Statistical Model Evaluation Using Reproducing Kernels and Stein's method

Heishiro Kanagawa

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
of
University College London.

Gatsby Computational Neuroscience Unit
University College London

Declaration

I, Heishiro Kanagawa, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Advances in computing have enabled us to develop increasingly complex statistical models. However, their complexity poses challenges in their evaluation. The central theme of the thesis is addressing intractability and interpretability in model evaluations. The key tools considered in the thesis are kernel and Stein's methods: Kernel methods provide flexible means of specifying features for comparing models, and Stein's method further allows us to incorporate model structures in evaluation.

The first part of the thesis addresses the question of intractability. The focus is on latent variable models, a large class of models used in practice, including factor models, topic models for text, and hidden Markov models. The kernel Stein discrepancy (KSD), a kernel-based discrepancy, is extended to deal with this model class. Based on this extension, a statistical hypothesis test of relative goodness of fit is developed, enabling us to compare competing latent variable models that are known up to normalization.

The second part of the thesis concerns the question of interpretability with two contributed works. First, interpretable relative goodness-of-fit tests are developed using kernel-based discrepancies developed in Chwialkowski et al. [2015], Jitkrittum et al. [2016, 2017b]. These tests allow the user to choose features for comparison and discover aspects distinguishing two models. Second, a convergence property of the KSD is established. Specifically, the KSD is shown to control an integral probability metric defined by a class of polynomially growing continuous functions. In particular, this development allows us to evaluate both unnormalized statistical models and sample approximations to posterior distributions in terms of moments.

Impact statements

This thesis develops novel evaluation techniques for statistical models. A major contribution of this thesis is enabling the evaluation of complex models that were previously difficult to assess, such as latent variable models.

The techniques developed in this thesis would bring about scientific advances. Many scientific and engineering disciplines depend on accurate descriptions of observed complex phenomena; these include physics, economics, and artificial intelligence research. The developments in this thesis enable researchers to inspect and improve their models, allowing them to treat a broader range of problems.

The use of statistical models is not limited to academia. Data-based decision-making is becoming increasingly popular and is expected to permeate critical applications, such as assigning medical treatments, examining criminal evidence, or policymaking. Unfortunately, inaccurate data models could misguide our decisions and have dire consequences. Therefore, the ability to inspect statistical models is critical, and the techniques introduced in this thesis contribute to this purpose.

To my family, and to my father in loving memory.

Acknowledgements

This thesis would have not been possible without the support of many people. Before all else, I want to express my gratitude to my supervisor, Arthur Gretton, for his excellent guidance throughout this journey. I have been fortunate to work with and learn from someone who really enjoys science and has a good sense of humour. I thank the Gatsby foundation for its financial support, without which I would not have been able to experience this incredible journey. I appreciate the Gatsby community for providing an inspiring, warm, and supporting environment. I thank the administrative team for their support. Thank you to Barry Fong for being friends with me in addition to the admin job. My special thanks go to my office mates: Antonin Schrab, Dimitri Meunier, Kevin Li, Kevin Huang, Pierre Grasser, Sanjeevan Ahilan, and Wenkai Xu, for allowing me to have random chats and listening to my complaints. I thank Anna Korba, Hugo Soulat, Iryna Korshunova, Kevin Li, Virginia Rutten, and Wenkai Xu for their emotional support; I am grateful to my collaborators: Wittawat Jitkrittum for being a great senpai, teaching me about many things including life, and listening to my complaints; Lester Mackey for rigorous feedback and teaching me Stein's method; Michael Shvartsman for introducing me to the world of cognitive science; Kenji Fukumizu for his thorough feedback and hosting me at the Institute of Mathematical Statistics; Kevin Li for inviting me to work on interesting problems and research chats. I thank William Dorrell for thorough proofreading. I thank my family (especially my late father) and my then-partner for their unconditional care and support. Finally, I would like to thank the two examiners, Prof. Brooks Paige and Prof. Gesine Reinert, for agreeing to assess my thesis and fruitful discussions.

UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3):			
a) **What is the current title of the manuscript?**	A kernel Stein test for comparing latent variable models		
b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if ‘Yes’, please give a link or doi):	https://arxiv.org/abs/1907.00586		
c) **Where is the work intended to be published?** (e.g. journal names)	The Journal of the Royal Statistical Society: Series B		
d) **List the manuscript’s authors in the intended authorship order:**	Heishiro Kanagawa, Wittawat Jitkrittum, Lester Mackey, Kenji Fukumizu, Arthur Gretton		
e) **Stage of publication** (e.g. in submission):	Under review		
2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4):			
Heishiro Kanagawa constructed the propose test and wrote the paper; Wittawat Jitkrittum contributed to the experiment and proofread the paper; Lester Mackey proofread and revised the paper; Kenji Fukumizu proofread the paper; Arthur Gretton contributed to the writing and revision of the paper			
3. **In which chapter(s) of your thesis can this material be found?**			
3			
4. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):			
Candidate:	Heishiro Kanagawa	**Date:**	25 August 2022
Supervisor/ Senior Author (where appropriate):	Arhur Gretton	**Date:**	25 August 2022

UCL Research Paper Declaration Form: referencing the doctoral candidate’s own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):	
a) What is the title of the manuscript?	Informative features for model comparison
b) Please include a link to or doi for the work:	https://proceedings.neurips.cc/paper/2018/file/550a141f12de6341fba65b0ad0433500-Paper.pdf
c) Where was the work published?	In Advances in Neural Information Processing Systems, 31 (pp. 816–827).
d) Who published the work? (e.g. OUP):	Curran Associates Inc.
e) When was the work published?	2018
a) List the manuscript’s authors in the order they appear on the publication:	Wittawat Jitkrittum,Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, Arthur Gretton
f) Was the work peer reviewed?	Yes
g) Have you retained the copyright?	Yes
h) Was an earlier form of the manuscript uploaded to a preprint server? (e.g. medRxiv; if ‘Yes’, please give a link or doi):	https://doi.org/10.48550/arXiv.1810.11630
[If no, please seek permission from the relevant publisher and check the box next to the below statement]:	
□	*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*
2. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4):	
Wittawat Jitkrittum designed the test statistics, conducted experiments, and wrote the paper. Heishiro Kanagawa contributed to the design of the test statistics and the proposed test, conducted experiments, and wrote the paper. Patsorn Sangkloy conducted the experiments concerning GAN models. James Hays, Bernhard Schölkopf, and Arthur Gretton proofread and revised the paper.	
3. In which chapter(s) of your thesis can this material be found?	
5	
4. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):	

Candidate:	Heishiro Kanagawa	**Date:**	25 August 2022
Supervisor/ Senior Author (where appropriate):	Arthur Gretton	**Date:**	25 August 2022

Contents

1 Introduction 13

1.1 Challenges with modern statistical models . . . 13

1.2 Structure of the Thesis . . . 16

2 Background 17

2.1 Reproducing kernel Hilbert space . . . 17

2.2 Integral probability metrics and the maximum mean discrepancy . . . 20

2.3 Stein's method and Stein discrepancies . . . 23

3 Comparing latent variable models using the kernel Stein discrepancy – Part 1 25

3.1 Introduction . . . 25

3.2 The kernel Stein discrepancy and latent variable models . . . 27

3.2.1 Stein operators and the kernel Stein discrepancy . . . 28

3.2.2 Kernel Stein discrepancies of latent variable models . . . 31

3.3 A relative goodness-of-fit test . . . 33

3.3.1 Problem setup . . . 33

3.3.2 Estimating kernel Stein discrepancies of latent variable models . . . 34

3.3.3 Test procedure . . . 35

3.3.4 Kernel choice . . . 37

3.4 Experiments . . . 39

3.4.1 Probabilistic Principal Component Analysis . . . 40

3.4.2 Dirichlet process mixtures . . . 45

3.4.3 Latent Dirichlet Allocation . . . 46

3.5 Conclusion . . . 52

3.6 Proofs . . . 53

3.6.1 Asymptotic normality of approximate U-statistics . . . 53

3.6.2 Variance of a U-statistic . . . 55

3.A Kernel Stein discrepancy for bounded domains . . . 58

3.B Numerically stable Stein operator . . . 58

3.B.1 Univariate case . . . 58

3.B.2 Multivariate case . . . 59

3.B.3 Application to latent variable models . . . 60

3.C Integrability condition for the KSD expression . . . 61

3.D Convergence assumption in the asymptotic normality proof 64
3.E The maximum mean discrepancy relative goodness-of-fit test 66
3.F MMD and KSD for Gaussian distributions 69
3.F.1 MMD 69
3.F.2 KSD 71
3.F.3 Kernel choice and KSD: Gaussian models and data 71
3.G The score formula for Dirichlet process models 73
3.H Invariance properties of kernel Stein discrepancy 74
3.H.1 Model invariance 74
3.H.2 Coordinate-choice independence 75
3.I Arvesen's formula for the jackknife variance estimator 77
3.J Additional experiments 80
3.J.1 PPCA: type-I errors and test power 80
3.J.2 LDA 82
3.J.3 Experiment: close models and type-I errors 83
3.J.4 Experiment: identical models 85

4 Comparing latent variable models using the kernel Stein discrepancy – Part 2 87

4.1 Introduction 87
4.2 U-statistics with random kernels 88
4.3 Conclusion 90
4.A Proof of asymptotic normality of random kernel U-statistics 91
4.B Details of the delta method 92
4.C Proof of the asymptotic normality of the test statistic 95

5 Interpretable features for model comparison 97

5.1 Introduction 97
5.2 Measures of Goodness of Fit 99
5.3 Proposal: Rel-UME and Rel-FSSD Tests 101
5.4 Experiments 103
5.4.1 Illustration of Rel-UME and Rel-FSSD Power Criteria 104
5.4.2 Test Powers on Toy Problems 105
5.4.3 Informative Power Objective 107
5.4.4 Testing GAN Models 107
5.4.5 Examining GAN Training 109
5.5 Conclusion 110
5.A Optimization of Test Locations in Rel-UME and Rel-FSSD 111
5.B Trained Models for Generating Smiling and Non-Smiling Images 111
5.C Proof of Theorem 5.1 112
5.D Details of Experiment 5: Examining GAN Training 113
5.E Comparing Different GAN Models Trained on MNIST 115

6 Controlling moments with kernel Stein discrepancies 119

6.1 Introduction 119

6.2 Background 121

6.2.1 Definitions and symbols specific to this chapter 121

6.2.2 Generators of Itô diffusions and their Stein equations 122

6.2.3 Diffusion Stein operators and their Stein equations 124

6.2.4 The diffusion kernel Stein discrepancy 126

6.3 Main results 126

6.3.1 Uniform integrability and DKSD bounds 127

6.3.2 The DKSD detects non-uniform integrability 129

6.3.3 Recommended kernel choice 131

6.3.4 The DKSD detects convergence 131

6.4 Experiments 132

6.4.1 The Langevin KSD 133

6.4.2 The DKSD and heavy-tailed distributions 135

6.4.3 Failure mode: distribution mixtures with isolated components 137

6.5 Conclusion 140

6.A Proofs of main results 141

6.A.1 Characterization of pseudo-Lipschitz metrics 141

6.A.2 Uniform integrability and a DKSD lower bound 143

6.A.3 The diffusion Stein operator and zero-mean functions 149

6.A.4 DKSD upper bounds 150

6.B Auxiliary results 151

6.B.1 Results from previous work 151

6.B.2 Miscellaneous results 152

6.B.3 Results concerning approximation 159

6.B.4 Results for DKSD 163

6.B.5 Polynomial functions are pseudo-Lipschitz functions 167

7 Conclusion and further research 171

Chapter 1

Introduction

Statistical models describe processes underlying data generation and express uncertainty associated with observations. These models are used to extract meaningful information and patterns from the data. Model assessment is therefore a critical operation, as the quality of our subsequent inference depends on the accuracy of the models.

This thesis addresses the problem of evaluating statistical models. Mathematically, a statistical model is a probability distribution over the space in which data lie – the language of probability allows us to characterize uncertainty stemming from our incomplete knowledge of the data-generating process. We therefore take a probabilistic approach, where we compare probability distributions using some *metric* that quantifies closeness of distributions. We emphasize, however, that there is no single *canonical* way of evaluating statistical models, since what makes a model preferable depends on the purpose. Some application domains might have specific metrics; for example, visual fidelity might be preferred in image synthesis, where probabilistic models are fitted to datasets of images to generate new content. Our goal is not to address specific applications, but to provide general tools that apply to a broad range of statistical applications. Statistical models are also called probabilistic (generative) models, and we use these terms interchangeably in what follows.

1.1 Challenges with modern statistical models

A standard approach to specify a statistical model is using a probability density function. A model may be designed by imposing such assumptions on the data generating process, such as unobserved variables causing the observed variables [Bishop, 2006] or conditional independence relations among variables (e.g., Bayes networks and Markov random fields) [Koller and Friedman, 2009]. Discrepancy measures such as the Kullback-Leibler divergence (or likelihood evaluation) allow us to quantify the mismatch of a model using a density function, and those are standard approaches when the density of a model is available and tractable. A challenge of modern statistical models is that their densities are typically intractable, especially those describing high-dimensional complex phenomena. For example, models with latent variables have density functions defined by intractable integrals; normalization constants are intractable to compute for models that only specify dependence among variables (e.g., Markov random fields); some models are specified only by sampling procedures and do not admit explicit densities (e.g., the generative adversarial networks [Goodfellow et al., 2014], simulation-based models [Lintusaari et al., 2017, Cranmer et al., 2020]). Intractability precludes the direct application of density-based evaluation techniques, and it is vital to develop alternative approaches that accommodate such complex models.

An alternative strategy is to inspect the expectations of test functions under distributions. These test functions may be interpreted as some features of interest; differences in expected features (e.g., coordinate functions yield expected locations) reveal disagreements between two distributions. Although the represented discrepancy can be enriched using many test functions, the number of test functions must be finite for numerical implementation. Their specification also requires knowledge of the distributions in question; without prior knowledge, such manual specification could be sub-optimal and does not generally come with guarantees, therefore calling for a principled procedure.

One viable solution is the maximum mean discrepancy (MMD) [Gretton et al., 2006, 2012a]. The MMD is a powerful approach emerging from the machine learning literature that allows us to use infinitely many test functions; the MMD takes the maximum difference over a class of functions called a reproducing kernel Hilbert space (RKHS) [Aronszajn, 1950]. An RKHS is determined by a reproducing kernel, a function that measures similarity between two points. Thus, one can specify a discrepancy measure by choosing an appropriate kernel. Indeed, the reproducing kernel theory has established theoretical properties guiding kernel choice: e.g., the MMD is a metric over probability distributions for characteristic kernels [Fukumizu et al., 2004, Sriperumbudur et al., 2010] and metrizes weak convergence under certain topological and kernel conditions [Sriperumbudur et al., 2010, Simon-Gabriel et al., 2020]. Moreover, a practically appealing property of the MMD is that we can estimate it straightforwardly from samples; the estimation only requires sample evaluations of the kernel function, resulting in tractable estimators such as two-sample U-statistics [Hoeffding, 1948, Kowalski and Tu, 2007] or V-statistics [von Mises, 1947]. Consequently, the numerical tractability of the MMD allows us to treat a wide range of statistical models.

Despite its practical and theoretical advantages, the MMD's performance critically depends on the choice of the kernel function and its ability to represent features relevant to the problem at hand. There have been extensive studies on tuning kernel parameters in the context of two-sample testing [Gretton et al., 2012b, Sutherland et al., 2016, Jitkrittum et al., 2016, 2017b, Liu et al., 2020]. An alternative emerged from the studies of goodness-of-fit testing based on Stein's method, which creates a bespoke kernel for a given model. Stein's method was originally developed to obtain explicit rates of convergence to normality [Stein, 1972]. The key construction in Stein's method is a Stein operator, which characterizes a distribution so that a function if modified by the operator, has zero expectation under the target. The combination of a reproducing kernel and a Stein operator induces a model-dependent kernel function that may be considered as representing tailored features for model criticism; the resulting discrepancy is called the kernel Stein discrepancy (KSD). The utility of the KSD has been vindicated by the KSD goodness-of-fit tests [Chwialkowski et al., 2016, Liu et al., 2016, Yang et al., 2018, Fernandez et al., 2020], where a model structure in the kernel boosts power in some situations. Remarkably, we can obtain a tractable Stein operator for density models having unknown normalization constants; this feature eliminates the need for sampling from models, a demanding requirement for such intractable models. The KSD has therefore resulted in diverse applications in recent years, including parameter inference for intractable models [Barp et al., 2019, Matsubara et al., 2021], sampling [Liu and Lee, 2017, Chen et al., 2018, 2019, Riabiz et al., 2021], and sample quality checks for Monte Carlo integration [Gorham and Mackey, 2017, Huggins and Mackey, 2018].

Objectives and contributions. The primary objective of this thesis is to improve model evaluation practices, emphasizing kernel-based methods. In this regard, the contribution of this thesis is twofold.

The first is to extend the KSD's reach, addressing the question of intractability. The KSD is limited to a class of models with explicit density functions up to normalization constants. While this is a larger class than previously treated, it still excludes a great majority of models used in practice – even relatively simple models such as topic models [Blei et al., 2003] for text or hidden Markov models [Rabiner, 1989]. This thesis deals with this challenge for the class of latent variable models and proposes a hypothesis test. We consider a test for model comparison (i.e., *relative* goodness-of-fit) rather than absolute evaluation as in the previous KSD tests [Chwialkowski et al., 2016, Liu et al., 2016, Yang et al., 2018]. Relative goodness-of-fit is more relevant to models of complex phenomena, where all models are imperfect; by contrast, absolute goodness-of-fit tests are preferred for simple phenomena (e.g., testing normality).

The second contribution concerns interpretability in model evaluation. We treat the following two problems:

1. The MMD and the KSD do not yield indications of how models disagree with the data.
2. It is challenging to interpret the features corresponding to a KSD due to the modification by a Stein operator; hence, it is unclear what to conclude when the KSD is small or decays to zero.

Jitkrittum et al. [2016] and Jitkrittum et al. [2017a] studied the first question in the context of two-sample and goodness-of-fit testing, respectively. Their approach is to construct explicit (kernel-based) feature dictionaries and maximize the test power to obtain interpretable features that distinguish the model from the data. In this thesis, this approach is extended to model comparison, which enables modelers to investigate what makes two competing models differ in terms of the fit to the data. The second question has been in part addressed using Stein's method. Gorham and Mackey [2017] showed that the KSD controls the bounded-Lipschitz metric; a KSD decay may be interpreted as diminishing expected differences of bounded Lipschitz functions. In this thesis, we extend this result to functions of polynomial growth. This development enables us to interpret the KSD in terms of moments, which are fundamental quantities in data analysis (e.g., mean and variance). In particular, besides statistical model evaluation, this work also contributes to Bayesian inference, as it enables assessing the quality of sample approximations to target posterior distributions in respect of moments.

1.2 Structure of the Thesis

We start in Chapter 2 with some brief background on kernel methods and Stein's method. The next three chapters concern statistical hypothesis tests for comparing statistical models: Chapter 3 presents a test for comparing latent variable models using the KSD; Chapter 4 treats the same problem but presents a simple alternative to the one taken in Chapter 3 – we explain why this approach fails; Chapter 5 addresses the lack of interpretability with discrepancy-based model comparison approaches and presents a new hypothesis test. Finally, in Chapter 6, we address the interpretability of the KSD by investigating its implication for moment convergence.

The four main chapters are based on the works completed over the course of this thesis. Chapter 3 and 4 are based on a submitted paper

> **Kanagawa, H.**, Jitkrittum, W., Mackey, L., Fukumizu, K., & Gretton, A. (Revision under review by the Journal of the Royal Statistical Society: Series B, 2019, July). A Kernel Stein Test for Comparing Latent Variable Models. arXiv: 1907.00586

Chapter 5 is adapted from the publication

> Jitkrittum, W., **Kanagawa, H.**, Sangkloy, P., Hays, J., Schölkopf, B., & Gretton, A. (2018). Informative Features for Model Comparison. In Advances in Neural Information Processing Systems, 31 (pp. 816–827).

