we need to show that the second expectation in Equation (2.8) is equal to zero. To do that in the first step we rewrite the expectation in the integral form $(\int |K_{\mathbf{x}}\zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot)| dp)$ 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 $C_0 = \sup \{\|g(x)\| : x \in \mathbb{R}^D\}$ and $C_1 = \sup \{|\nabla \cdot g(x)| : x \in \mathbb{R}^D\}$. 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:

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

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 $\int K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot) dp = \lim_{r \rightarrow \infty} \int 1_{B_r} K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot) dp$. Then we have:

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

Now, after this proof, we can rewrite the Equation (2.8) as follows:

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

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 $x \sim \pi_x$, $\|\mathbf{r}_{p,\pi}(\cdot, \mathbf{x})\|_{\mathcal{G}^{D_y}}^2 = 0$ if and only if $p_{\cdot|x} = \pi_{\cdot|x}$.

The rest can be proved as the previous part. $\square$

A.1.2 Proof of Proposition 2

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

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

From the requirement 4 in Theorem 1, and due to the Bochner integrability of $K_{\mathbf{x}} \zeta_{\rho(\cdot|\mathbf{x})}(\mathbf{y}, \cdot)$, 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 $K_{\mathbf{x}'}^* K_{\mathbf{x}} = K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') I$ and I is an

identity operator. Thus $b_\rho((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}'))$ can be written as:

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

□

A.2 Proofs of Theoretical Results for Chapter 4

This section provides detailed proofs for the theoretical results presented in Chapter 4. Each proof follows a structured approach, leveraging key properties of the diffusion-weighted Kernel Stein Discrepancy (KSD) and generalized Bayesian inference. For clarity, we include intermediate lemmas and derivations where necessary.

A.2.1 Proof of Proposition 4

Proof. This is the reversible specialization ($a(x) = w(x)I_d$) of the diffusion–Stein identity in Gorham et al. (2019, Sec. 2; App. A); hence the claim *holds based on* that reference. For completeness, we give a short argument under Assumption 1.

Let $\theta \in \Theta$ be arbitrary and let $g \in \mathcal{G}(k)^d$ be admissible. By the divergence-form definition,

$$(\mathcal{A}_{p_\theta}^w g)(x) = \frac{1}{p_\theta(x)} \nabla \cdot (w(x) g(x) p_\theta(x)).$$

For $R > 0$, write $B_R = \{x : \|x\| \leq R\}$ and denote by $n(x)$ the outward unit normal on ∂B_R . Applying the divergence theorem on B_R (or more generally on $\Omega_R := \mathcal{X} \cap B_R$, in which case the extra boundary integral over $\partial\mathcal{X}$ vanishes by the admissibility/boundary condition), we obtain

$$\int_{B_R} (\mathcal{A}_{p_\theta}^w g)(x) p_\theta(x) dx = \int_{B_R} \nabla \cdot (w(x) g(x) p_\theta(x)) dx = \int_{\partial B_R} n(x)^\top w(x) g(x) p_\theta(x) dS(x).$$

Assumption 1(4) ensures the boundary integral tends to 0 as $R \rightarrow \infty$. The regularity and integrability in Assumption 1(i)–(iii) (bounded $w, \nabla w$; RKHS derivatives controlled by the C^2 kernel; and $\mathbb{E}_{p_\theta} \|s_{p_\theta}(X)\| < \infty$) justify the divergence theorem and allow us to pass the limit $R \rightarrow \infty$ by dominated convergence. Therefore,

$$\mathbb{E}_{X \sim p_\theta} [(\mathcal{A}_{p_\theta}^w g)(X)] = \int_{\mathcal{X}} (\mathcal{A}_{p_\theta}^w g)(x) p_\theta(x) dx = \lim_{R \rightarrow \infty} \int_{\partial B_R} n^\top w g p_\theta dS = 0.$$

Since θ was arbitrary, the identity holds for all $\theta \in \Theta$. $\square$

A.2.2 Proof of Proposition 5

Proof. This is the reversible specialization ($a(x) = w(x)I_d$) of the diffusion–Stein RKHS representation in Gorham et al. (2019, Sec. 2; App. A); hence the result *holds based on* that reference. We record a brief argument under Assumptions 1 and 3.

Write $\mathcal{A}_{p_\theta}^{w,x}$ for the operator acting on the x –argument. By Assumption 3(2), the coordinatewise contraction

$$K_{p_\theta}^w(x, x') := (\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'}) K(x, x') \in \mathbb{R}^{d \times d}$$

is well defined and continuous. Define the linear functional $\Lambda_\theta(g) := \mathbb{E}_{X \sim \mathbb{Q}}[(\mathcal{A}_{p_\theta}^w g)(X)]$.