Chapter 6 builds largely on an unpublished ongoing work and in small part the following workshop contribution

> Wenliang, L. K. & **Kanagawa, H.** (2021, December). Blindness of score-based methods to isolated components and mixing proportions. In NeurIPS Workshop "Your Model is Wrong: Robustness and misspecification in probabilistic modeling".

Other contributions. The following published works are not included in this thesis:

1. Wenliang, L. K., Moskovitz, T., **Kanagawa, H.**, & Sahani, M. (2020, February). Amortised Learning by Wake-Sleep. In Proceedings of the 37th international conference on machine learning, ICML 2020.
2. Jitkrittum, W., **Kanagawa, H**., & Schölkopf, B. (2020, June). Testing Goodness of Fit of Conditional Density Models with Kernels. In Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020.
3. Xu, L., **Kanagawa, H.**, & Gretton, A. (2021). Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation. In Advances in Neural Information Processing Systems 35.

Chapter 2

Background

This chapter presents a brief overview of the theory of reproducing kernel Hilbert space (RKHS), integral probability metrics, and Stein's method. The material presented in this chapter will be the basis of the following chapters and assumed throughout the thesis. In the process, we introduce our notation.

2.1 Reproducing kernel Hilbert space

We briefly recall the definition and key properties of an reproducing kernel Hilbert space (RKHS). We refer the reader to Berlinet and Thomas-Agnan [2004] and Steinwart and Christmann [2008] for comprehensive treatment of the subject.

Definition 2.1 (Reproducing kernel Hilbert spaces)**.** Let $\mathcal{X}$ be a non-empty set. A reproducing kernel Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ is a real Hilbert space of functions on $\mathcal{X}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ satisfying the following condition: for any $x \in \mathcal{X}$, there exists an element $\varphi_x \in \mathcal{H}$ such that for any $f \in \mathcal{H}$,

$$f(x) = \langle f, \varphi_x \rangle_{\mathcal{H}}. \tag{2.1}$$

The property (2.1) is called the reproducing property, as the function value $f(x)$ is reproduced by the inner product with φ_x. The map $x \mapsto \varphi_x$ transforms the input x into a vector φ_x in $\mathcal{H}$; in the context of machine learning, this process is interpreted as extracting *features* relevant to the problem. Thus, the map and the RKHS are called a feature map and a feature space, respectively. The reproducing property (2.1) indicates that an RKHS function is simply a linear function of the feature vector φ_x with weight $f \in \mathcal{H}$.

Given an RKHS, we can define a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, y) := \varphi_x(y) = \langle \varphi_x, \varphi_y \rangle_{\mathcal{H}}.$$

The function k is called the reproducing kernel of the RKHS $\mathcal{H}$. From this definition, it follows that a reproducing kernel k satisfies the following conditions: (a) $k(x, y) = k(y, x)$ for any

$x, y \in \mathcal{X}$, and (b) for any $n \geq 1$ and $\{a_1, \ldots, a_n\} \subset \mathbb{R}$,

$$\sum_{i=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

When a real-valued function on $\mathcal{X} \times \mathcal{X}$ satisfies these conditions, it is called a positive semi-definite kernel. In what follows, following the literature, a positive semi-definite kernel is simply called positive definite. We have seen that an RKHS $\mathcal{H}$ defines a positive definite kernel. Conversely, the Moore–Aronszajn theorem states that any positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ admits a unique RKHS [Aronszajn, 1950] – we may therefore specify an RKHS by choosing a positive definite kernel k. Because of this correspondence, we denote the RKHS of a positive definite kernel k by $\mathcal{H}_k$ when we emphasize the dependency on the kernel. From the above feature viewpoint, this implies that we can specify a feature map $x \mapsto \varphi_x$ using a positive definite kernel $k(x, \cdot)$; this feature map is called the *canonical feature map* [Steinwart and Christmann, 2008, Lemma 4.19]. Although the explicit computation of $\langle \varphi_x, \varphi_y \rangle_{\mathcal{H}}$ seems intractable if $\mathcal{H}$ is high-dimensional, we can perform this operation simply by computing a positive kernel. Thus, a positive definite kernel allows us to consider rich, nonlinear features φ_x, while their similarities can be measured tractably. This view proved very useful in machine learning, where flexible, nonlinear algorithms were created by considering classical linear counterparts (e.g., principal component analysis, and linear regression) in RKHSs [see, e.g., Hofmann et al., 2008, for a review].

To illustrate the concepts introduced above, let us consider a homogeneous polynomial kernel on $\mathbb{R}^D \times \mathbb{R}^D$

$$\begin{aligned} k(x, y) &= \langle x, y \rangle^p \\ &= \sum_{(i_1, \ldots, i_p) \in \{1, \ldots, D\}^p} x_{i_1} x_{i_2} \cdots x_{i_p} \cdot y_{i_1} y_{i_2} \cdots y_{i_p} \end{aligned}$$

with x_i denoting the i-th entry of x, $\langle x, y \rangle = x^\top y := \sum_{i=1}^{D} x_i y_i$, and integer $p \geq 1$. In this case, the corresponding feature map φ_x maps x to the concatenation of all ordered products of the entries of x [Poggio, 1975, Hoffman et al., 2010]. The feature represents interactions between coordinates, which is absent in the linear case ($p = 1$). The explicit computation of the feature map is prohibitive in higher dimensions, while the complexity of the kernel remains linear in the dimension D. This example shows that a reproducing kernel allow us to obtain rich features while their similarity can be measured tractably.

A hallmark of RKHSs is that we can construct a function space with desired properties by designing an appropriate kernel. This point may be informally described as follows: since a function in an RKHS is a linear function of the canonical feature map $k(x, \cdot)$, it inherits the properties (preserved under linear transformations) of the kernel. In the following, we review kernel properties and their implications for RKHS functions.

Growth. We begin with conditions to characterize the growth rate of functions in an RKHS $\mathcal{H}_k$. Any function $f \in \mathcal{H}_k$ satisfies, by the reproducing property and the Cauchy-Schwarz inequality, the following relation:

$$\begin{aligned}|f(x)| &\le \|f\|_{\mathcal{H}_k}\|k(x,\cdot)\|_{\mathcal{H}_k}\\ &= \|f\|_{\mathcal{H}_k}\sqrt{k(x,x)},\end{aligned}$$

where $\|f\|_{\mathcal{H}_k} = \sqrt{\langle f, f\rangle_{\mathcal{H}_k}}$. The inequality above shows that any function in a closed ball $\mathcal{B}_R(\mathcal{H}_k) \coloneqq \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le R\}$ grows at the rate at most of $R\sqrt{k(x,x)}$.

Differentiability. Next, we look at the differentiability of RKHS functions. Let $\mathcal{X}$ be an open subset of $\mathbb{R}^D$ ($D \ge 1$) and k be a kernel on $\mathcal{X} \times \mathcal{X}$. Before we present the characterization, we set up our notation. Let ∂_d denote the partial derivative operator with respect to the d-th coordinate. Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)$ be a multi-index where each element is a non-negative integer, and $|\boldsymbol{\alpha}| \coloneqq \sum_{d=1}^{D} \alpha_d = \alpha \ge 0$. We write $\partial^{\boldsymbol{\alpha}} = \partial_1^{\alpha_1} \cdots \partial_D^{\alpha_D}$; analogously, we define an differential operator $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}}$acting on a kernel function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ on to be $\partial_1^{\alpha_1} \cdots \partial_D^{\alpha_D} \partial_{1+D}^{\alpha_1} \cdots \partial_{2D}^{\alpha_D}$ by regarding k as a function on $\mathbb{R}^{2D}$, For $m \ge 0$, we define a kernel k to be m-times continuously differentiable if $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k$ exists and continuous for all multi-indices $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \le m$ [Steinwart and Christmann, 2008, Definition 4.35]. We denote by $C^{(m,m)}$ the set of m-times continuously differentiable kernels. With these definitions, the differentiability of RKHS functions is summarized as follows:

Lemma 2.2 (Corollary 4.36 of Steinwart and Christmann [2008])**.** *Let $\mathcal{X} \subset \mathbb{R}^D$ be an open set, $m \ge 0$, and k be m-times continuously differentiable kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_k$. Then, every $f \in \mathcal{H}_k$ is m-times continuously differentiable, and for $x \in \mathcal{X}$ and a multi-index* $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d, \dots \alpha_D)$ with each α_d negative integer and $|\boldsymbol{\alpha}| \le m$, we have

$$|\partial^{\boldsymbol{\alpha}} f(x)| \le \|f\|_{\mathcal{H}_k}\sqrt{\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k(x,x)}.$$

Universality. Function approximation is a ubiquitous task in statistics and machine learning. Universality is a concept describing the approximation capacity of a given function class. For our purposes, we present the notion of C_0-universality. Let $\mathcal{X}$ be a locally compact Hausdorff space (such as $\mathbb{R}^D$) and C_0 be the space of real-valued functions vanishing at infinity equipped with the uniform norm $\|f\|_\infty = \sup_{x\in\mathcal{X}}|f(x)|$; A function f is said to vanish at infinity if for each $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathcal{X}$ such that $|f(x)| \le \varepsilon$ for any $x \in \mathcal{X} \setminus K_\varepsilon$. An RKHS $\mathcal{H} \subset C_0$ is called C_0-universal if it is dense in C_0 with respect to the uniform norm; i.e., for each $f \in C_0$ and $\varepsilon > 0$, there exists a function $f_\varepsilon \in \mathcal{H}$ such that $\|f - f_\varepsilon\|_\infty \le \varepsilon$ [Carmeli et al., 2010, Sriperumbudur et al., 2011]. The definition of C_0-universality is known to be equivalent to other concepts [see Sriperumbudur et al., 2010, for a review]. The first of these is L^p-universality [Carmeli et al., 2010, Theorem 4.1], i.e., the density of $\mathcal{H}$ in $L^p(\mathcal{X},\mu)$ with respect to the p-norm $\|f\|_{L^p(\mathcal{X},\mu)} = \left(\int |f(x)|^p \mathrm{d}\mu(x)\right)^{1/p}$ for all Borel probability measures μ and some $p \in (1,\infty]$. Here, $L^p(\mathcal{X},\mu)$ is the Banach space of p-integrable μ-measurable functions [Steinwart and Christmann, 2008]; we sometimes omit the space $\mathcal{X}$ and write $L^p(\mu)$ if it is clear from the context. The second equivalent concept is the integrally strictly positive definiteness of the kernel [Sriperumbudur et al., 2010]. An kernel is said to be ISPD if it is measurable and $\int k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) > 0$ for any non-zero finite signed measure μ.

Expectation and kernel mean embedding. We finally show that the expectation of a function in an RKHS can be characterized by a tool called kernel mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]. Kernel mean embedding is a technique to represent a probability distribution in an RKHS. Formally, let P be a probability measure on $\mathcal{X}$, and k be a positive definite kernel. The kernel mean embedding μ_P of P is defined as the expectation of the canonical feature map

$$\mu_P(\cdot) := \int k(x,\cdot)\mathrm{d}P(x) = \mathbb{E}_{X\sim P}[k(X,\cdot)].$$

Let $\mathcal{P}_{\sqrt{k}} = \{P \in \mathcal{P} : \mathbb{E}_{X\sim P}\big[\sqrt{k(X,X)}\big] < \infty\}$ with $\mathcal{P}$ the set of all probability measures. If $P \in \mathcal{P}_{\sqrt{k}}$, then the kernel mean embedding μ_P exists and belongs to the RKHS $\mathcal{H}_k$ [Fukumizu et al., 2004, Sriperumbudur et al., 2010]. In particular, for any $f \in \mathcal{H}_k$, we have

$$\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mu_P\rangle_{\mathcal{H}_k}.$$

Intuitively, since an RKHS function is a linear function of the canonical feature, taking the inner product between its weight f and the expected feature μ_P yields the expectation of the function.

According to Berlinet and Thomas-Agnan [2004, p. 189], the study of kernel mean embedding was originated by Denis Bosq and C. Guilbart. Mean embedding allows us to manipulate probability distributions using various vector operations resulting from the Hilbert space structure; e.g., we can quantify the similarity between probability distributions using the norm, as introduced in the next section. The technique has resulted in a wide range of applications such as two-sample testing [Gretton et al., 2006, 2012a], independence testing [Gretton et al., 2007], and nonparametric Bayesian inference [Fukumizu et al., 2013]; see the survey by Muandet et al. [2017] for other applications.

2.2 Integral probability metrics and the maximum mean discrepancy

Various statistical tasks can be formulated using a distance over probability distributions, including parameter inference and hypothesis testing. Given a family of probability measures $\{P_\theta\}_{\theta\in\Theta}$, the task of parameter inference is to choose an appropriate parameter θ_n given a sample $\{x_1,\ldots,x_n\}$. If the sample is generated according to a law R, an appropriate choice might be the closest one to R in some distance $d(P_\theta, R)$. A distance can also be used to specify a statistical hypothesis. For example, the problem of goodness-of-fit testing is to test the hypothesis that a model P is equal to the unknown distribution R underlying an observed sample; this hypothesis may be equally written as $H_0 : d(P,R) = 0$ for some distance d powerful enough to distinguish any distributions.

One practical class of distance is integral probability metrics [IPMs, Müller, 1997]. For a set

$\mathcal{F}$ of real-valued measurable functions on a measure space $\mathcal{X}$, the IPM between two probability distributions P, Q on $\mathcal{X}$ is defined as the worst-case difference of integrals:

$$d_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right| .$$

An IPM is an intuitive summary of the difference between two distributions: if we think of $\mathcal{F}$ as a set of features of interest, then the IPM characterizes expected disagreements in those features and summarizes them by the worst-case error. Note that an IPM is in general a pseudo metric and becomes a metric if and only if the function class $\mathcal{F}$ separates two distributions, i.e., $\mathbb{E}_{X \sim P}[f(X)] \neq \mathbb{E}_{Y \sim Q}[f(Y)]$ for some $f \in \mathcal{F}$ [Müller, 1997]. As in goodness-of-fit testing mentioned above, ensuring the separability is of theoretical interest. Following are examples of function classes ensuring this condition:

1. $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$, where $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$. This choice yields the total variation metric.
2. $\mathcal{F} = \{1_{(-\infty, t]} : t \in \mathbb{R}\}$, where $\mathcal{X} = \mathbb{R}$ and 1_A denotes the indicator function of a set A. The corresponding distance $d_{\mathcal{F}}$ is called the Kolmogorov metric . The Kolmogorov metric characterizes the maximum discrepancy between cumulative distribution functions and is used for the Kolmogorov-Smirnov test [Kolmogorov, 1933, Smirnov, 1948]
3. $\mathcal{F} = \{f : \|f\|_\infty + \|f\|_{\mathrm{L}} \leq 1\}$, where $\|f\|_{\mathrm{L}} := \sup_{x \neq y} |f(x) - f(y)| / \tilde{d}(x, y)$ and $\tilde{d}$ is a metric on $\mathcal{X}$ (so that $\mathcal{X}$ is a metric space). The $d_{\mathcal{F}}$ is known as the bounded-Lipschitz metric (or the Dudley metric) [Dudley, 2002, Chapter 11]. The Dudley metric is known to metrize weak convergence (or narrow convergence) [Dudley, 2002, Section 11.3]. Here, the weak convergence of a sequence of probability measures $\{P_1, P_2, \dots\}$ is defined as having $\int f \mathrm{d}P_n \to \int f \mathrm{d}P$ for any continuous bounded function f.

Despite well-understood theoretical properties, not all IPMs are suitable to statistical applications. IPMs may not admit computable forms due to the optimization formulation and may also be challenging to estimate [Sriperumbudur et al., 2010]. For example, for two distributions on $\mathbb{R}^D$, the Dudley metric above may be estimated using samples; however, the convergence rate depends on D and can be slow for a large D [Sriperumbudur et al., 2012].

An IPM can be constructed using an RKHS to overcome these challenges. The maximum mean discrepancy (MMD) [Gretton et al., 2006, 2012a] is an IPM defined by the unit ball in an RKHS $\mathcal{H}_k$:

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right| .$$

An attractive property of the MMD is that it admits a closed-form expression. If $P, Q \in \mathcal{P}_{\sqrt{k}}$, one can show that the MMD is given by their mean embeddings: the supremum is attained by a function $f^* \propto \mu_P - \mu_Q$, yielding

$$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k} . \tag{2.2}$$

The optimal function f^* is called the witness function [Gretton et al., 2012a, Section 2.3]. The

MMD uses the RKHS norm to measure the departure of the witness f^* from zero. A different norm gives rise to a different discrepancy [Chwialkowski et al., 2015, Jitkrittum et al., 2016]; we will see this in Chapter 5.

A distinctive property of the MMD is that it admits tractable and well-studied estimators. The expression (2.2) leads to a closed-form solution involving kernel expectations [Gretton et al., 2012a, Lemma 6]

$$\begin{aligned}\mathrm{MMD}^2(P, Q) = \mathbb{E}_{X,X' \sim P \otimes P}[k(X, X')] + \mathbb{E}_{Y,Y' \sim Q \otimes Q}[k(Y, Y')] \\ - 2\mathbb{E}_{X,Y \sim P \otimes Q}[k(X, Y)]\end{aligned}, \tag{2.3}$$

where $X \otimes Y \sim P_1 \otimes P_2$ means that X and Y are independent and $X \sim P_1$, $Y \sim P_2$. Given mutually independent samples $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and $\{y_i\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} Q$, we can estimate the squared MMD with a two-sample U-statistic [Hoeffding, 1948, Kowalski and Tu, 2007, p. 131]

$$\widehat{\mathrm{MMD}^2}(P, R) = \frac{1}{\binom{n}{2}} \frac{1}{\binom{m}{2}} \sum_{j_1 < j_2} \sum_{i_1 < i_2} \ell(x_{i_1}, x_{i_1}; y_{j_1}, y_{j_2})$$

with

$$\ell(x, x'; y, y') = k(x, x') + k(y, y') - \frac{1}{2}\{k(x, y) + k(x, y') + k(x', y) + k(x', y')\}.$$

Note that this statistic is equal to the unbiased estimator of Gretton et al. [2012a, Eq. 3]. Alternatively, one can estimate the MMD using a V-statistic [von Mises, 1947]

$$\mathrm{MMD}^2(P_n, Q_m) = \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(y_i, y_j),$$

where $P_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$, $Q_m = m^{-1} \sum_{j=1}^m \delta_{y_j}$, and δ_x denotes the Dirac measure having unit mass at x. Note that these two estimators are asymptotically equivalent. In fact, when the kernel is bounded or translation invariant (i.e., $k(x, y) = \phi(x - y)$ for some function ϕ), then their difference decays at a rate of $O(m^{-1} + n^{-1})$. Moreover, it is known that the V-statistic estimator achieves $\sqrt{mn/(m + n)}$-consistency at a rate independent of D for bounded kernels [Sriperumbudur et al., 2012, Corollary 3.5].

The MMD is a metric on probability measures if the kernel function is *characteristic*. A kernel k is characteristic if and only if the mean map $P \mapsto \mu_P(\cdot) = \mathbb{E}_{X \sim P}[k(X, \cdot)]$ is injective [Fukumizu et al., 2004, Sriperumbudur et al., 2010]. It is easy to see that the MMD's separability follows from the expression (2.2). Examples of characteristic kernels are the exponentiated quadratic kernel $k(x, y) = \exp\left(-\|x - y\|_2^2 / 2\lambda^2\right)$ for any $\lambda > 0$ and the Matérn class kernels [Matérn, 1986, Stein, 1999]; see Sriperumbudur et al. [2010] for other examples. It can intuitively be understood that the RKHS defined by a characteristic kernel is rich enough to distinguish any two distributions. An obvious example of non-characteristic kernels is the linear kernel $k(x, y) = \langle x, y \rangle$, as the mean embedding of P simply becomes the mean of P; in this case, the MMD is only informative of the mean difference (the RKHS consists of linear functions). A concept related to the richness of an RKHS is that of universality introduced in the previous section. The relation between characteristic and universal kernels has been investigated by Sriperumbudur et al. [2011]. For example, for bounded continuous translation-invariant kernels on $\mathbb{R}^D$ (such as the aforementioned two characteristic kernels), C_0-universality and characteristicness are equivalent [Sriperumbudur et al., 2011, Section 4.3].

2.3 Stein's method and Stein discrepancies

Stein's method is a technique to compare distributions, introduced in the seminal paper by Stein [1972]. Stein's method provides a characterization of probability distributions and enables us to upper bound an IPM. This section serves as a brief introduction to the subject. We refer the reader to the expository papers by Ross [2011] and Anastasiou et al. [2021] for more detailed descriptions of the technique; the latter reference also provides an overview of applications in computational statistics.

A starting point of Stein's method is identifying an operator characterizing a probability distribution. Formally, for a distribution P on $\mathcal{X}$, let $\mathcal{T}_P$ be a linear operator that acts on a set $\mathcal{G}(\mathcal{T}_P)$ of functions on $\mathcal{X}$ such that

$$\mathbb{E}_{X\sim P}\big[\mathcal{T}_P g(X)\big] = 0 \text{ for each } g \in \mathcal{G}(\mathcal{T}_P). \tag{2.4}$$

Such an operator $\mathcal{T}_P$ and a set $\mathcal{G}(\mathcal{T}_P)$ are respectively called a *Stein operator* and a *Stein set*; the identity of the form (2.4) is known as Stein's identity. For simplicity, we assume that $\mathcal{T}_P g$ is a real-valued function in the following.

Using a Stein operator of P, one can measure the dissimilarity between Q and P by examining the deviation of the expectation $\mathbb{E}_{X\sim Q}\big[\mathcal{T}_P g(X)\big]$ from zero, where $g \in \mathcal{G}(\mathcal{T}_P)$. Following this idea, for any subset $\mathcal{G} \subset \mathcal{G}(\mathcal{T}_P)$, one can construct a discrepancy measure

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}) = \sup_{g\in\mathcal{G}} \left|\mathbb{E}_{X\sim Q}\big[\mathcal{T}_P g(X)\big]\right|. \tag{2.5}$$

This worst-case discrepancy measure is called a *Stein discrepancy*, introduced by Gorham and Mackey [2015] (the term usage differs from Ledoux et al. [2015]). Remarkably, by choosing an appropriate Stein operator and a Stein set, one can construct a computable Stein discrepancy [Gorham and Mackey, 2015, Chwialkowski et al., 2016, Liu et al., 2016, Oates et al., 2017, Gorham et al., 2019].

This thesis builds on a class of computable Stein discrepancies, the *kernel Stein discrepancy* (KSD) [Chwialkowski et al., 2016, Liu et al., 2016, Oates et al., 2017]. This section aims to provide an overview of the idea and thus only describes an informal description of the KSD; see Chapter 3 for its rigorous construction. The KSD is an instance of the MMD; it is similarly constructed by specifying the Stein set $\mathcal{G}$ to be the unit ball $\mathcal{B}_1(\mathcal{H}_k)$ of an RKHS. As we have seen in Section 2.1, an RKHS function g is a linear function of the canonical feature $\varphi_x = k(x, \cdot)$. As a Stein operator $\mathcal{T}_P$ is a linear operator, we informally obtain

$$\mathcal{T}_P g(x) = \mathcal{T}_P\langle g, \varphi_x\rangle = \langle g, \mathcal{T}_P\varphi_x\rangle.$$

This relation implies that the Stein-modified function $\mathcal{T}_P g(x)$ is a linear function of the modified feature $\mathcal{T}_P \varphi_x$. This feature induces a new kernel function,

$$h_P(x, y) = \langle \mathcal{T}_P \varphi_x, \mathcal{T}_P \varphi_y \rangle,$$

and $\mathcal{T}_P g(x)$ is an element of the RKHS determined by h_P. In particular, we have $\mathbb{E}_{X\sim P}[h_P(X, y)] = 0$ for each $y \in \mathcal{X}$, and that all elements of the RKHS has zero-mean under the target P (under appropriate conditions). This viewpoint turns the KSD $\mathcal{S}\big(Q, \mathcal{T}_P, \mathcal{B}_1(\mathcal{H}_k)\big)$ into a special MMD defined by kernel h_P, resulting in a closed-form expression:

$$\mathrm{KSD}\,(P\|Q)^2 = \mathbb{E}_{X,X'\sim Q\otimes Q}\big[h_P(X, X')\big].$$

As with the MMD, if h_P is possible to evaluate, the KSD admits tractable estimators. In contrast to the MMD, the KSD does not involve integrals with respect to the target P. This feature is particularly useful when sampling from P is challenging; an example of this situation is where the target is defined by a density with an unknown normalizing constant. Chapter 3 introduces a KSD that deals with this class of distributions.

Stein discrepancies and IPMs are closely related. The key to connecting these is the Stein equation

$$\mathcal{T}_P g_f = f - \mathbb{E}_{X\sim P}[f(X)], \tag{2.6}$$

where f is a function of interest, and $g_f \in \mathcal{G}(\mathcal{T}_P)$ is a solution to the Stein equation (2.6). The existence of a solution depends on the properties of the test function f and the operator $\mathcal{T}_P$. If we can take expectations, then we obtain

$$\mathbb{E}_{Y\sim Q}[f(Y)] - \mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}\big[\mathcal{T}_P g_f(Y)\big].$$

For a function class $\mathcal{F}$, this relation yields

$$\sup_{f\in\mathcal{F}} \left|\mathbb{E}_{Y\sim Q}[f(Y)] - \mathbb{E}_{X\sim P}[f(X)]\right| = \sup_{g\in\mathcal{G}_{\mathcal{F}}} \left|\mathbb{E}_{Y\sim Q}\big[\mathcal{T}_P g(Y)\big]\right|, \tag{2.7}$$

where $\mathcal{G}_{\mathcal{F}} = \{g_f : f \in \mathcal{F}\}$ is the set of solutions to the Stein equation. The key point of Stein's method is that the study of the IPM $d_{\mathcal{F}}(P, Q)$ can be reduced to the evaluation of the Stein discrepancy on the RHS of (2.7). The evaluation is typically performed by upper bounding the Stein discrepancy, as a solution g_f to the Stein equation is often not explicit and only known up to some regularity properties. By Stein's method, in Chapter 6, we obtain a upper bound on the IPM defined by a class of pseudo-Lipschitz functions; the upper bound is given by a kernel Stein discrepancy.

Chapter 3

Comparing latent variable models using the kernel Stein discrepancy – Part 1

Summary We propose a kernel-based nonparametric test of relative goodness of fit, where the goal is to compare two models, both of which may have unobserved latent variables, such that the marginal distribution of the observed variables is intractable. The proposed test generalizes the recently proposed kernel Stein discrepancy (KSD) tests [Chwialkowski et al., 2016, Liu et al., 2016, Yang et al., 2018] to the case of latent variable models, a much more general class than the fully observed models treated previously. The new test, with a properly calibrated threshold, has a well-controlled type-I error. In the case of certain models with low-dimensional latent structure and high-dimensional observations, our test significantly outperforms the relative Maximum Mean Discrepancy test, which is based on samples from the models and does not exploit the latent structure.

3.1 Introduction

A major approach to statistical modeling is the use of variables representing quantities that are unobserved but thought to underlie the observed data: well-known instances include probabilistic PCA [Roweis, 1997, Tipping and Bishop, 1999], factor analysis [see, e.g., Basilevsky, 1994], mixture models [see, e.g., Gilks et al., 1995], topic models for text [Blei et al., 2003], and hidden Markov models (HMMs) [Rabiner, 1989]. The hidden structure in these generative models serves multiple purposes: it allows interpretability and understanding of model features (e.g., the topic proportions in a latent Dirichlet allocation (LDA) model of text), and it facilitates modeling by leveraging simple low-dimensional dynamics of phenomena observed in high dimensions (e.g., HMMs with a low dimensional hidden state). Statistical modelers ultimately use such models to reason about the data; to guarantee the validity of the inference, modelers desire to choose accurate models and therefore are in need of model diagnostics.

This chapter addresses the problem of evaluating and comparing generative probabilistic models, in cases where the models have a latent variable structure, and the marginals over the observed data are intractable. In this scenario, one strategy for evaluating a generative model is to draw samples from it and to compare these samples to the modeled data using a two-sample test: for instance, Lloyd and Ghahramani [2015] use a test based on the maximum mean discrepancy (MMD) [Gretton et al., 2012a]. This approach has two disadvantages, however: it is not computationally efficient due to the sampling step, and it does not take advantage of the information that the model supplies, for instance the dependence relations among the variables.

Recently, an alternative model evaluation strategy based on Stein's method [Stein, 1972, Chen, 1975, Stein, 1986, Barbour, 1988, Götze, 1991] has been proposed, which directly employs a closed-form expression for the unnormalized model. Stein's method is a technique from probability theory developed to prove central limit theorems with explicit rates of convergence [see, e.g., Ross, 2011]. The core of Stein's method is that it characterizes a distribution with a *Stein operator*, which, when applied to a function, causes the expectation of the function to be zero under the distribution. For our purposes, we will use the result that a model-specific Stein operator may be defined, to construct a measure of the model's discrepancy. Notably, Stein operators may be obtained without computing the normalizing constant.

Stein operators have been used to design integral probability metrics (IPMs) [Müller, 1997] to test the goodness of fit of models. IPMs specify a *witness function* which has a large difference in expectation under the sample and model, thereby revealing the difference between the two. When a Stein operator is applied to the IPM function class, the expectation under the model is zero, leaving only the expectation under the sample. A Stein-modified $W^{2,\infty}$ Sobolev ball was used as the witness function class in [Gorham and Mackey, 2015, Gorham et al., 2019]. Subsequent work in [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] used as the witness function class a Stein-transformed reproducing kernel Hilbert ball, as introduced by Oates et al. [2017]: the resulting goodness-of-fit statistic is known as the kernel Stein discrepancy (KSD). Conditions for using the KSD in convergence detection were obtained by Gorham and Mackey [2017]. While the foregoing work applies in continuous domains, the approach may also be used for models on a finite domain, where Stein operators [Ranganath et al., 2016, Yang et al., 2018, Bresler and Nagaraj, 2019, Reinert and Ross, 2019, Hodgkinson et al., 2020, Shi et al., 2022] and associated goodness-of-fit tests [Yang et al., 2018] have been established. Note that it is also possible to use Stein operators to construct feature dictionaries for comparing models, rather than using an IPM: examples include a test based on Stein features constructed in the sample space so as to maximize test power [Jitkrittum et al., 2017b, 2018] and a test based on Stein-transformed random features [Huggins and Mackey, 2018]. While the aforementioned tests address simple hypotheses, composite tests that use Stein characterizations have been proposed for specific parametric families including gamma [Henze et al., 2012, Betsch and Ebner, 2019c] and normal distributions [Betsch and Ebner, 2019b, Henze and Visagie, 2019], and general univariate parametric families [Betsch and Ebner, 2019a] (note that these tests are not based on IPMs).

While an absolute test of goodness of fit may be desirable for models of simple phenomena, it will often be the case that in complex domains, no model will fit the data perfectly. In this setting, it is more constructive to ask which model fits better, either within a class of models or in comparing different model classes. A likelihood ratio test would be an alternative choice for this task, since it is the uniformly most powerful test [Lehmann and Romano, 2005], but this would require the normalizing constants for both models. A purely sample-based relative goodness of fit test was proposed by Bounliphone et al. [2016], based on comparing maximum mean discrepancies between the samples from two rival models with a reference real-world sample. A second relative test was proposed by Jitkrittum et al. [2018], generalizing Jitkrittum et al. [2017b] and learning the Stein features for which each model outperforms the other.

A major limitation of the foregoing Stein tests is that they all require the likelihood in closed form, up to normalization: if latent variables are present, they must be explicitly marginalized out. While certain previous works on Stein's method for model comparison did account for the presence of latent variables, they did so by explicitly marginalizing over these variables in closed form. Two examples are the Gaussian mixtures studied by Gorham et al. [2019] and the Gaussian-Bernoulli restricted Boltzmann machine studied by Liu et al. [2016, Section 6], where there are a small number of hidden binary variables. In many cases of interest, this closed-form marginalization is not possible.

In the present work, we introduce a relative goodness-of-fit test for latent variable models, which does not require exact evaluation of the unnormalized observed-data marginals. Our test compares models by computing approximate kernel Stein discrepancies, where we represent the distributions over the latent variables by samples. Our approach differs from Bayesian model selection [Jeffreys, 1961, Schwarz, 1978, Kass and Raftery, 1995, Watanabe, 2013] in which posterior odds (or Bayes factors) are reported. As in our proposed test, these quantities can be computed using Monte Carlo techniques [see, e.g., Friel and Wyse, 2012, for a review], but they do not come with calibrated thresholds to control false rejection rates. Our interest is in the fit of models, measured in kernel Stein discrepancy, and in the associated frequentist test of relative goodness of fit. Additionally, in contrast to the aforementioned quantities, our discrepancy measure does not require the likelihood function to be normalized (see Section 3.3).

We recall the Stein operator and kernel Stein discrepancy in Section 3.2, and the notion of relative tests in Section 3.3. Our main theoretical contributions, also in Section 3.3, are two-fold: first, we derive an appropriate test threshold to account for the randomness in the test statistic caused by sampling the latent variables. Second, we provide guarantees that the resulting test has the correct Type-I level (i.e., that the rate of false positives is properly controlled) and that the test is consistent under the alternative: the number of false negatives drops to zero as we observe more data. Finally, in Section 3.4, we demonstrate our relative test of goodness-of-fit on a variety of latent variable models. Our main point of comparison is the relative MMD test [Bounliphone et al., 2016], where we sample from each model. We demonstrate that the relative Stein test outperforms the relative MMD test in the particular case where the low dimensional structure of the latent variables can be exploited.

3.2 The kernel Stein discrepancy and latent variable models

In this section, we recall the definition of the Stein operator as used in goodness-of-fit testing, as well as the kernel Stein discrepancy, a measure of goodness-of-fit based on this operator. We will then introduce latent variable models, which will bring us to the setting of relative goodness of fit with competing models in Section 3.3.

Before proceeding, we call attention to our setting: in this chapter, we treat both continuous- and discrete-valued observations, as formally defined at the outset of Section 3.2.1. It is our intention to study these two data modalities as they admit the same treatment. The subsequent definitions and analysis of our test are independent of whether a continuous or discrete Stein operator is used, besides in experiments concerning discrete-valued observations. Thus, the detail about discrete models in Section 3.2.1 may be initially skipped if desired.

3.2.1 Stein operators and the kernel Stein discrepancy

Let $\mathcal{X}$ be the space in which the data takes values; for $D \geq 1$, the space $\mathcal{X}$ is either the Euclidean space $\mathbb{R}^D$ or a finite lattice $\{0, \ldots, L-1\}^D$ for some $L > 1$. Depending on $\mathcal{X}$, we shall assume that the densities below are all defined with respect to the Lebesgue measure or the counting measure; i.e., the term *density* includes probability mass functions (pmfs).

Continuous-valued observations. Suppose that we are given data $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$ from an unknown distribution R, and we wish to test the goodness of fit of a model P. We first consider the case where the probability distributions P, R are defined on $\mathbb{R}^D$ and have respective probability densities p, r, where all density functions considered in this chapter are assumed strictly positive and continuously differentiable. We treat the case of densities defined on bounded domains in the supplement, Section 3.A. For differentiable density functions, we define the *score function*,

$$\mathbf{s}_p(x) \in \mathbb{R}^D := \frac{\nabla p(x)}{p(x)} = \nabla \log p(x),$$

where the gradient operator is $\nabla := [\partial_1, \ldots, \partial_D]^\top$. The score is independent of the normalizing constant for p, making it computable even when p is known only up to normalization. Using this score, we define the *Langevin Stein operator* on a space $\mathcal{F}$ of differentiable functions from $\mathbb{R}^D$ to $\mathbb{R}^D$ [Gorham and Mackey, 2015, Oates et al., 2017],

$$[\mathcal{T}_P f](x) = \langle \mathbf{s}_p(x), f(x) \rangle + \langle \nabla, f(x) \rangle, \quad f \in \mathcal{F}.$$

A kernel discrepancy may be defined based on the Stein operator [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017], which allows us to measure the departure of a distribution R from a model P. We define $\mathcal{F}$ to be a space comprised of D-dimensional vectors of functions $f = (f_1, \ldots f_D)$ where the d-th function f_d is in a reproducing kernel Hilbert space (RKHS) [Aronszajn, 1950, Steinwart and Christmann, 2008, Definition 4.18] with a positive definite kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (we use the same kernel for each dimension). The inner product on $\mathcal{F}$ is $\langle f, g \rangle_{\mathcal{F}} := \sum_{d=1}^D \langle f_d, g_d \rangle_{\mathcal{F}_k}$, and $\mathcal{F}_k$ denotes an RKHS of real-valued functions with kernel k.

The (Langevin) *kernel Stein discrepancy* (KSD) between P and R is defined as

$$\mathrm{KSD}\left(P\|R\right) = \sup_{\|f\|_{\mathcal{F}}\leq 1} |\mathbb{E}_{x\sim R}\mathcal{T}_P f(x) - \mathbb{E}_{y\sim P}\mathcal{T}_P f(y)|. \tag{3.1}$$

Under appropriate conditions on the kernel and measure P, the expectation $\mathbb{E}_{y\sim P}\mathcal{T}_P f(y) = 0$ for any $f \in \mathcal{F}$. To ensure this property, we will require that $k \in C^{(1,1)}$, the set of continuous functions on $\mathcal{X} \times \mathcal{X}$ with continuous first derivatives and that $\mathbb{E}_{y\sim P}\left[\|\mathbf{s}_p(y)\|_2\right] < \infty$ with $\|\cdot\|_2$ the Euclidean norm. We further assume that the following tail condition holds outside a bounded set : $p(x)\sqrt{k(x,x)} \leq C\|x\|_2^{\delta}$ for some constants $C > 0$ and $\delta > D - 1$ [see the clarification by South et al., 2021, p.12, on the tail condition for the Stein's identity]. With the vanishing expectation $\mathbb{E}_{y\sim P}\mathcal{T}_P f(y) = 0$, the KSD reduces to $\mathrm{KSD}\left(P\|R\right) = \sup_{\|f\|_{\mathcal{F}}\leq 1} |\mathbb{E}_{x\sim R}\mathcal{T}_P f(x)|$. The use of an RKHS as the function class yields a closed form expression of the discrepancy by the kernel trick [Chwialkowski et al., 2016, Gorham and Mackey, 2017, Proposition 2],

$$\mathrm{KSD}^2\left(P\|R\right) = \mathbb{E}_{x,x'\sim R\otimes R}[h_p(x,x')],$$

if $\mathbb{E}_{x\sim R}[h_p(x,x)^{1/2}] < \infty$. Here, the symbol $R \otimes R$ denotes the product measure of two copies of R (so x and x' are independent random variables identically distributed with the law R). The function h_p (called a *Stein kernel*) is expressed in terms of the RKHS kernel k and the score function $\mathbf{s}_p$,

$$h_p(x,x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x')k(x,x') + \mathbf{s}_p(x)^\top k_1(x',x) + \mathbf{s}_p(x')^\top k_1(x,x') + k_{12}(x,x'),$$

where we have defined

$$\begin{aligned} k_1(a,b) &:= \nabla_x k(x,x')|_{x=a,x'=b}, \\ k_{12}(a,b) &:= \nabla_x^\top \nabla_{x'} k(x,x')|_{x=a,x'=b}. \end{aligned}$$

For a given i.i.d. sample $\{x_i\}_{i=1}^n \sim R$, the discrepancy has a simple closed-form finite sample estimate,

$$\mathrm{KSD}^2\left(P\|R\right) \approx \frac{1}{n(n-1)} \sum_{i\neq j} h_p(x_i,x_j), \tag{3.2}$$

which is a U-statistic [Hoeffding, 1948]. When the kernel is integrally strictly positive definite (ISPD) [Sriperumbudur et al., 2011, Section 2], and R admits a density r that satisfies $\mathbb{E}_{x\sim R}\left\|\nabla \log\big(p(x)/r(x)\big)\right\|_2 < \infty$, we have that $\mathrm{KSD}\left(P\|R\right) = 0$ iff $P = R$ [Barp et al., 2019, Proposition 1]. The earlier results of Chwialkowski et al. [2016] and Liu et al. [2016] require more stringent integrability conditions. Gorham and Mackey [2017, Theorem 7] showed that KSD can distinguish any Borel measure R from P by assuming conditions such as distant dissipativity (satisfied by finite Gaussian mixtures) [Gorham et al., 2019, Section 3]. However, such conditions may be difficult to validate for latent variable models. Thus, hereafter, we assume the former condition on the data distribution R.