Let $r_x(\cdot) := (\mathcal{A}_{p_\theta}^{w,x} \otimes I) K(x, \cdot) \in \mathcal{G}(K)$. Then, by the vector-valued reproducing property, $(\mathcal{A}_{p_\theta}^w g)(x) = \langle g, r_x \rangle_{\mathcal{G}(K)}$. Cauchy–Schwarz and Assumption 3(3) give

$$|\Lambda_\theta(g)| = |\mathbb{E} \langle g, r_X \rangle_{\mathcal{G}(K)}| \leq \|g\|_{\mathcal{G}(K)} \left(\mathbb{E} \|r_X\|_{\mathcal{G}(K)}^2 \right)^{1/2} = \|g\|_{\mathcal{G}(K)} \left(\mathbb{E} \operatorname{tr}(K_{p_\theta}^w(X, X)) \right)^{1/2},$$

so Λ_θ is bounded. By the Riesz representation theorem, there exists a unique $h_\theta \in \mathcal{G}(K)$ such that $\Lambda_\theta(g) = \langle g, h_\theta \rangle_{\mathcal{G}(K)}$ for all g , and

$$\|h_\theta\|_{\mathcal{G}(K)}^2 = \mathbb{E}_{X, X' \sim \mathbb{Q}}[\operatorname{tr} K_{p_\theta}^w(X, X')].$$

Therefore,

$$\operatorname{KSD}_w(\mathbb{P}_\theta \| \mathbb{Q}) = \sup_{\|g\|_{\mathcal{G}(K)} \leq 1} |\Lambda_\theta(g)| = \|h_\theta\|_{\mathcal{G}(K)}, \quad \operatorname{KSD}_w^2(\mathbb{P}_\theta \| \mathbb{Q}) = \mathbb{E}_{X, X' \sim \mathbb{Q}}[\operatorname{tr} K_{p_\theta}^w(X, X')].$$

Replacing $\mathbb{Q}$ by the empirical measure $\mathbb{Q}_n$ yields the V–statistic in (4.11). $\square$

A.2.3 Proof of Proposition 6

Proof. Recall the diffusion Stein operator in divergence form

$$(\mathcal{A}_{p_\theta}^w g)(x) = \frac{1}{p_\theta(x)} \nabla_x (w(x) g(x) p_\theta(x)) = w(x) \{s_{p_\theta}(x)^\top g(x) + \nabla \cdot g(x)\} + \nabla w(x)^\top g(x).$$

For a matrix–valued kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{d \times d}$, define the coordinatewise contraction

$$K_{p_\theta}^w(x, x') := (\mathcal{A}_{p_\theta}^{w,x} \otimes \mathcal{A}_{p_\theta}^{w,x'}) K(x, x').$$

Let

$$\psi_\theta(x) := w(x) s_{p_\theta}(x) + \nabla w(x) \in \mathbb{R}^d, \quad \operatorname{div}_x K(x, x') := \nabla_x \cdot K(x, x') \in \mathbb{R}^d,$$

where $(\mathbf{div}_x K(x, x'))_m := \sum_{\ell=1}^d \partial_{x_\ell} K_{\ell m}(x, x')$ (columnwise divergence). We also write

$$\mathrm{tr}(\nabla_x \nabla_{x'}^\top K(x, x')) := \sum_{m=1}^d \sum_{\ell=1}^d \partial_{x_\ell} \partial_{x'_m} K_{\ell m}(x, x').$$

A standard coordinatewise application of $\mathcal{A}_{p_\theta}^w$ to K in each argument yields

$$\begin{aligned} K_{p_\theta}^w(x, x') &= \underbrace{\psi_\theta(x)^\top K(x, x') \psi_\theta(x')}_{(T1)} + \underbrace{\psi_\theta(x)^\top (\mathbf{div}_{x'} K(x, x')) w(x')}_{(T2)} \\ &+ \underbrace{\psi_\theta(x')^\top (\mathbf{div}_x K(x, x')) w(x)}_{(T3)} + \underbrace{w(x) w(x') \mathrm{tr}(\nabla_x \nabla_{x'}^\top K(x, x'))}_{(T4)}. \end{aligned}$$