Discrete-valued observations. We next recall the kernel Stein discrepancy in the discrete setting where distributions are defined on $\mathcal{X} = \{0, \ldots, L-1\}^D$ with $L > 1$, as introduced by Yang et al. [2018]. In place of derivatives, we specify Δ_k as the cyclic forward difference w.r.t. k-th coordinate: $\Delta_k f(x) = f(x^1, \ldots, \tilde{x}^k, \ldots, x^D) - f(x^1, \ldots, x^k, \ldots, x^D)$ where $\tilde{x}^k = x^k + 1 \mod L$, with the corresponding vector-valued operator $\Delta = (\Delta_1, \ldots, \Delta_D)$. The inverse operator Δ_k^{-1} is given by the backward difference $\Delta_k^{-1} f(x) = f(x^1, \ldots, x^k, \ldots, x^D) - f(x^1, \ldots, \bar{x}^k, \ldots, x^D)$, where $\bar{x}^k = x^k - 1 \mod L$, and $\Delta^{-1} = (\Delta_1^{-1}, \ldots, \Delta_D^{-1})$. The score is then $\mathbf{s}_p(x) := p(x)^{-1} \Delta p(x)$, where it is assumed that the pmf is strictly positive (i.e., it is never zero). The difference Stein operator is then defined as $\mathcal{A}_P f(x) = \operatorname{tr}\left[f(x)\mathbf{s}_p(x)^\top + \Delta^{-1} f(x)\right]$, where it can be shown that $\mathbb{E}_{x\sim P}[\mathcal{A}_P f(x)] = 0$ [Yang et al., 2018, Theorem 2] (note that we include a trace for consistency with the continuous case–this does not affect the test statistic [Yang et al., 2018, Eq. 10]). We have defined the Stein operator and the score function slightly differently from Yang et al. [2018]; the change is only in their signs, but this results in the same discrepancy. The difference Stein operator is not the only allowable Stein operator on discrete spaces: other alternatives are given by Yang et al. [2018, Theorem 3], Hodgkinson et al. [2020], and Shi et al. [2022]. Although we focus on the Stein operator above, in practice, one might want to consider different Stein operators depending on the application. For instance, the score function $\mathbf{s}_p$ can be numerically unstable, as it contains the reciprocal $1/p(x)$; this can occur when the support of the model is severely mismatched to that of the data. In this particular case, one might choose the Barker-Stein operator proposed by Shi et al. [2022], an instance of the Zanella-Stein operator of Hodgkinson et al. [2020, Example 2]. See Appendix 3.B for details. We compare this operator to the difference operator in an experiment where this mismatch occurs (Section 3.4.3.3).

As in the continuous case, the KSD can be defined as an IPM, given a suitable choice of reproducing kernel Hilbert space for the discrete domain. An example of kernel is the exponentiated Hamming kernel, $k(x, x') = \exp(-d_H(x, x'))$, where $d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x^d \neq x'^d)$. The population KSD is again given by the expectation of the Stein kernel, $\mathrm{KSD}^2(P\|R) = \mathbb{E}_{(x,x')\sim R\otimes R}[h_p(x, x')]$, where h_p is defined as

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^\top k_1(x', x) + \mathbf{s}_p(x')^\top k_1(x, x') + k_{12}(x, x'),$$

and the kernel gradient is replaced by the inverse difference operator, e.g., $k_1(x, x') = \Delta_x^{-1} k(x, x')$, where Δ_x^{-1} indicates that the operator Δ^{-1} is applied with respect to the argument x. From Yang et al. [2018, Lemma 8], we have that $\mathrm{KSD}(P\|R) = 0$ iff $P = R$, under the conditions that the probability mass functions for P and R are positive and that the Gram matrix defined over all the configurations in $\mathcal{X}$ is strictly positive definite (i.e., the kernel is integrally strictly positive definite). One can define a kernel satisfying the required condition, for example, by embedding $\mathcal{X}$ into $\mathbb{R}^{L\times D}$ with one-hot encoding and using a Taylor-type kernel such as the exponentiated quadratic kernel [Christmann and Steinwart, 2010, Theorem 2.2].

3.2.2 Kernel Stein discrepancies of latent variable models

Our objective is to use the KSD to evaluate latent variable models, and here we formally specify our target model class. Let $\mathcal{L}_{\mathcal{X}|\mathcal{Z}} = \{p(\cdot|z) : z \in \mathcal{Z}\}$ be a family of probability density functions on $\mathcal{X}$ (called likelihood functions), which are indexed by elements of a set $\mathcal{Z}$. A latent variable model P is specified by such a family $\mathcal{L}_{\mathcal{X}|\mathcal{Z}}$ and a (prior) probability measure P_Z over $\mathcal{Z}$. The combination of these defines the marginal density function $p(x) = \int p(x|z)\mathrm{d}P_Z(z)$ and the posterior distribution $P_Z(\mathrm{d}z|x) = \{p(x|z)/p(x)\}P_Z(\mathrm{d}z)$; The distribution P induced by the former acts as a model of the distribution R underlying the observation, and the latter enables us to draw an inference over the unobserved variable.

Remark 3.1. In our notation, the variable z can represent multiple latent variables. The likelihood $p(x|z)$ often contains parameters, but the dependency on these is suppressed here. If a prior is defined on a parameter, we may treat it as a latent variable; this consideration is relevant to predictive distributions. The likelihood and the prior in a model may be conditioned on some fixed data (i.e., they can be posterior predictive distributions), which we require to be independent of the data used for testing – in such a case, we omit the dependency on the held-out data. For examples, we refer the reader to Section 3.4.

The definition of the KSD remains the same for latent variable models, but an additional difficulty arises in its estimation. Unfortunately the U-statistic estimator given in (3.2) requires the score function of the marginal p, which is challenging to obtain due to the intractability of marginalizing out the latent variable. We will address this challenge by rewriting the score function in terms of the posterior distribution of the latent. In the following, we focus on continuous variable models, but the same conclusion holds for discrete counterparts by replacing gradient operation with cyclic differences.

Under a regularity condition, the score function can be expressed as

$$\mathbf{s}_p(x) = \mathbb{E}_{z|x}[\mathbf{s}_p(x|z)], \tag{3.3}$$

where $\mathbf{s}_p(x|z)$ is the score function of the conditional $p(x|z)$; i.e., $\mathbf{s}_p(x|z) = p(x|z)^{-1}\nabla_x p(x|z)$ for continuous-valued x. The reasoning is as follows:

$$\begin{aligned}\frac{\nabla_x p(x)}{p(x)} &= \frac{1}{p(x)}\int \nabla_x p(x|z)\mathrm{d}P_Z(z)\\ &= \int \frac{\nabla_x p(x|z)}{p(x|z)}\cdot\frac{p(x|z)\mathrm{d}P_Z(z)}{p(x)} = \mathbb{E}_{z|x}[\mathbf{s}_p(x|z)],\end{aligned}$$

where we have assumed the exchangeability of differentiation and integration: $\nabla_x p(x) = \int \nabla_x p(x|z)\mathrm{d}P_Z(z)$. The identity (3.3) is an analogue of Fisher's identity [Fisher, 1925, Dempster et al., 1977], which pertinently formed the basis for Stein control variate methodology in [Friel et al., 2016] and Bayesian model selection with Hyvärinen score [Dawid and Musio, 2015, Shao et al., 2019]. Note that the conditional score $\mathbf{s}_p(x|z)$ is typically possible to evaluate. For example, consider the following simple form of an exponential family density $p(x|z) \propto \exp(T(x)\eta(z))$ defined on $\mathbb{R}^D$ with $T(x) : \mathbb{R}^D \to \mathbb{R}$ and $\eta : \mathcal{Z} \to \mathbb{R}$; for this density, $\mathrm{s}_p(x|z) = \eta(z)\nabla_x T(x)$. As can be seen in this example, the normalizing constant of the likelihood $p(x|z)$ is not required.

With this identity, the KSD is rewritten as follows.

Lemma 3.2. *Let*

$$\begin{aligned} H_p[(x,z),(x',z')] = \mathbf{s}_p(x|z)^\top \mathbf{s}_p(x'|z')k(x,x') + \mathbf{s}_p(x|z)^\top k_1(x',x) \\ + k_1(x,x')^\top \mathbf{s}_p(x'|z') + k_{12}(x,x'). \end{aligned} \tag{3.4}$$

Assume $\mathbb{E}_{(x,z),(x',z')\sim\tilde{R}\otimes\tilde{R}}|H_p[(x,z),(x',z')]| < \infty$ *with the joint distribution* $\tilde{R}(\mathrm{d}(x,z)) = P_Z(\mathrm{d}z|x)R(\mathrm{d}x)$. *If the formula (3.3) holds, then,*

$$\mathrm{KSD}^2\,(P\|R) = \mathbb{E}_{(x,z),(x',z')\sim\tilde{R}\otimes\tilde{R}} H_p[(x,z),(x',z')].$$

Proof. Substituting the formula (3.3) in the definition of KSD gives the required equation by the Tonelli-Fubini theorem. □

Remark 3.3. The integrability assumption holds trivially if the input space $\mathcal{X}$ is finite, while care needs to be taken otherwise. The condition can be checked by examining the absolute integrability of each term in (3.4). The integrability assumption on the fourth term is mild, and is satisfied by common kernels, e.g., the exponentiated quadratic or the inverse multi-quadratic kernels. The condition on the other terms needs to be checked on a model-by-model basis. It can be shown that the example models in Section 3.4 satisfy the assumption (please see Section 3.C in the supplementary material for details).

The new KSD expression is an expectation of a computable symmetric kernel, and constructing an unbiased estimate is straightforward once we obtain a sample. In practice, when the model is complex, sampling from the posterior distribution generally requires simulation, as the posterior is not available in closed form. Therefore, we propose to approximate the expectation by Markov Chain Monte Carlo (MCMC) methods and construct an approximate U-statistic estimator as follows. Let $\mathbf{z}_i^{(t)} = \left(z_{i,1}^{(t)}, \cdots, z_{i,m}^{(t)}\right) \in \mathcal{Z}^m$ be a latent sample of size m drawn by an MCMC method having $P_Z(\cdot|x_i)$ as its invariant measure after t burn-in iterations. Let $\bar{\mathrm{s}}_p(x_i|\mathbf{z}_i^{(t)}) = \frac{1}{m}\sum_{j=1}^m \mathrm{s}_p(x_i|z_{i,j}^{(t)})$. Given a joint sample $\left\{\left(x_i, \mathbf{z}_i^{(t)}\right)\right\}_{i=1}^n$, we estimate the KSD by

$$U_n^{(t)}(P) := \frac{1}{n(n-1)}\sum_{i\neq j} \bar{H}_p\left[\left(x_i,\mathbf{z}_i^{(t)}\right),\left(x_j,\mathbf{z}_j^{(t)}\right)\right], \tag{3.5}$$

where

$$\begin{aligned} \bar{H}_p\left[\left(x_i,\mathbf{z}_i^{(t)}\right),\left(x_j,\mathbf{z}_j^{(t)}\right)\right] = \bar{\mathrm{s}}_p(x_i|\mathbf{z}_i^{(t)})^\top \bar{\mathrm{s}}_p(x_j|\mathbf{z}_j^{(t)})k(x_i,x_j) + \bar{\mathrm{s}}_p(x_i|\mathbf{z}_i^{(t)})^\top k_1(x_j,x_i) \\ + k_1(x_i,x_j)^\top \bar{\mathrm{s}}_p(x_j|\mathbf{z}_j^{(t)}) + k_{12}(x_i,x_j), \end{aligned}$$

and the sum is taken over all distinct sample pairs. If $P_Z^{(t)}(\mathrm{d}\mathbf{z}|x)$ denotes the distribution of an MCMC sample $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_m^{(t)})$, then this estimator is indeed a U-statistic, but its expectation is that of kernel $\bar{H}_p$ with respect to $P_Z^{(t)}(\mathrm{d}\mathbf{z}|x)R(\mathrm{d}x)$ instead of $P_Z(\mathrm{d}\mathbf{z}|x)R(\mathrm{d}x)$. Thus, the estimator is biased against the target estimand, the model's KSD, for a finite burn-in period t, and can therefore be seen an approximation to the *true* U-statistic $U_n^{(\infty)}$. Designing a statistical test requires understanding the behavior of the statistic (3.5), and we will provide its analysis in the next section. Although we focus on MCMC for its approximate unbiasedness in our proposed test, different posterior approximations may be considered in other applications; for example, with a more computationally efficient approach (e.g., variational approximation), the new KSD expression in Lemma 3.2 might allow us to consider parameter estimation for unnormalized statistical models with latent variables [Barp et al., 2019].

3.3 A relative goodness-of-fit test

We now address the setting of statistical testing for model comparison. We begin this section with our problem settings and notation, and then define a test by showing the asymptotic normality of approximate U-statistics.

3.3.1 Problem setup

We consider the case where we have two latent variable models P and Q, and we wish to determine which is a closer approximation of the distribution R generating our data $\{x_i\}_{i=1}^n$. The respective density functions of the models are given by the integrals $p(x) = \int p(x|z)\mathrm{d}P_Z(z)$ and $q(x) = \int q(x|w)\mathrm{d}Q_W(w)$. As with P, the latent variable w is assumed to take values in a set $\mathcal{W}$ with prior Q_W. We assume that $p(x)$ and $q(x)$ cannot be tractably evaluated, even up to their normalizing constants. Our goal is to determine the *relative* goodness-of-fit of the models by comparing each model's discrepancy from the data distribution. Our problem is formulated as the following hypothesis test:

$$\begin{aligned} H_0 &: \mathrm{KSD}\,(P\|R) \leq \mathrm{KSD}\,(Q\|R) \ \ \text{(null hypothesis)}, \\ H_1 &: \mathrm{KSD}\,(P\|R) > \mathrm{KSD}\,(Q\|R) \ \ \text{(alternative)}. \end{aligned} \tag{3.6}$$

In other words, the null hypothesis is that the fit of P to R (in terms of KSD) is as good as Q, or better. Note that the KSD in (3.6) is defined by a particular reproducing kernel, and thus different kernels yield distinct hypotheses. For kernel selection, we refer the reader to Section 3.3.4.

We next provide an overview of the formal assumptions made throughout this chapter. Let $(\Omega, \mathcal{S}, \Pi)$ be a probability space, where Ω is a sample space, $\mathcal{S}$ is a σ-algebra, and Π is a probability measure. All random variables (for example, data points x_i and draws $\mathbf{z}_i^{(t)}$ from a Markov chain sampler) are understood as measurable functions from the sample space Ω. The input space $\mathcal{X}$ is equipped with the Borel σ-algebra generated by its standard topology. We assume that $\mathcal{Z}, \mathcal{W}$ are Polish spaces with the Borel σ-algebras defined by their respective topologies, on which the priors $P_Z,\ Q_W$ are defined. Finally, we require that the two models are distinct; i.e., their marginal densities disagree on a set of positive R-measure.

3.3.2 Estimating kernel Stein discrepancies of latent variable models

The hypotheses in (3.6) can be equally stated in terms of the difference of the (squared) KSDs, $\mathrm{KSD}^2\left(P\|R\right) - \mathrm{KSD}^2\left(Q\|R\right)$, which motivates us to design a test statistic by estimating each term. Let $U_n^{(t)}(P,Q) \coloneqq U_n^{(t)}(P) - U_n^{(t)}(Q)$ be the difference of KSD estimates, where, $U_n^{(t)}(Q)$ is defined as for $U_n^{(t)}(P)$ in (3.5). Note that $U_n^{(t)}(P,Q)$ is an *approximate* U-statistic (in the sense of the final paragraph in Section 3.2.2) defined by the difference kernel

$$\bar{H}_{p,q}[(x,\mathbf{z},\mathbf{w}),(x',\mathbf{z}',\mathbf{w}')] \coloneqq \bar{H}_p[(x,\mathbf{z}),(x',\mathbf{z}')] - \bar{H}_q[(x,\mathbf{w}),(x',\mathbf{w}')]$$

evaluated on the joint sample $\left\{\left(x_i,\mathbf{z}_i^{(t)},\mathbf{w}_i^{(t)}\right)\right\}_{i=1}^n$. The statistic takes as input random variables with evolving laws, and defining a test require us to understand the behavior of such statistics. This section delivers an analysis in a general setting.

We first characterize the asymptotic distribution of an approximate U-statistic. The following theorem shows that such a statistic is asymptotically normal around the expectation of the true U-statistic provided its bias vanishes fast.

Theorem 3.4 (Asymptotic normality)**.** *Let $\{\gamma_t\}_{t=1}^\infty$ be a sequence of Borel probability measures on a Polish space $\mathcal{Y}$ and γ be another Borel probability measure. Let $\{Y_i^{(t)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \gamma_t$, and for a symmetric function $h:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$, define a U-statistic and its mean by*

$$U_n^{(t)} = \frac{1}{n(n-1)}\sum_{i\neq j} h\left(Y_i^{(t)},Y_j^{(t)}\right),\ \theta_t = \mathbb{E}_{(Y,Y')\sim\gamma_t\otimes\gamma_t}[h(Y,Y')].$$

Let $\theta = \mathbb{E}_{(Y,Y')\sim\gamma\otimes\gamma}[h(Y,Y')]$. Let $\nu_t \coloneqq \mathbb{E}_{(Y,Y')\sim\gamma_t\otimes\gamma_t}\left[|\tilde{h}_t(Y,Y')|^3\right]^{1/3}$ with $\tilde{h}_t = h-\theta_t$, and assume $\limsup_{t\to\infty}\nu_t<\infty$. Assume that $\sigma_t^2 = 4\mathrm{Var}_{Y'\sim\gamma_t}\left[\mathbb{E}_{Y\sim\gamma_t}[h(Y,Y')]\right]$ converges to a constant σ^2. Assume that we have the double limit $\sqrt{n}(\theta_t-\theta)\to 0$; i.e., for any $\varepsilon>0$, there exists $N\geq 1$ such that $\sqrt{n}(\theta_t-\theta)\leq\varepsilon$ for any $n,t\geq N$. Then, if $\sigma>0$, we have

$$\sqrt{n}\left(U_n^{(t)}-\theta\right)\xrightarrow{\mathrm{d}}\mathcal{N}(0,\sigma^2)\text{ as }n,t\to\infty,$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. In the case $\sigma=0$, $\sqrt{n}(U_n^{(t)}-\theta)\to 0$ in probability.

The proof is in Section 3.6 in the supplement. Note that in the preceding and following results, the limit of n and t is taken simultaneously rather than sequentially, such that the condition $\sqrt{n}(\theta_t-\theta)\to 0$ holds: see discussion below and in Section 3.3.3. By letting $Y_i^{(t)} = \left(x_i,\mathbf{z}_i^{(t)},\mathbf{w}_i^{(t)}\right)$ and $h=\bar{H}_{p,q}$ in the foregoing theorem, we obtain the same conclusion for the difference estimate $U_n^{(t)}(P,Q)$.

The asymptotic normality allows us to define a test procedure. Theorem 3.4 involves unknown variance σ^2, however; in order to construct a test, we need to be able to estimate it consistently. For our test, we propose to use the following jackknife variance estimator

$$v_{n,t} \coloneqq (n-1)\sum_{i=1}^n\left(U_{n,-i}^{(t)}-U_n^{(t)}\right)^2 \tag{3.7}$$

where $U_n^{(t)}$ is defined as in Theorem 3.4, and $U_{n,-i}^{(t)}$ the U-statistic computed on the sample with the i-th data point removed. We defer the discussion on this choice until we introduce our test procedure in Section 3.3.3. Here, we present the required consistency, the proof of which can be found in Appendix 3.6.2 (see Lemma 3.8).

Lemma 3.5. *Define symbols as in Theorem 3.4 and the jackknife variance estimator as in (3.7). Assume*

$$\limsup_{t\to\infty} \mathbb{E}_{(Y,Y')\sim\gamma_t\otimes\gamma_t}[h(Y,Y')^4] < \infty.$$

Let $\sigma^2 = \lim_{t\to\infty}\sigma_t^2$ *where* $\sigma_t^2 = 4\zeta_{1,t} = 4\mathrm{Var}_{Y\sim\gamma_t}\left[\mathbb{E}_{Y'\sim\gamma_t}\left[h(Y,Y')\right]\right]$. *Then, we have the double limit* $\mathbb{E}\left(v_{n,t}-\sigma^2\right)^2 \to 0$ *as* $n,t\to\infty$. *In particular, the limit holds regardless of the growth rate of* t *as a function of* n.

We have shown that the jackknife estimator allows consistent estimation of the asymptotic variance of $U_n^{(t)}(P,Q)$. Using the results obtained in this section, we present our test procedure in the next section.

3.3.3 Test procedure

We are finally ready to define the test procedure. Recall that our objective is to compare model discrepancies, which can be accomplished by estimating the difference $\mathrm{KSD}^2\left(P\|R\right) - \mathrm{KSD}^2\left(Q\|R\right)$. The previous section has established the asymptotic normality of the difference estimate $U_n^{(t)}(P,Q)$ and provides a consistent estimator of its asymptotic variance. Therefore, we define our test statistic to be

$$T_{n,t} = \sqrt{n}\frac{U_n^{(t)}(P,Q)}{\sigma_{n,t}}, \tag{3.8}$$

where $\sigma_{n,t} = \left(v_{n,t}\right)^{1/2}$ with $v_{n,t}$ the jackknife variance estimator in (3.7) computed using the joint sample $\left\{\left(x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)}\right)\right\}_{i=1}^n$ and kernel $h = \bar{H}_{p,q}$. The following property follows from Theorem 3.4 and Slutsky's lemma [see e.g., van der Vaart, 2000, p. 13] along with the consistency of $\sigma_{n,t}$ from Lemma 3.5.

Corollary 3.6. *Let* $\mu_{P,Q} = \mathrm{KSD}^2\left(P\|R\right) - \mathrm{KSD}^2\left(Q\|R\right)$. *Let* $\mathbf{y}_1^{(t)}$ *and* $\mathbf{y}_2^{(t)}$ *be i.i.d. variables;* $\mathbf{y}_1^{(t)}$ *represents a copy of random variables* $(x, \mathbf{z}^{(t)}, \mathbf{w}^{(t)})$*; the variables* $\mathbf{z}^{(t)}, \mathbf{w}^{(t)}$ *are draws from the respective Markov chains of* P *and* Q *conditioned on* x *after* t *burn-in steps, and* x *obeys* R *(the starting points* $\mathbf{z}^{(1)}, \mathbf{w}^{(1)}$ *for the two Markov chains are shared). Assume* $\limsup_{t\to\infty}\mathbb{E}[\bar{H}_{p,q}[\mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}]^4] < \infty$. *If the assumptions in Theorem 3.4 hold for the statistic* $U_n^{(t)}(P,Q)$ *with asymptotic variance* $\sigma_{P,Q}^2 > 0$, *we have* $\sqrt{n}(U_n^{(t)}(P,Q) - \mu_{P,Q})/\sigma_{n,t} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ *as* $n,t\to\infty$.

Remark 3.7. Corollary 3.6 holds for any choice of the Markov chain sample size $m \geq 1$. However, in practice, a small value of m leads to large variance of the score estimates $\bar{\mathbf{s}}_p, \bar{\mathbf{s}}_q$, and hence the test statistic $T_{n,t}$, which results in a conservative test. To improve the test's sensitivity, we therefore recommend using as large an m as possible.

Corollary 3.6 leads to the following simple model comparison test (summarized in Algorithm 1): for a given significance level $\alpha \in (0,1)$, we compare the test statistic $T_{n,t}$ against the $(1-\alpha)$-quantile $\tau_{1-\alpha}$ of the standard normal, and reject the null if $T_{n,t}$ exceeds $\tau_{1-\alpha}$. By this design, under the null hypothesis $H_0 : \mu_{P,Q} \leq 0$, we have $\lim_{n,t\to\infty} \Pi(T_{n,t} > \tau_{1-\alpha}|H_0) \leq \alpha$, and the test is therefore asymptotically level α for each fixed R satisfying H_0 [Lehmann and Romano, 2005, Definition 11.1.1]. On the other hand, under any fixed alternative $H_1 : \mu_{P,Q} > 0$, it follows from $\sqrt{n}\mu_{P,Q} \to \infty$ $(n \to \infty)$ that we have $\lim_{n,t\to\infty} \Pi(T_{n,t} > \tau_{1-\alpha}|H_1) = 1$, indicating that the test is consistent in power.

Algorithm 1: Test procedure

```
Input: Data $\{x_i\}_{i=1}^n$, models $P$, $Q$, and significance level $\alpha$
Result: Test the null $H_0$
/*Form a joint sample $\{(x_i, \mathbf{z}_i^{(t)}, \mathbf{w}_i^{(t)})\}_{i=1}^n$                */
for $i \leftarrow 1$ to $n$ do
    Generate $m$ samples $\mathbf{z}_i^{(t)} = (z_{i,1}^{(t)}, \ldots, z_{i,m}^{(t)})$ after $t$ burn-in steps with an MCMC
      algorithm to simulate $P_Z(\mathrm{d}z|x_i)$;
    Generate $m$ samples $\mathbf{w}_i^{(t)} = (w_{i,1}^{(t)}, \ldots, w_{i,m}^{(t)})$ after $t$ burn-in steps with an MCMC
      algorithm to simulate $Q_W(\mathrm{d}w|x_i)$;
end
$\tau_{1-\alpha} \leftarrow (1-\alpha)$-quantile of $\mathcal{N}(0,1)$;
/*Compute test statistic $T_{n,t}$ in equation (3.8)                                    */
Compute KSD difference estimate $U_n^{(t)}(P,Q)$;
Compute variance estimate $v_{n,t}$;
/*Direct computation of $T_{n,t} = \sqrt{n}U_n^{(t)}(P,Q)/\sqrt{v_{n,t}}$ can be numerically
  unstable                                                                              */
if $U_n^{(t)}(P,Q) > (\sqrt{v_{n,t}}/\sqrt{n}) \cdot \tau_{1-\alpha}$ then Reject the null $H_0$ ;
```

We remark that the above analysis will not apply in particular extreme cases, where both models are identical, or both perfectly match the data distribution. When these occur, then $\sigma_{P,Q} = 0$ and $\mu_{P,Q} = 0$ (note that if $\mu_{p,Q} \neq 0$, the test statistic diverges as the sample size increases). Applying our procedure as above to this setting, the normal approximation might fail to correctly capture the variability of the test statistic, and the type-I error could exceed the significance level. To detect this failure mode, we would need to independently check that the two models are not identical, either by inspection or via two-sample testing. A more systematic treatment could be performed, e.g., by preventing degeneracy using a sample splitting technique as proposed by Schennach and Wilhelm [2017], and we leave this refinement for future work.

We empirically found that our choice of the variance estimator acted as a safeguard against the failure mode mentioned above. The jackknife estimator is nonnegative, while individually estimating the variances and covariance of the two U-statistics might yield a negative estimate. The jackknife is also known to overestimate the variance [Efron and Stein, 1981], and its use may result in a more conservative test. This estimator is not the only allowable choice, as the variance estimation in U-statistics has been extensively studied [for other concrete estimators see, e.g., Maesono, 1998, and references therein]. In our preliminary analysis, we considered two other estimators, but the jackknife estimator controlled type-I errors better than these alternatives in the *near degenerate* case. For details, we refer the reader to experiments in Sections 3.J.3 and 3.J.4 in the supplement.

The limiting behaviors of the test are only guaranteed when an appropriate double limit is taken with respect to the burn-in size t and the sample size n. Theorem 3.4 suggests that the bias of the statistic $U_n^{(t)}(P, Q)$ should decay faster than $1/\sqrt{n}$ in the limit of t. Our practical recommendation is to take a burn-in period as long as the computational budget allows; this heuristic is justified if the bias vanishes as $t \to \infty$. For $\mathrm{KSD}^2 (P\|R)$ and its estimate, the bias is due to that of the score estimate $\mathbf{s}_p^{(t)}(x) = \mathbb{E}_{\mathbf{z}|x}^{(t)}[\bar{\mathbf{s}}_p(x|\mathbf{z})]$, where $\mathbb{E}_{\mathbf{z}|x}^{(t)}$ denotes the expectation with respect to $P_Z^{(t)}(\mathrm{d}\mathbf{z}\,|x)$. If the score's bias is confirmed to converge to zero, we can check the bias of the KSD estimate by examining the convergence of $\mathbb{E}_{(x,x')\sim R\otimes R}[h_{p,t}(x,x')]$, with $h_{p,t}(x,x')$ a Stein kernel defined by the approximate score $\mathbf{s}_p^{(t)}$. The convergence of $\mathbf{s}_p^{(t)}$ can be established by assuming appropriate conditions on $\mathbf{s}_p(x|z)$ and the sampler; for instance, for the exponential family likelihood $p(x|z) \propto \exp(T(x)\eta(z))$, if the natural parameter η is a continuous bounded function, the weak convergence of the sampler implies the desired convergence (the score $\mathbf{s}_p(x|z)$ is factorized as $\eta(z)\nabla T(x)$). The quantification of the required growth rate of t relative to n needs more stringent conditions on the employed MCMC sampler, which we discuss in the supplement, Section 3.D. Admittedly, it is often not straightforward to theoretically establish an explicit relation between the growth rates of t and n. We therefore experimentally evaluate the finite-sample performance of our test in Section 3.4.

The overall computational cost of the proposed test is $O\{n^2 + n(t+m)\}$, assuming that the cost of sampling a latent is constant. The test statistic in (3.8) requires evaluating the U-statistic kernel $\bar{H}_{p,q}$ on all distinct sample pairs. Note that we need to perform this computation only once if we memoize the evaluated values; in particular, the cost of the variance estimate (3.5) can be made $O(n^2)$ with memoization. Thus, assuming that we have evaluated and stored the score values $\left\{\bar{\mathbf{s}}_p\left(x_i|\mathbf{z}_i^{(t)}\right), \bar{\mathbf{s}}_q\left(x_i|\mathbf{w}_i^{(t)}\right)\right\}_{i=1}^n$, the cost of evaluating the U-statistic kernel is $O(n^2)$, but this operation can be easily parallelized over sample pairs. The additional $O\{n(t+m)\}$ cost comes from evaluating the approximate score functions, as it requires running Markov chains for each data point (see the loop between Lines 1-4 in Algorithm 1). We can improve the sample-size n dependency in score evaluation by parallelization, since MCMC can be performed independently over sample points x_i.

3.3.4 Kernel choice

A discrepancy measure such as KSD encodes a particular sense of how two distributions differ. In the case of KSD, the magnitude of this discrepancy is affected not only by evaluated models but also the choice of a reproducing kernel. Ideally, we should choose a kernel that makes the KSD reflect the discrepancy of features relevant to the problem at hand. We provide general guidance on kernel selection as follows:

Continuous observations: As mentioned in Section 3.2, ISPD kernels enable the KSD to distinguish any two distributions satisfying certain regularity conditions. Of ISPD kernels, in the light of practical performances reported in prior work in goodness-of-fit testing

[Gorham and Mackey, 2017] and distribution approximation [Chen et al., 2019, Riabiz et al., 2021], we advocate for the use of the preconditioned IMQ kernel [Chen et al., 2019]

$$k(x, x') = \left(c^2 + \|\Lambda^{-1/2}(x - x')\|_2^2\right)^{-\beta} \tag{3.9}$$

with Λ a strictly positive definite matrix, and scalars $c > 0$ and $0 < \beta < 1$; as a default choice, we recommend to take $\beta = 1/2$ and $c = 1$. Following the kernel method literature, we recommend to choose the pre-conditioner Λ in a data-dependent way so that the KSD can capture relevant features of the data. We suggest two default options: the median heuristic, where $\Lambda = \lambda^2 I$ with $\lambda = \mathrm{median}\{\|x_i - x_j\|_2 : 1 \leq i < j \leq n\}$ and I the identity matrix; the sample covariance $\Lambda = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top/(n-1)$ with $\bar{x} = \sum_{i=1}^n x_i/n$, which should be suitably regularized. Each of these choices has its own merits, as we illustrate in a simple example with Gaussian distributions in the supplement (Section 3.F.2). Moreover, in general, the KSD is not invariant to a change of coordinates representing the data. The above choices partially address this issue, as they ensure that the KSD is invariant to rotation and displacement [see Section 3.H.2 in the supplement; Matsubara et al., 2021, Section 5.1] . For continuous observations, we additionally need to examine the integrability of the Stein kernel to use the KSD expression of Lemma 3.2. To this end, one might want to make an assumption about the tail decay of the data distribution. The integrability condition can be alternatively enforced by reweighting the reproducing kernel so that the Stein kernel is uniformly bounded; i.e., for a kernel k, define a new kernel k_w by $k_w(x, x') = k(x, x')w(x)w(x')$ where $w : \mathcal{X} \to (0, \infty)$ is some decreasing function dominating the growth of the score function. We discuss how to choose such w in the supplement, Section 3.C. This reweighting might reduce the sensitivity of the KSD and break the aforementioned property of coordinate-choice independence, however.

Discrete observations: We have given a condition for a kernel to be ISPD at the end of Section 3.2; e.g., the exponentiated quadratic kernel on one-hot encoding, which can be efficiently implemented in a sparse tensor format. In general, however, it is challenging to compute such an ISPD kernel for discrete objects in high-dimensions. Note that ISPD-ness is only required to distinguish *any* two distributions. In practice, we only require the KSD to capture aspects relevant to model evaluation, and might therefore choose a kernel insensitive to some differences, as long as they represent computationally affordable alternatives suited to the given problem. An instructive example is testing on distributions over graphs. Graphs of V nodes can be represented as adjacency matrices that are elements of $\{0, 1\}^{V \times V}$. The Dirac delta kernel that examines if two graphs are identical is an ISPD kernel but computationally intractable (no polynomial algorithm is known). This notion of graph identification is in practice too restrictive, and therefore one typically uses kernels that convey other relevant graph properties [see e.g., Borgwardt et al., 2020, for more details]. We also demonstrate this trade-off in our experiment with latent Dirichlet allocation models in Section 3.4.3, where we can ignore the sequential structure of the data.

Use of multiple kernel functions: As we have seen, there are often multiple choices of the kernel function, and they might represent distinct features. Our recommendation is to test the hypotheses corresponding to the kernel choices simultaneously, as it makes evaluation more rigorous. However, one has to correct for multiple comparisons such as controlling the family-wise error rate. It should be noted that a correction typically makes the test more conservative as the number of kernels grows. The user thus needs to control the number of kernels to be used (e.g., using a handful of values of scale parameter λ for the above IMQ kernel with $\Lambda = \mathrm{diag}(\lambda, \dots, \lambda)$).

Finally, we note that specific kernels can be employed that encode domain-specific expertise in particular problem settings: for instance, kernels have been defined on groups [Fukumizu et al., 2008] and graphs [Borgwardt et al., 2020]. KSDs and associated statistical tests can likewise be defined for certain of these cases [e.g., Xu and Matsuda, 2020]. That being said, it may sometimes be preferable to favor an MMD with goal-specific features over an omnibus KSD test.

3.4 Experiments

We evaluate the proposed test (LKSD, hereafter) through simulations. Our goal is to show the utility of the KSD in model comparison. To this end, we compare our test with the relative MMD test [Bounliphone et al., 2016], a kernel-based frequentist test that supports a great variety of latent variable models. Note that this test and ours address different hypotheses, as the MMD and LKSD tests use different discrepancy measures; it is indeed possible that they reach conflicting conclusions (e.g., a model is better in terms of KSD but worse in MMD). To align the judgement of both tests, we construct problems using models with controllable parameters; for a given class, a reference distribution, from which a sample is drawn, is chosen by fixing the model parameter; two candidates models are then formed by perturbing the reference's parameter such that a larger perturbation yields a worse model for both tests. We show that there are cases where the MMD fails to detect model differences whereas the KSD succeeds. For completeness, we provide the detail of our implementation of the MMD test in the supplement (Section 3.E), since it requires modification to yield satisfactory performance in our setting. Code to reproduce all the results is available at `https://github.com/noukoudashisoup/lkgof`.

Following are details shared by the experiments below. All results below are based on 300 trials, except for the experiment in Section 3.4.2 (see the section for details). In the light of the discussion in Section 3.3, unless specified, the MMD test draws $n_{\mathrm{model}} = m + t$ samples for each model so that its cost matches the additional computation afforded to LKSD test, which is of $O\{n(m + t)\}$ from MCMC.

3.4.1 Probabilistic Principal Component Analysis

We first consider a simple model in which the score of its marginal is tractable. This allows us to separately assess the impact of employing a score function approximation. Probabilistic Principal Component Analysis (PPCA) models serve this purpose since the marginals are given by Gaussian distributions. Let $\mathcal{X} = \mathbb{R}^D$ and $\mathcal{Z} = \mathbb{R}^{D_z}$ with $1 \leq D_z < D$. A PPCA model $\mathrm{PPCA}(A, \psi)$ is defined by

$$p(x|z, A, \psi) = \mathcal{N}(Az, \psi^2 I_x),\ P_Z = \mathcal{N}(\mathbf{0}, I_z),$$

where $A \in \mathbb{R}^{D\times D_z}$, $I_x \in \mathbb{R}^{D\times D}$, $I_z \in \mathbb{R}^{D_z \times D_z}$ are the identity matrices, ψ is a positive scalar, and $\mathbf{0}$ is a vector of zeros. The conditional score function is $\mathbf{s}_p(x|z) = -(x - Az)/\psi^2$. In particular, the marginal density is given by $p(x) = \mathcal{N}(\mathbf{0}, AA^\top + \psi^2 I_x)$.