Consequently, the (population) MS-KSD admits the kernel form $\mathrm{KSD}_w^2(\mathbb{P}_\theta \parallel \mathbb{Q}) = \mathbb{E}_{X, X' \sim \mathbb{Q}}[\mathrm{tr} K_{p_\theta}^w(X, X')]$, and the empirical version is the V-statistic

$$L_n(\theta) := \mathrm{KSD}_w^2(\mathbb{P}_\theta \parallel \mathbb{Q}_n) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathrm{tr} K_{p_\theta}^w(X_i, X_j).$$

Exponential family specialization. Assume the (possibly curved) exponential family

$$\log p_\theta(x) = \xi(\theta)^\top t(x) - a(\theta) + b(x),$$

with natural parameter $\xi(\theta) \in \mathbb{R}^p$, sufficient statistic $t : \mathcal{X} \rightarrow \mathbb{R}^p$, and carrier b .

The score with respect to x is

$$s_{p_\theta}(x) = \nabla_x \log p_\theta(x) = J_t(x)^\top \xi(\theta) + \nabla b(x),$$

where $J_t(x) \in \mathbb{R}^{p \times d}$ is the Jacobian of t . For each data point X_i , define

$$w_i := w(X_i), \quad A_i := w_i J_t(X_i)^\top \in \mathbb{R}^{d \times p}, \quad b_i := w_i \nabla b(X_i) + \nabla w(X_i) \in \mathbb{R}^d,$$

so that $\psi_\theta(X_i) = A_i \xi(\theta) + b_i$. Write

$$\begin{aligned} K_{ij} &:= K(X_i, X_j) \in \mathbb{R}^{d \times d}, \quad d_{ij} := \mathbf{div}_x K(X_i, X_j) \in \mathbb{R}^d, \\ d'_{ij} &:= \mathbf{div}_{x'} K(X_i, X_j) \in \mathbb{R}^d, \quad H_{ij} := \nabla_x \nabla_{x'}^\top K(X_i, X_j). \end{aligned}$$

Insert the decomposition (T1)–(T4) into $L_n(\theta)$. The only *quadratic* dependence on $\xi(\theta)$ comes from (T1):

$$\psi_\theta(X_i)^\top K_{ij} \psi_\theta(X_j) = \xi^\top A_i^\top K_{ij} A_j \xi + \xi^\top (A_i^\top K_{ij} b_j + A_j^\top K_{ij} b_i) + b_i^\top K_{ij} b_j.$$

Terms (T2) and (T3) contribute *linear* forms in ξ , and (T4) is ξ -independent.

Therefore

$$L_n(\theta) = \xi(\theta)^\top \Gamma_n \xi(\theta) + \xi(\theta)^\top \tau_n + c_n,$$

with

$$\Gamma_n := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n A_i^\top K_{ij} A_j \in \mathbb{R}^{p \times p} \quad (\text{symmetric, PSD since } K \text{ is operator-valued PD}),$$

$$\tau_n := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left\{ A_i^\top K_{ij} b_j + A_j^\top K_{ij} b_i + A_i^\top d'_{ij} w_j + A_j^\top d_{ij} w_i \right\} \in \mathbb{R}^p,$$

$$c_n := \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left\{ b_i^\top K_{ij} b_j + b_i^\top d'_{ij} w_j + b_j^\top d_{ij} w_i + w_i w_j \text{tr}(H_{ij}) \right\} \in \mathbb{R}.$$

(PSD of Γ_n : for any $v \in \mathbb{R}^p$, $v^\top \Gamma_n v = \frac{1}{n^2} \sum_{i,j} (A_i v)^\top K_{ij} (A_j v) \geq 0$ by the PD property of K .) This establishes that the empirical MS-KSD is a quadratic form in the natural parameter $\xi(\theta)$.

MS–KSD–Bayes posterior (Gaussian prior). With a Gaussian prior $\pi_0(\xi) = \mathcal{N}(\mu_0, \Sigma_0)$ on the natural parameter and generalized posterior

$$\pi_n^w(\xi) \propto \pi_0(\xi) \exp \{ -\alpha n L_n(\theta(\xi)) \},$$

completion of the square yields a Gaussian posterior

$$\Sigma_n = (\Sigma_0^{-1} + 2\alpha n \Gamma_n)^{-1}, \quad \mu_n = \Sigma_n (\Sigma_0^{-1} \mu_0 - \alpha n \tau_n),$$

which is exactly the form stated in Proposition 6. The normalizing constant depends only on c_n and is immaterial. In the *natural* exponential family case $\xi(\theta) = \theta$, this posterior is over θ . $\square$

Bibliography

- Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61. 121
- Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR. 122
- Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928. 124, 125