While the posterior in this model is tractable, it is instructive to see how KSD estimation is performed by MCMC. By using an MCMC method, such as the Metropolis Adjusted Langevin Algorithm (MALA) [Besag, 1994, Roberts and Tweedie, 1996] or Hamiltonian Monte Carlo (HMC) [Duane et al., 1987, Neal, 2011], we obtain latent samples $\mathbf{z}_i \in \mathcal{Z}^m$ for each x_i, which forms a joint sample $\{(x_i, \mathbf{z}_i)\}_{i=1}^n$; samples $\mathbf{z}_i$ are used to compute a score estimate at each point x_i,

$$\bar{\mathbf{s}}_p(x_i|\mathbf{z}_i) = -\left\{ x_i - A\Big(\frac{1}{m}\sum_{j=1}^m z_{i,j}\Big)\right\}/\psi^2,$$

and these approximate score values are used to compute the U-statistic estimate in (3.5). By choosing suitably decaying kernels (Section 3.C), we can guarantee the integrability condition in Lemma 3.2. The vanishing bias assumption in Theorem 3.4 corresponds to the convergence in mean, which can be measured by the Kantorovich–Rubinstein distance [Kantorovich, 2006] (also known as the L^1-Wasserstein distance [see, e.g., Villani, 2009, Chapter 6]). Note that the negative logarithm of the unnormalized posterior density is strongly convex, and its gradient is Lipschitz; the strong convexity- and Lipschitz constants are independent of x. Therefore, using HMC for example, by appropriately choosing a duration parameter and a discretization step size, we can show that the bias of the above score estimate diminishes uniformly over x [Bou-Rabee et al., 2020].

3.4.1.1 Type-I error and test power

We investigate the finite-sample performance of the proposed test in terms of type-I error and power rates. We generate data from a PPCA model $R = \mathrm{PPCA}(A, \psi)$. The dimensions of the observable and the latent are set to $D = 100$, $D_z = 10$, respectively. Each element of the weight matrix A is drawn from a uniform distribution $U[0, 1]$ and fixed. The variance parameter ψ is set to 1. As PPCA models have tractable marginals, we also compare our test with the KSD test using exact score functions (i.e. no MCMC simulation), which serves as the performance upper-bound. The MCMC sampler we use is HMC; more precisely, we use the NumPyro [Phan et al., 2019] implementation of No-U-Turn Sampler (NUTS) [Hoffman and Gelman, 2014]; we take $t = 200$ burn-in samples and $m = 500$ consecutive draws for computing a score estimate

$\bar{\mathbf{s}}_p$.

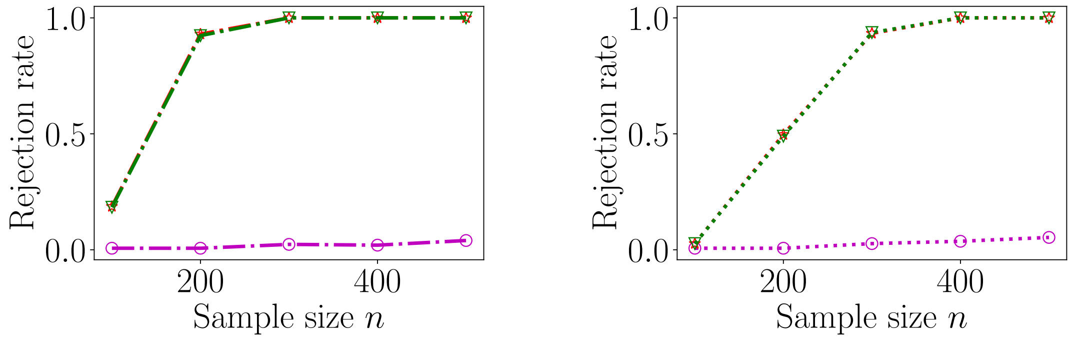
We use two kernel functions: (a) the exponentiated quadratic (EQ) kernel $k(x, x') = \exp\{-\|x - x'\|_2^2/(2\lambda^2)\}$, and (b) the IMQ kernel (3.9) with $\beta = 0.5, c = 1$, and $\Lambda = \lambda^2 I$. All three tests use the same kernel function, which allows us to investigate the effect of using the Stein-modified kernel. The length scale parameter λ is set to the median of the pairwise (Euclidean) distances of holdout samples from R so that the parameter (and thus the hypothesis) is fixed across trials. We include the EQ kernel in our comparison, as the population MMD is possible to compute, allowing us to verify the hypothesis in advance.

We simulate null and alternative cases by perturbing the weight parameter A; we add a positive value $\delta > 0$ to the $(1, 1)$-entry of A. Let us denote a perturbed weight by A_δ. Note that the data PPCA model has a Gaussian marginal $\mathcal{N}(\mathbf{0}, AA^\top + \psi^2 I_x)$. Therefore, this perturbation gives a model $\mathcal{N}(\mathbf{0}, A_\delta A_\delta^\top + \psi^2 I_x)$, where the first row and column of $A_\delta A_\delta^\top$ deviate from those of $AA^\top$. The perturbation is additive and increasing in δ, as each element of A is positive. We create a problem by specifying perturbation parameters (δ_P, δ_Q) for (P, Q). For the EQ-kernel MMD, we numerically confirmed that the perturbation gives a worse model for a larger perturbation. While the population KSD is not analytically tractable, this perturbation affects the score function through the covariance matrix, and the same behavior is expected for KSD; see Section 3.F in the supplement for details.

Table 3.1: Type-I errors the MMD test of Bounliphone et al. [2016], the proposed LKSD test, and the KSD test in PPCA Problem 1. Rejection rates are computed on 300 trials with significance level $\alpha = 0.05$. The columns EQ-med and IMQ-med denote EQ and IMQ kernels with the median bandwidth, respectively.

Sample size n	Rejection rates					
		EQ-med			IMQ-med	
	MMD	LKSD	KSD	MMD	LKSD	KSD
100	0.000	0.013	0.000	0.000	0.010	0.000
200	0.000	0.000	0.000	0.000	0.000	0.000
300	0.003	0.007	0.000	0.003	0.003	0.000
400	0.003	0.007	0.000	0.003	0.000	0.000
500	0.007	0.013	0.000	0.007	0.007	0.000

Problem 1 (null) We create a null scenario by choosing $(\delta_P, \delta_Q) = (1, 1 + 10^{-5})$ (P has a smaller covariance perturbation and is closer to R than Q). For different null settings, we refer the reader to Section 3.J.3 in the supplement. We run the tests with significance levels $\alpha = 0.01, 0.05$. Table 3.1 reports the finite-sample size of the three tests for significance level $\alpha = 0.05$. The result for $\alpha = 0.01$ is omitted as none of the tests rejected the hypotheses. The size of the proposed LKSD test is indeed controlled. The extremely small type-I errors of the KSD tests are caused by the sensitivity of KSD to this perturbation; the population KSD value is negative and far from zero, and the test statistics easily fall in the acceptance region. The other two tests also have their error rates lower than the significant level. Note that their test thresholds are determined by treating the population discrepancy differences as zero, resulting in conservative tests.

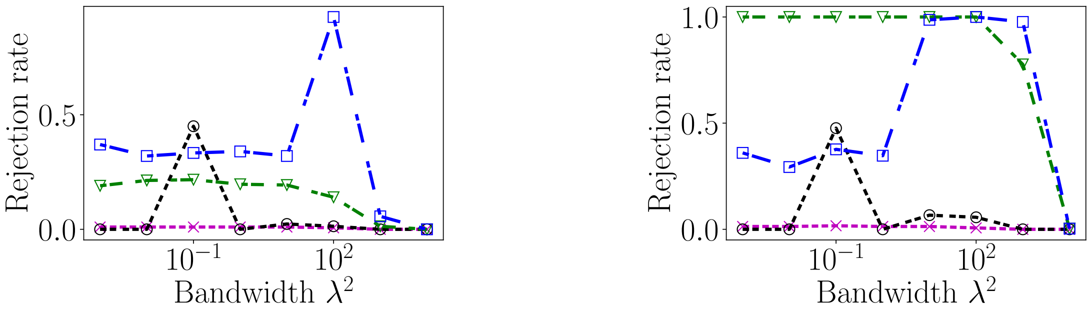


(a) : EQ kernel with median scaling.

(b) IMQ kernel with median scaling.

Figure 3.1: Power curves of the MMD test of Bounliphone et al. [2016], the proposed LKSD test, and the KSD test with the exact score function in PPCA Problem 2. The perturbation parameters are set as $(\delta_P, \delta_Q = 2, 1)$. each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: $\bigtriangledown$ (the LKSD test); ☆ (the KSD test); ◯ (the relative MMD test).

Problem 2 (alternative) We investigate the power of the proposed test. We set up a alternative scenario by fixing $\delta_P = 2$ for P and $\delta_Q = 1$ for Q. The significance level α is fixed at 0.05. All the other parameters are chosen as in Problem 1. Figure 3.1 shows the plot of the test power against the sample size in each problem. The KSD reaches a near 100 percent rejection rate relatively quickly, indicating that information from the score function is helpful for these problems. The effect of the score approximation is negligible in this experiment, as the power curve of the LKSD test overlaps with that of KSD. The power of the MMD test is substantially lower than the other tests, indicating that the MMD is insensitive to this perturbation to the covariance.



(a) $n = 100$

(b) $n = 300$

Figure 3.2: Power curves of the proposed LKSD test and the MMD test in PPCA Problem 2. The perturbation parameters are set as $(\delta_P, \delta_Q = 2, 1)$. each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: $\bigtriangledown$ (LKSD test with IMQ kernel); □ (LKSD test with EQ kernel); ◯ (MMD test with IMQ kernel); $\times$ (MMD test with EQ kernel).

3.4.1.2 Effect of kernel parameter choice

Dependency on scaling parameter. Using Problem 2 above, we examine how the test power is affected by the scaling parameter. We use the EQ and IMQ kernels as above, and choose their scaling parameter λ^2 from $\{10^{-3}, 10^{-2}, \ldots, 10^3\}$. For each $n \in \{100, 300\}$ we run 300 trials and estimate the test power of the LKSD and MMD tests. Figure 3.2 plots the power curves of the tests. We can see that the high-power region of the EQ kernel is localized while the IMQ kernel's power curves are flat, indicating that the IMQ kernel does not depend on the parameter as much as the EQ. Therefore, for this problem, the IMQ kernel can be seen as more robust against misspecification of the scaling parameter. Nonetheless, with the right choice of the scaling parameter, the EQ kernel yields higher power for both MMD and KSD tests. It can be considered that the distinction arises because of the local nature of the difference between the two distributions; the EQ kernel is more sensitive in choosing features used to compute the KSD (see Section 3.F).

Different parameterization. We also consider a different parameter choice for the preconditioning matrix. Here, we compare the median-scaled IMQ kernel with the same kernel having a covariance preconditioning matrix, as suggested in Section 3.3.4. Figure 3.3 shows the power curves of the three tests. Here, the relation between the MMD and KSD tests is overturned, and the KSD test struggles to detect the perturbation to the covariance. This result demonstrates that certain kernel choices can make the testing problem more challenging than others. Using multiple kernels, rather than relying on a single choice, could therefore robustify the evaluation, at the expense of a loss of power due to multiple testing correction.

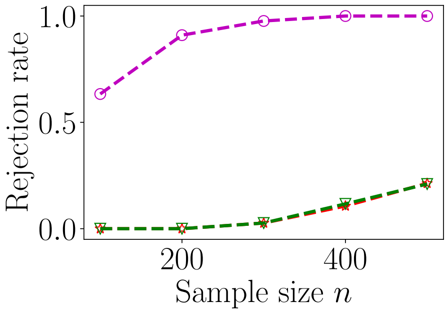
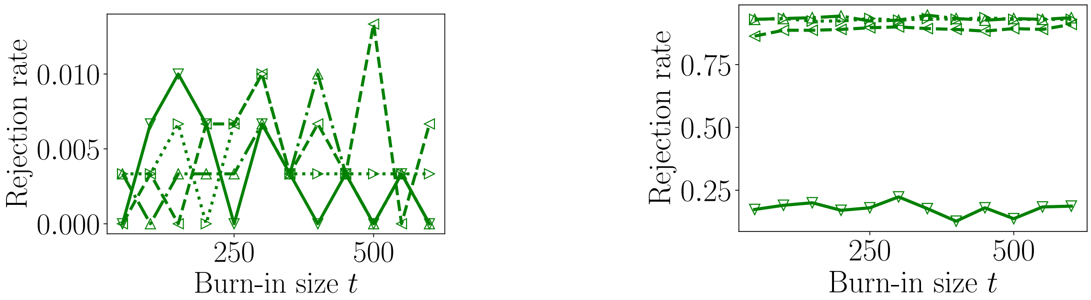


Figure 3.3: Power curves of the MMD test, the proposed LKSD test, and the KSD test in PPCA Problem 2. All the test use the covariance-preconditioned IMQ kernel. The perturbation parameters are set as $(\delta_P, \delta_Q = 2, 1)$. Each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: $\bigtriangledown$ (the LKSD test); ☆ (the KSD test); $\bigcirc$ (the relative MMD test).

3.4.1.3 Quality of Markov chain samplers

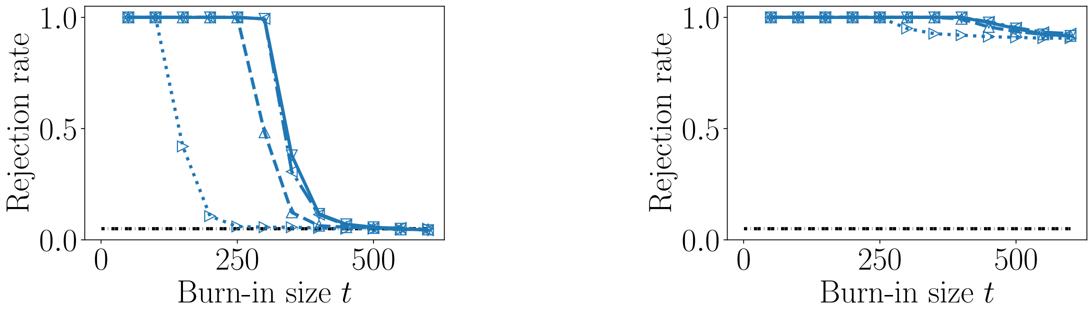
The asymptotic property of our test (Corollary 3.6) hinges on the quality of the Markov chain samplers. This section studies the effect of these Markov chains on the inference. We vary the burn-in size t and the score approximation sample size m, which is expected to affect the type-I error rate and the power of the test. In the experiments below, we set $\alpha = 0.05$. We choose t from $\{50, 100, \ldots, 600\}$ and m from $\{1, 10, 100, 1000\}$.



(a) Problem 1 (null H_0 is true).

(b) Problem 2 (alternative H_1 is true).

Figure 3.4: The effect of MCMC quality on the test's performance. Rejection rates against burn-in size t with varying Markov chain sample size m. PPCA Problems 1 and 2 with $\alpha = 0.05$. Both samplers use NUTS. Markers: ▽ ($m = 1$); ◁ ($m = 10$); △ ($m = 100$); ▷ ($m = 1000$).



(a) $n = 100$

(b) $n = 300$

Figure 3.5: The effect of a poor MCMC sampler on the test. Type-I error rates against the burn-in size t with varying Markov chain sample size m. PPCA Problem 1 (the null H_0 is true). The dark dashed line indicates the significance level $\alpha = 0.05$. The samplers for P and Q are respectively MALA and NUTS. Markers: ▽ ($m = 1$); ◁ ($m = 10$); △ ($m = 100$); ▷ ($m = 1000$).

In our first experiment, as in the previous sections, we use the NUTS with the same initialization strategy for both models. With $n = 300$, we run the test using Problems 1 and 2 above. Figure 3.4 shows rejection rates of the test for different settings of t and m. In both cases, the burn-in length t does not affect the test's performance, indicating the fast convergence of the sampler. The importance of a larger value of m can be seen when the alternative hypothesis holds, since the test power improves as m increases. The improved performance is likely due to reduced variance.

We next consider a slow-converging sampler for which the burn-in length t becomes crucial. We consider the null case (Problem 1) and replace the sampler for the first model P with MALA. We set the step size for the MALA sampler to make its convergence slow; we use the step size of $10^{-4}D_z^{-1/3}$. We initialize the two samplers differently to make sure that the resulting distributions differ when the samplers have not converged: the MALA sampler for P is initialized with samples from a Gaussian $\mathcal{N}\{(1, \ldots, 1), I_z\}$ and the NUTS sampler for Q a uniform distribution $U[-2, 2]^{D_z}$. Figure 3.5 demonstrates the relation between type-I error rates and choices of t and m. In contrast to the previous experiment, the burn-in has a clear effect on the type-I error: insufficient burn-in leads to uncontrolled error rates. The right panel ($n = 300$) shows that the test has substantially higher type-I error rates than in the left ($n = 100$). Comparison between these cases illustrates that a larger sample size n requires more intensive burn-in, as the test becomes more confident to reject. A large value of m improves the test as in the previous experiment. It can be understood that the contribution of burn-in samples is negligible in the score approximation. Although our analysis in Corollary 3.6 requires long burn-in, taking large m appears to be more important in practice, especially under a computational budget constraint. This experiment thus confirms the importance of the quality of the sampler.

3.4.2 Dirichlet process mixtures

Our next experiment applies our test to a Dirichlet process mixtures (DPM) model. Let $\psi(x|z)$ be a probability density function on $\mathbb{R}^D$. We consider a mixture density

$$\int \psi(x|z)\mathrm{d}F(z)$$

where F is a Borel probability measure on a Polish space $\mathcal{Z}$. A DPM model [Ferguson, 1983] places a Dirichlet process prior $\mathrm{DP}(a)$ on the mixing distribution F. Thus, a DPM model $\mathrm{DPM}(a)$ assumes the following generative process:

$$x_i|z_i, \phi, F \overset{\text{ind.}}{\sim} \psi(x|z_i),\ z_i|F \overset{\text{i.i.d.}}{\sim} F,\ F \sim \mathrm{DP}(a).$$

Here, a is a finite Borel measure on $\mathcal{Z}$. Note that the marginal density is given by

$$\mathbb{E}_F\left[\int \psi(x|z)\mathrm{d}F(z)\right].$$

Although the prior has an infinite-dimensional component, the required conditional score function is simply $\mathrm{s}_\psi(x|z,\phi)$; thus we only need to sample from a finite-dimensional posterior $P_Z(\mathrm{d}z|x)$. Practically, DPM models are of limited use without observations to condition them; i.e., we are interested in their predictive distributions. If a model is conditioned on held-out data $\mathcal{D}$, then the predictive density $p(x|\mathcal{D})$ is $\mathbb{E}_{F|\mathcal{D}}\left[\int \psi(x|z)\mathrm{d}F(z)\right]$, and its score is given by the expectation of $\mathrm{s}_\psi(x|z)$ with respect to the posterior

$$\frac{\psi(x|z,\phi)}{p(x|\mathcal{D})}\bar{F}_{\mathcal{D}}(\mathrm{d}z)$$

with $\bar{F}_{\mathcal{D}}$ the mean measure of $P_F(\mathrm{d}F|\mathcal{D})$. Sampling from the posterior can be performed with a combination of the Metropolis-Hastings algorithm and Gibbs sampling [see e.g., Ghosal and van der Vaart, 2017, Chapter 5]. For the score formula and the MCMC procedure, we refer the reader to Section 3.G in the supplement. By setting ψ to an isotropic normal density, for example, we can guarantee the integrability assumption in Lemma 3.2 (see Section 3.C).

Experiment details. We consider the following simple Gaussian DPM model $\mathrm{GDPM}(\mu)$,

$$x_i \overset{\text{ind.}}{\sim} \mathcal{N}(z_i, 2I),\ z_i \overset{\text{i.i.d.}}{\sim} F, F \sim \mathrm{DP}(a),\ a = \mathcal{N}(\mu, I),$$

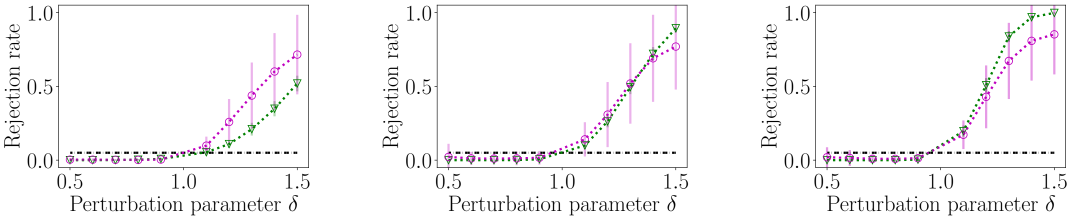
where $\mu \in \mathbb{R}^D$, and I is the identity matrix of size $D \times D$. Note that without conditioning on observations, the model's marginal density is simply a Gaussian distribution $\mathcal{N}(\mu, 3I)$, which does not require approximation.

We therefore compare predictive distributions, i.e., we compare two GDPM models conditioned on *training data* $\mathcal{D}_{\mathrm{tr}} = \{\tilde{x}_i\}_{i=1}^{n_{\mathrm{tr}}} \overset{\text{i.i.d.}}{\sim} R$. We consider two GDPM models with wrong priors, where their prior means are shifted. Specifically, we take $R = \mathrm{GDPM}(\mathbf{0})$, and two models chosen as $Q = \mathrm{GDPM}(\bar{\mathbf{1}})$ and $P = \mathrm{GDPM}(\delta\bar{\mathbf{1}})$ with $\bar{\mathbf{1}} = \mathbf{1}/\sqrt{D}$. Unlike the preceding experiments, we condition the two models on the training data, and obtain the predictive distributions, denoted by $P_{\mathcal{D}_{\mathrm{tr}}}$ and $Q_{\mathcal{D}_{\mathrm{tr}}}$, respectively; our problem is thus the comparison between $P_{\mathcal{D}_{\mathrm{tr}}}$ and $Q_{\mathcal{D}_{\mathrm{tr}}}$. The distributions now require simulating their posterior, and we use a random-scan Gibbs sampler and the Metropolis algorithm with a burn-in period $t = 1,000$ and the size of the latents $m = 500$. For sampling observables from the models, we use a random-scan Gibbs sampler with a burn-in period $2,000$. We expect that if the training sample size n_{tr} is small, a larger perturbation would give a worse model as the effect of the prior is still present; we thus set $n_{\mathrm{tr}} = 5$. Due to the small sample size, the expected model relation might not hold, depending on the draw of $\mathcal{D}_{\mathrm{tr}}$. Therefore, we examine the rejection rates of the LKSD and MMD tests, averaged over 50 draws; for each draw of $\mathcal{D}_{\mathrm{tr}}$, we estimate the rejection rates based on 100 trials. Our problem is formed by varying the perturbation scale δ for $P_{\mathcal{D}_{\mathrm{tr}}}$, which is chosen from a regular grid $\{0.5, 0.6, \cdots, 0.9, 1.1, \cdots, 1.5\}$. This construction gives a null case when $\delta < 1$, the alternative otherwise. We set the dimension D to 10 and the significance level α to 0.05. As in Section 3.4.1, we use the IMQ kernel with median scaling.

Figure 3.6 reports the rejection rates of the two tests for each of $n \in \{50, 100, 200\}$. Note that the curves in the graph do not represent type-I errors nor power, as they are rejection rates *averaged* over draws $\mathcal{D}_{\mathrm{tr}}$, each of which forms a different problem. It can be seen that on average, both tests have correct sizes ($\delta < 1$). In the alternative regime ($\delta < 1$), the LKSD test underperforms the MMD with a small sample size ($n = 50$); however, its improvement in power is faster and exceeds the MMD at $n = 200$. These results imply that the LKSD estimate has a large variance for a small sample size, whereas its estimand (the population difference) is also larger, and thus the mean of the test statistic diverges faster. Thus, it may be understood that the KSD is more sensitive to model differences in this setting.

3.4.3 Latent Dirichlet Allocation

Our final experiment studies the behavior of the LKSD test on discrete data using Latent Dirichlet Allocation (LDA) models. LDA is a mixed-membership model [Airoldi et al., 2014] for grouped discrete data such as text corpora. We follow Blei et al. [2003] and use the terminology of text data for ease of exposition. Accordingly, the following terms are defined using our notation. A word is an element in a discrete set (a vocabulary) $\{0, \ldots, L-1\}$ of size L. A document x is a sequence of D words, i.e., $x \in \{0, \ldots, L-1\}^D$ is a D-dimensional discrete vector. A



(a) $n = 50$ (b) $n = 100$ (c) $n = 200$

Figure 3.6: Comparison in Gaussian Dirichlet mixture models. Rejection rates plotted against the perturbation parameter δ. The sample size n is chosen from $\{50, 100, 200\}$. The rejection rates are averaged over draws of $\mathcal{D}_{\mathrm{tr}}$. The supposed null and alternative regimes are $\delta < 1$ and $\delta > 1$, respectively. Markers: ▽ (the LKSD test); ◯ (the relative MMD test). The dark dashed line indicates the significance level $\alpha = 0.05$. The errorbars indicate the standard deviations of the estimated rejections rates.

prominent feature of LDA is that it groups similar words assuming they come from a shared latent *topic*, which serves as a mixture component. An LDA model assumes the following generative process on a corpus of documents $\{x_i\}_{i=1}^n$:

1. For each document $i \in \{1, \ldots, n\}$, generate a distribution over K topics $\theta_i \overset{\text{i.i.d.}}{\sim} \mathrm{Dir}(a)$ (the Dirichlet distribution), where θ_i is a probability vector of size $K \geq 1$.
2. For the j-th word $x_i^j, j \in \{1, \ldots, D\}$ in a document i,
   (a) Choose a topic $z_i^j \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(\theta_i)$.
   (b) Draw a word from $x_i^j \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(b_k)$, where b_k is the distribution over words for topic k, and the topic assignment $z_i^j = k$.

Here, $a = (a_1, \ldots, a_K)$ is a vector of positive real numbers, and $b = (b_1, \ldots, b_K)^\top \in [0,1]^{K \times L}$ represents a collection of K distributions over L words. In summary, an LDA model $P = \mathrm{LDA}(a, b)$ assumes the factorization

$$\prod_{i=1}^{n} p(x_i|z_i, \theta_i; a, b) p(z_i, \theta_i; a, b) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{D} p(x_i^j|z_i^j, b) p_z(z_i^j|\theta_i) \right\} p_\theta(\theta_i|a),$$

where z_i and θ_i act as latent variables.

Because of the independence structure over words, the conditional score function is simply given as

$$\mathbf{s}_p(x|z, \theta, a, b) = \mathbf{s}_p(x|z, b) = \left(\frac{p(\tilde{x}^j|z^j, b)}{p(x^j|z^j, b)} - 1 \right)_{j=1,\ldots,D}, \text{ where } \tilde{x}^j = x^j + 1 \mod L.$$

Score approximation requires the posterior distribution $p(z|x; a, b)$ with respect to z. Marginalization of θ renders latent topics dependent on each other, and thus the posterior is intractable. A latent topic is conjugate to the corresponding topic distribution given all other topics. Therefore, an MCMC method such as collapsed Gibbs sampling allows us to sample from $p(z|x; a, b)$. As the observable and the latent are supported on finite sets, the use of Lemma 3.2 is justified;

Table 3.2: Rejection rates of the MMD test and the LKSD test in LDA experiments. Each result is based on 300 trials.

(a) Type-I errors of the KSD and MMD tests in LDA Problem 1; $(\delta_P, \delta_Q) = (0.5, 0.6)$. The significance level $\alpha = 0.05$.

Sample size n	Rejection rates	
	MMD	LKSD
100	0.003	0.013
200	0.010	0.007
300	0.007	0.003
400	0.003	0.007
500	0.007	0.010

(b) Power of the KSD and MMD tests in LDA Problem 2; $(\delta_P, \delta_Q) = (1.0, 0.5)$. The significance level α is chosen from $\{0.01, 0.05\}$.

Sample size n	Rejection rates			
	Level $\alpha = 0.01$		Level $\alpha = 0.05$	
	MMD	LKSD	MMD	LKSD
100	0.000	0.010	0.007	0.070
200	0.003	0.030	0.010	0.183
300	0.000	0.097	0.003	0.283
400	0.000	0.197	0.010	0.463
500	0.000	0.280	0.007	0.570

the finite moment assumptions in Corollary 3.6 are guaranteed; and the consistency of the population mean and variance of the test statistic follows from the convergence of $\mathbb{E}^{(t)}_{\mathbf{z}|x}[\bar{\mathbf{s}}_p(x|\mathbf{z})]$ and $\mathbb{E}^{(t)}_{\mathbf{w}|x}[\bar{\mathbf{s}}_q(x|\mathbf{w})]$ for each $x \in \mathcal{X}$. Note that we have implicitly ordered the vocabulary set to define the score function $\mathcal{X}$. A naive ordering might induce a discrepancy measure not useful for model comparison with respect to a given dataset (e.g., the score function might not vary on the data points). One might consider data-dependent ordering such as sorting by the word frequency in the given data. Investigating ordering choices appropriate for model evaluation is an interesting research topic and remains an open question.

3.4.3.1 Synthetic data – prior sparsity perturbation

In the two problems below, we observe a sample $\{x_i\}_{i=1}^n$ from an LDA model $R = \mathrm{LDA}(a, b)$. The number of topics is $K = 3$. The hyper-parameter a is chosen as $a = (a_0, a_0, a_0)$; for model R, we set $a_0 = 0.1$. Each of three rows in $b = (b_1,b_2, b_3)^\top \in [0, 1]^{3\times L}$ is fixed at a value drawn from the symmetric Dirichlet distribution with all the concentration parameters one, and the vocabulary size is $L = 10,000$. Each $x_i \in \{0, \dots, L-1\}^D$ is a document consisting of $D = 50$ words.

We design problems by perturbing the sparsity parameter a_0. Recall that $\mathrm{Dir}(a)$ is a distribution on the $(K-1)$ - probability simplex. A small $a_0 < 1$ makes the prior $p_\theta(\theta_i|a) = \mathrm{Dir}(a)$ concentrate its mass on the vertices of the simplex; the case $a_0 = 1$ corresponds to the uniform distribution on the simplex; choosing $a_0 > 1$ leads to the prior mass concentrated on the center of the simplex. The data distribution R (with $a_0 = 0.1$) is thus intended to draw sparse topic proportions θ_i, and a document x_i is likely to have words from a particular topic. By increasing a_0, we can design a departure from this behavior. Therefore, as in the PPCA experiments, we additively perturb a_0 with parameters (δ_P, δ_Q) for respective candidate models (P, Q).

As LDA disregards word order, we need a kernel that respects this structure. We use the Bag-of-Words (BoW) IMQ kernel $k(x, x') = (1 + \|B(x) - B(x')\|_2^2)^{-1/2}$; it is simply the IMQ kernel computed in the BoW representation $B(x) \in \{0, 1, 2, \dots, D\}^L$ whose ℓ-th entry (counting from 0 to $L-1$) is the count of the occurrences of word $\ell \in \{0, \dots, L-1\}$ in a document x. By Lemma 3.17 in Section 3.H.1, this choice ensures that arbitrary reordering of text sequences does not change the KSD value; i.e., the KSD does not assess models by their ability to generate sequences. We also tested differing input-scaling values and found that the bandwidth of the IMQ kernel did not have a significant effect on the test power (Section 3.J.2.2).

For score estimation in the LKSD test, we use a random scan Gibbs sampler; we generate $m = 1,000$ latent samples after $t = 4,000$ burn-in iterations.

Problem 1 (null) We create a null situation by having $(\delta_P, \delta_Q) = (0.5, 0.6)$. In this case, Q's prior on θ is less sparse than that of P. Table 3.2a shows the size of the different tests for significance levels $\alpha = 0.05$; the result for $\alpha = 0.01$ is omitted as both tests did not reject the hypothesis. It can be seen that the rejection rates of both tests are bounded by the nominal level.

Problem 2 (alternative) We consider an alternative case in which the sparsity parameters are chosen as $(\delta_P, \delta_Q) = (1.0, 0.5)$. Here, the model Q is expected to have less mixed topic proportions. Table 3.2b demonstrates the power of the MMD and LKSD tests. The power of the LKSD test improves as the sample size n increases, whereas the MMD has almost no power in this case. In this problem, the topics b are not sparse enough for each topic to have a sufficiently distinctive vocabulary. Thus, the problem is challenging for the MMD, as it is unable to find distinguishing words, in addition to the high-dimensionality. By contrast, the KSD is able to distinguish the models by taking advantage of their underlying structure.

3.4.3.2 Synthetic data – topic perturbation

We provide a negative example to illustrate a failure mode of the LKSD test for discrete data. The data is generated as in the previous section, whereas we construct two models differently. We set up a model by perturbing the topics of the data model R. That is, a model is given by $\mathrm{LDA}(a, b_\delta)$ with $b_\delta = (1-\delta)b + \delta b_{\mathrm{ptb}}$ with $0 < \delta < 1$. We choose b_{ptb} as we did for b; the value is drawn independently of b. We set the perturbation parameter for Q as $\delta = 0.01$ and vary it for P, where the value is chosen from $\{0.06, 0.11, \ldots, 0.51\}$. Thus, P is morphed from b to b_{ptb} and therefore expected to underperform Q as perturbation δ increases. We run trials with $n = 300$. For score estimation, we take $m = 10,000$ and $t = 4,000$.

Figure 3.7 shows the plot of rejection rates against perturbation parameters. We see that the power of the LKSD test degrades as the perturbation increases. As P's topic becomes close to b_{ptb}, some words in the target's topic b become rare and therefore fall in the low probability region of P. This situation leads to increasing variance of the test statistic as δ increases, because the score function contains the reciprocal $1/p(x)$. The LKSD test can therefore fail when the support of the model is severely mismatched to that of the data, since the high variance of the statistic makes is difficult to detect significant departures from the null. Note that this observation does not apply to the continuous counterpart as the score can be written as the gradient of the logarithm of the density, which is typically numerically stable.

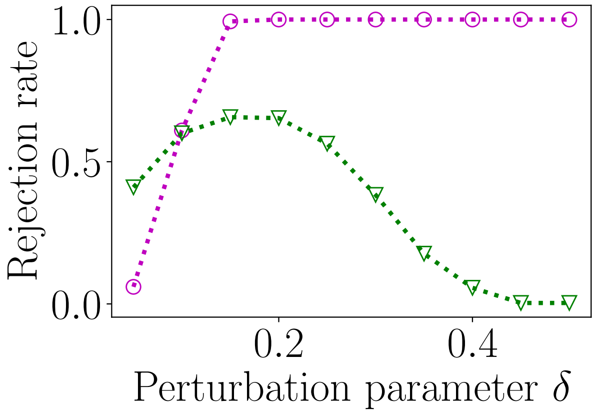


Figure 3.7: Power estimates plotted against perturbation parameters δ. The significance level $\alpha = 0.05$; the sample size $n = 300$. Markers: $\bigtriangledown$ (the LKSD test); $\bigcirc$ (the MMD test).

3.4.3.3 Comparing topic models for arXiv articles

Our final experiment investigates the test's performance using the arXiv dataset [Cornell University, 2020]. The dataset consists of meta information of scholarly articles on the e-print service arXiv. We treat the abstract of an article as a document, and use paper categories to set up a problem. Specifically, we construct a problem by choosing three paper categories for model P, Q and the data distribution R. Unlike the preceding experiments, for a model category, we fit an LDA model to the dataset of abstracts in the category. As the KSD requires the number of words to be fixed, then for a given data category, we extract abstracts of length no less than $D = 100$ and subsample excess words. This process yields a dataset of articles of equal length D; for each trial, we obtain the data $\{x_i\}_{i=1}^{n}$ by subsampling from the larger set of articles. Thus, our problem is to compare two LDA models trained on different article sets, and assess their fit to the dataset.

In the following experiments, we examine the power of LKSD and MMD tests. We vary the sample size n from 100 to 500. We fix the dataset category to stat.TH (statistics theory) and inspect two combinations of model categories. To train an LDA model $\mathrm{LDA}(a, b)$, we use the Gensim implementation [Rehurek and Sojka, 2011] of the variational algorithm of Hoffman et al. [2010]. For sparsity parameters a, we use the parameter returned by this algorithm; we point-estimate topics b using the mean of the topics under the variational distribution. The number of topics is set to 100. The vocabulary set is comprised of words that appear in the abstracts of three chosen categories. As in the previous experiments, we use the IMQ-BoW kernel for both tests. We fix the significance level α at 0.05.

As we have seen the numerical instability issue in the previous section, we also consider an alternative KSD that is stable but computationally more expensive, as mentioned in 3.2.1 and the supplement (Section 3.B). For this, we take a burn-in size $t = 500$ and a Markov chain size $m = 1,000$. We denote this method by LKSD-stable.

Probability theory vs Statistical methodology. We choose math.PR (mathematics probability theory) for P and stat.ME (statistics methodology) for Q. In addition to the taxonomic proximity

Table 3.3: Rejection rates of the MMD test and the LKSD test in the math.PR vs. stat.ME experiment. Each result is based on 300 trials.

Sample size n	Rejection rates				
	MMD	MMD-extra	LKSD	LKSD-extra	LKSD-stable
100	0.150	0.157	0.333	0.673	0.437
200	0.160	0.167	0.807	0.880	0.845
300	0.197	0.207	0.913	0.980	0.950
400	0.180	0.187	0.950	0.986	0.970
500	0.267	0.263	0.966	0.993	0.983

to stat.TH, the category stat.ME has a larger proportion of articles shared with the target category: $3,121$ of $18,973$ (stat.ME) vs. $2,884$ of $46,769$ (math.PR). Thus, we expect Q to outperform P. This combination results in a vocabulary set of size $L = 126,190$. For score estimation, we set the burn-in length t to 500 and the Markov chain sample size m to $5,000$. Additionally, we run the LKSD test with $m = 15,000$ (labeled LKSD-extra) and the MMD test with the model sample size $n_{\mathrm{model}} = 10,000$ (labeled MMD-extra). The n_{model} is thresholded at $10,000$ as the computational cost exceeds that of the LKSD test (in fact, sampling in this case makes the MMD by an order of magnitude slower due to the large vocabulary size).

Table 3.3 summarizes the result. The MMD test underperforms all the KSD-based tests; extra sampling did not lead to a significant improvement. We can see that increasing the Markov chain size m boosts the LKSD test, as it reduces the variance of the score estimator. The low power of the MMD test indicates that the model difference is too subtle to discern from the word compositions of generated documents; the LKSD tests offers a different viewpoint based on the model information.

Table 3.4: Rejection rates of the MMD test and the LKSD test in the cs.LG vs. stat.ME experiment. Each result is based on 300 trials.

Sample size n	Rejection rates		
	MMD	LKSD	LKSD-stable
100	1	0.000	0.287
200	1	0.007	0.643
300	1	0.013	0.833
400	1	0.013	0.873
500	1	0.113	0.923

Machine learning vs Statistical Methodology. Our second experiment uses cs.LG (computer science machine learning) for P, while Q uses the same category as the previous experiment. With this combination, the vocabulary size L is $208,671$. By the same reasoning as above, the second model Q is expected to be better than P. We run the same tests as above and compare their performances.

Table 3.4 shows the result. This experiment serves as a negative case study for the LKSD test: the MMD tests achieved power 1 for all sample-size choices (MMD-extra is omitted here), whereas the power of the LKSD test does not exceed even the significance level α for most sample size settings (LKSD-extra is omitted as increasing the Markov chain size did not improve the power). We attribute this failure to the unmatched support of the model P in the test distribution. This reasoning is supported by the high power of the MMD, as the BoW feature easily detects deviation of document patterns in this case. Thus, as we noted in the synthetic experiment in Section 3.4.3.2, the LKSD test fails when there is a severe mismatch in data and model support. The stable LKSD test approaches the same level as the MMD at $n = 500$, but still underperforms. While stable, the KSD used for this test can also suffer from the mismatch of the support, since it depends on the same density ratio as in the unstable counterpart.

3.5 Conclusion

We have developed a test of relative goodness of fit for latent variable models based on the kernel Stein discrepancy. The proposed test applies to a wide range of models, since the requirements of the test are mild: (a) models have MCMC samplers for inferring their latent variables, and (b) likelihoods have evaluable score functions. The proposed test complements existing model evaluation techniques by providing a different means of model comparison, which takes advantage of the known model structure. Our experimental results confirm this view – the relative MMD test was unable to detect subtle differences between models in several of our benchmark experiments.

Our asymptotic analysis of the test statistic indicates that the test could suffer from bias if the mixing of the deployed MCMC sampler is slow. Removing the assumptions on the bias and the moments in Theorem 3.4 is certainly desirable; we envision that the recent development of unbiased MCMC [Jacob et al., 2020] could be used to construct an alternative unbiased KSD estimator, and leave this possibility as future work. While we have focused on comparing two models, extensions to ranking multiple models are possible as in [Lim et al., 2019]. Finally, the technique used in this chapter can be applied to other Stein discrepancies requiring the score function [Barp et al., 2019, Xu and Matsuda, 2020]; one interesting application would be the KSD for directional data [Xu and Matsuda, 2020], where densities with computable normalizing constants are scarce.

3.6 Proofs

This section provides proofs for the results concerning the asymptotic normality of our test statistic: (a) Theorem 3.4, and (b) an estimator of the variance of a U-statistic and its consistency.

3.6.1 Asymptotic normality of approximate U-statistics

Theorem 3.4 (Asymptotic normality)**.** *Let $\{\gamma_t\}_{t=1}^{\infty}$ be a sequence of Borel probability measures on a Polish space $\mathcal{Y}$ and γ be another Borel probability measure. Let $\{Y_i^{(t)}\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \gamma_t$, and for a symmetric function $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, define a U-statistic and its mean by*

$$U_n^{(t)} = \frac{1}{n(n-1)} \sum_{i \neq j} h\big(Y_i^{(t)}, Y_j^{(t)}\big),\ \theta_t = \mathbb{E}_{(Y,Y') \sim \gamma_t \otimes \gamma_t}[h(Y, Y')].$$

Let $\theta = \mathbb{E}_{(Y,Y')\sim\gamma\otimes\gamma}[h(Y,Y')]$. Let $\nu_t := \mathbb{E}_{(Y,Y')\sim\gamma_t\otimes\gamma_t}\left[|\tilde{h}_t(Y,Y')|^3\right]^{1/3}$ with $\tilde{h}_t = h - \theta_t$, and assume $\limsup_{t\to\infty} \nu_t < \infty$. Assume that $\sigma_t^2 = 4\mathrm{Var}_{Y'\sim\gamma_t}\left[\mathbb{E}_{Y\sim\gamma_t}[h(Y,Y')]\right]$ converges to a constant σ^2. Assume that we have the double limit $\sqrt{n}(\theta_t - \theta) \to 0$; i.e., for any $\varepsilon > 0$, there exists $N \geq 1$ such that $\sqrt{n}(\theta_t - \theta) \leq \varepsilon$ for any $n, t \geq N$. Then, if $\sigma > 0$, we have

$$\sqrt{n}\left(U_n^{(t)} - \theta\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2) \text{ as } n, t \to \infty,$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. In the case $\sigma = 0$, $\sqrt{n}(U_n^{(t)} - \theta) \to 0$ in probability.

Proof. Recall that $(\Omega, \mathcal{S}, \Pi)$ is the underlying probability space, and $U_n^{(t)}$ is a random variable on it. We show that the cumulative distribution function (CDF) of $\sqrt{n}(U_n^{(t)} - \theta)$ converges to that of a normal distribution.

First we consider the case $\sigma > 0$. Note that we can express the CDF as

$$\begin{aligned}\Pi\left[\sqrt{n}\left(\frac{U_n^{(t)} - \theta}{\sigma}\right) < \tau\right] &= \Pi\left[\sqrt{n}\left(\frac{U_n^{(t)} - \theta_t + \theta_t - \theta}{\sigma_t}\right) < \frac{\sigma}{\sigma_t}\tau\right] \\ &= F_{n,t}\left(\frac{\sigma\tau - \sqrt{n}(\theta_t - \theta)}{\sigma_t}\right)\end{aligned}$$

where $F_{n,t}$ denotes the CDF of $\sqrt{n}(U_n^{(t)} - \theta_t)/\sigma_t$. Let Φ be the CDF of the standard Gaussian distribution. Then, for $\tau \in \mathbb{R}$,

$$\begin{aligned}\left|\Pi\left[\sqrt{n}\left(\frac{U_n^{(t)} - \theta}{\sigma}\right) < \tau\right] - \Phi(\tau)\right| \leq & \underbrace{\left|F_{n,t}\left(\frac{\sigma\tau - \sqrt{n}(\theta_t - \theta)}{\sigma_t}\right) - \Phi\left(\frac{\sigma\tau - \sqrt{n}(\theta_t - \theta)}{\sigma_t}\right)\right|}_{\text{(i)}} \\ & + \underbrace{\left|\Phi\left(\frac{\sigma\tau - \sqrt{n}(\theta_t - \theta)}{\sigma_t}\right) - \Phi(\tau)\right|}_{\text{(ii)}}.\end{aligned}$$

We show that both terms on the RHS converge to zero simultaneously. Let $\nu = \limsup_{t\to\infty} \nu_t$

and δ be a fixed constant such that $0 < \delta < \sigma$. Note that by the convergence assumptions on ν_t and σ_t, there exists $t_{\nu,\delta} \geq 1$ such that $\nu_t < \nu + \delta < \infty$ for any $t \geq t_{\nu,\delta}$; and there exists $t_{\sigma,\delta} \geq 1$ such that $\sigma_t > \sigma - \delta > 0$ for any $t \geq t_{\sigma,\delta}$. By the Berry-Esseen bound for U-statistics [Callaert and Janssen, 1978], for $t \geq \max(t_{\nu,\delta}, t_{\sigma,\delta})$ and any $n \geq 2$, the term (i) is bounded as

$$\begin{aligned}\text{(i)} &\leq \sup_{\tau'} \left| F_{n,t}(\tau') - \Phi(\tau') \right| \\ &\leq C \left(\frac{\nu_t}{\sigma_t} \right)^3 n^{-\frac{1}{2}} \\ &< C \left(\frac{\nu + \delta}{\sigma - \delta} \right)^3 n^{-\frac{1}{2}},\end{aligned}$$

where C is a universal constant. For the term (ii), by the continuity of Φ and our assumptions on $\sqrt{n}(\theta_t - \theta)$ and σ_t, we can make the term (ii) arbitrarily small. Formally, for $\epsilon/2 > 0$, we can take $N_{\epsilon/2} \geq 1$ such that the term (ii) is bounded by $\epsilon/2$ for any $n, t \geq N_{\epsilon/2}$. Thus, for any $\epsilon > 0$, choosing n and t such that

$$n \geq \max\left(\left(\frac{2C}{\epsilon} \cdot \frac{\nu + \delta}{\sigma - \delta} \right)^2, N_{\epsilon/2} \right),$$

and $t \geq \max(t_{\nu,\delta}, t_{\sigma,\delta}, N_{\epsilon/2})$, we have

$$\left| \Pi \left[\sqrt{n} \left(\frac{U_n^{(t)} - \theta}{\sigma} \right) < \tau \right] - \Phi(\tau) \right| \leq \text{(i)} + \text{(ii)} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Next, for the case $\sigma = 0$, consider the squared error $n\mathbb{E}(U_n^{(t)} - \theta)^2$, which is decomposed as

$$n\mathbb{E}(U_n^{(t)} - \theta)^2 = n\mathbb{E}(U_n^{(t)} - \theta_t)^2 + n(\theta_t - \theta)^2.$$

The first term is the variance of the U-statistic $U_n^{(t)}$, and so according to Hoeffding [1948, Eq. 5.18], we have, for any $n \geq 2$,

$$\begin{aligned}n\mathbb{E}(U_n^{(t)} - \theta_t)^2 &= \frac{(n-2)}{(n-1)} \underbrace{4\mathrm{Var}_{Y \sim \gamma_t}\left[\mathbb{E}_{Y' \sim \gamma_t}[h(Y, Y')] \right]}_{\sigma_t^2} + \frac{2}{(n-1)} \mathrm{Var}_{Y,Y' \sim \gamma_t \otimes \gamma_t}[h(Y, Y')] \\ &\leq \sigma_t^2 + \frac{2}{n-1} \mathrm{Var}_{Y,Y' \sim \gamma_t \otimes \gamma_t}[h(Y, Y')].\end{aligned}$$

We have $w = \limsup_{t \to \infty} \mathrm{Var}_{Y,Y' \sim \gamma_t \otimes \gamma_t}[h(Y, Y')] < \infty$ by the finiteness of (the limit supremum of) the third central moment, and $\sigma_t^2 \to \sigma^2 = 0$ by assumption. Therefore, for any $\epsilon > 0$, we can take $t_{\epsilon,v} \geq 1$ such that

$$\mathrm{Var}_{Y,Y' \sim \gamma_t \otimes \gamma_t}[h(Y, Y')] < w + 1, \text{and } \sigma_t^2 \leq \frac{\epsilon}{4},$$

for any $t \geq t_{\epsilon,v}$. Choosing $n \geq 4(w + 1)/\epsilon + 1$, we have $n\mathbb{E}(U_n^{(t)} - \theta_t)^2 \leq \epsilon/2$. For the second term, we can take $N_{\epsilon/2} \geq 1$ such that $n(\theta_t - \theta)^2 \leq \epsilon/2$ for any $n, t \geq N_{\epsilon/2}$. Thus, having

$$n, t \geq \max\Big(t_{\epsilon,v}, 4\frac{w+1}{\epsilon} + 1, N_{\epsilon/2}\Big)$$

leads to $n\mathbb{E}(U_n^{(t)} - \theta)^2 \leq \epsilon$. We have shown $n\mathbb{E}(U_n^{(t)} - \theta)^2 \to 0$, which implies $\sqrt{n}(U_n^{(t)} - \theta) \to 0$ in probability. □

3.6.2 Variance of a U-statistic

We first recall known facts about U-statistics. For an i.i.d. sample $\{y_i\}_{i=1}^n \sim R$, let us define a U-statistic

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} h(y_i, y_j),$$

where h is a symmetric measurable kernel. According to Hoeffding [1948, Eq. 5.18], the variance of U_n is

$$\mathrm{Var}[U_n] = \frac{4(n-2)}{n(n-1)}\zeta_1 + \frac{2}{n(n-1)}\zeta_2, \tag{3.10}$$

where $\zeta_1 = \mathrm{Var}_{y \sim R}\left[\mathbb{E}_{y' \sim R}\left[h(y, y')\right]\right]$ and $\zeta_2 = \mathrm{Var}_{y,y' \sim R \otimes R}\left[h(y, y')\right]$. To obtain the asymptotic variance of the $\sqrt{n}(U_n - \mathbb{E}[U_n])$, we only need the first term ζ_1 as $n\mathrm{Var}[U_n] \to 4\zeta_1$ $(n \to \infty)$, assuming $\mathbb{E}_{(y,y') \sim R \otimes R}[h(y, y')^2] < \infty$.

Recall that our test statistic in Section 3.3.3 requires a consistent estimator of the asymptotic variance in Corollary 3.6 (see also Theorem 3.4). To accommodate the setting of our test, we consider the following situation: we are given samples $\{y_i^{(t)}\}_{i=1}^n \sim R_t$ where $\{R_t\}_{t=1}^\infty$ is a sequence of distributions approximating R. This defines a sequence of U-statistics

$$U_n^{(t)} = \frac{1}{n(n-1)} \sum_{i \neq j} h\big(y_i^{(t)}, y_j^{(t)}\big),$$

and the variance $\mathrm{Var}[U_n^{(t)}]$ is given by the corresponding parameters $\zeta_{1,t}$ and $\zeta_{2,t}$ (see Eq. (3.10)). In the following, we address the estimation of $\sigma^2 = \lim_{t \to \infty} \sigma_t^2$ with $\sigma_t^2 = 4\zeta_{1,t}$, assuming that the limit exists. As Theorem 3.4 suggests, the quantity σ^2 is the asymptotic variance of $\sqrt{n}(U_n^{(t)} - \mathbb{E}[U_n])$.

In the main body, we define our test using the jackknife estimator

$$\sigma_{n,t} := \sqrt{(n-1)\sum_{i=1}^{n}\big(U_{n,-i}^{(t)} - U_n^{(t)}\big)^2} \tag{3.11}$$

(see also the test statistic $T_{n,t}$ in Section 3.3.3), where $U_{n,-i}^{(t)}$ the U-statistic computed on the sample with the i-th data point removed. In the following, we provide a consistency proof for this estimator.

Jackknife estimator and its consistency. We first revisit the definition of the jackknife variance estimator and its properties. For simplicity, we first drop the dependency on t. The jackknife estimator of the variance of a (scaled) U-statistic $\sqrt{n}(U_n - \mathbb{E}[U_n])$ is defined as

$$v_n^J = (n-1)\sum_{i=1}^{n}(U_{n,-i} - U_n)^2, \tag{3.12}$$

where $U_{n,-i}$ is the U-statistic computed with the sample with the i-th data point removed. According to Arvesen [1969, Eq. 25, see also Section 3.I], the jackknife estimator has the expansion

$$v_n^J = \sum_{c=0}^{2} a_{n,c}\hat{U}_c, \tag{3.13}$$

where

$$a_{n,c} = \frac{n-1}{n}\binom{n-1}{2}^{-2}\{n\mathbb{I}(c>0) - 4\}\binom{n}{c}\binom{n-c}{2-c}\binom{n-2}{2-c}$$

with $\mathbb{I}(\cdot)$ the indicator function; each term in the sum is a U-statistic

$$\hat{U}_c = \binom{n}{4-c}^{-1}\sum_{(\alpha,\beta,\gamma)\in C_{n,4-c}} h_{\mathrm{sym}}(y_{\alpha_1},\ldots,y_{\alpha_c},y_{\beta_1},\ldots,y_{\beta_{2-c}},y_{\gamma_1},\ldots,y_{\gamma_{2-c}}), \tag{3.14}$$

where the sum is over all combinations of $(4-c)$ integers chosen from $\{1,\ldots,n\}$. If indices end with 0 as in $(\alpha_1,\ldots,\alpha_0)$, it should be understood that the corresponding variables are omitted. The function $h_{\mathrm{sym}}(y_{\alpha_1},\ldots,y_{\alpha_c},y_{\beta_1},\ldots,y_{\beta_{2-c}},y_{\gamma_1},\ldots,y_{\gamma_{2-c}})$ in (3.14) is defined as a symmetric kernel

$$\sum_{\sigma\in\Sigma(\alpha,\beta,\gamma)}\frac{\tilde{h}(y_{\sigma(\alpha_1)},\ldots,y_{\sigma(\alpha_c)},y_{\sigma(\beta_1)},\ldots,y_{\sigma(\beta_{2-c})})\tilde{h}(y_{\sigma(\alpha_1)},\ldots,y_{\sigma(\alpha_c)},y_{\sigma(\gamma_1)},\ldots,y_{\sigma(\gamma_{2-c})})}{(4-c)!}$$

with $\tilde{h} = h - \mathbb{E}[U_n]$ and $\Sigma(\alpha,\beta,\gamma)$ the set of all permutations of given $4-c$ integers $(\alpha_1,\ldots,\alpha_c,\beta_1,\ldots,\beta_{2-c},\gamma_1,\ldots,\gamma_{2-c})$. Note that we have $\mathbb{E}[\hat{U}_c] = \zeta_c$ with $\zeta_0 = 0$. For large n, the coefficient $a_{n,c}$ behaves as

$$\begin{aligned} a_{n,c} &\approx 4\frac{1}{c!}\left(\frac{1}{(2-c)!}\right)^2 n^{1-c}\ (c\geq 1),\\ a_{n,0} &= O(1). \end{aligned}$$

Therefore, if $\mathbb{E}h(y,y')^2 < \infty$, the estimator v_n^J converges to the asymptotic variance $4\zeta_1$ of the U-statistic $\sqrt{n}(U_n - \mathbb{E}[U_n])$, a.s.

Next, we recover the dependency on t and define a jackknife estimator

$$v_{n,t}^J = (n-1)\sum_{i=1}^{n}\left(U_{n,-i}^{(t)} - U_n^{(t)}\right)^2,$$

which is the estimator in (3.12) computed on the sample $\{y_i^{(t)}\}_{i=1}^n$. The following lemma provides the required consistency.

Lemma 3.8. *Assume*

$$\limsup_{t\to\infty} \mathbb{E}_{(y_1,y_2)\sim R_t\otimes R_t}[h(y_1,y_2)^4] < \infty.$$

Let $\sigma^2 = \lim_{t\to\infty}\sigma_t^2$ *where* $\sigma_t^2 = 4\zeta_{1,t} = 4\mathrm{Var}_{y\sim R_t}\left[\mathbb{E}_{y'\sim R_t}\left[h(y,y')\right]\right]$. *Then, we have the double limit* $\mathbb{E}\left(v^J_{n,t}-\sigma^2\right)^2 \to 0$ *as* $n,t\to\infty$.

Proof. Note that we have the following relation

$$\begin{aligned}\mathbb{E}[(v^J_{n,t}-\sigma^2)^2] &= \mathbb{E}[(v^J_{n,t}-\mathbb{E}[v^J_{n,t}])^2] + (\mathbb{E}[v^J_{n,t}]-\sigma^2)^2\\ &= \underbrace{\mathbb{E}[(v^J_{n,t}-\mathbb{E}[v^J_{n,t}])^2]}_{\text{variance}} + \underbrace{\{\mathbb{E}[v^J_{n,t}]-\sigma_t^2\}^2}_{\text{(squared) bias}} -2\left(\mathbb{E}[v^J_{n,t}]-\sigma_t^2\right)\left(\sigma_t^2-\sigma^2\right) + \left(\sigma_t^2-\sigma^2\right)^2.\end{aligned} \tag{3.15}$$

The decomposition indicates that as long as the bias and the variance terms in (3.15) decay as $n,t\to\infty$, the estimator $v^J_{n,t}$ serves as a consistent estimator of σ^2 (note that we have $\sigma_t^2-\sigma^2\to 0$ by assumption). In the following, we show that the assertion holds.

For the bias term, note that by the decomposition (3.13), we have

$$\mathbb{E}[v^J_{n,t}] = \sigma_t + O(n^{-1})\zeta_{2,t}.$$

By the assumption on the fourth moment, the limit supremum of $\zeta_{2,t} = \mathrm{Var}_{R_t\otimes R_t}[h(y_1,y_2)^2]$ is finite. Therefore, for any $\epsilon>0$, we can take $N_{b,\epsilon}\geq 1$ such that $(\mathbb{E}[v^J_{n,t}]-\sigma_t)^2<\epsilon$ for $n,t\geq N_{b,\epsilon}$.

For the variance term, using the decomposition (3.13), we have

$$\begin{aligned}\mathbb{E}[(v^J_{n,t}-\mathbb{E}[v^J_{n,t}])^2] &= \sum_{c,c'=0}^{2} a_{c,n}a_{c',n}\mathrm{Cov}[\hat{U}_c^{(t)},\hat{U}_{c'}^{(t)}]\\ &\leq \left(\sum_{c=0}^{2} a_{c,n}\mathrm{Var}[\hat{U}_c^{(t)}]^{1/2}\right)^2,\end{aligned}$$

where $\hat{U}_c^{(t)}$ is the U-statistic $\hat{U}_c$ computed with the sample $\{y_i^{(t)}\}_{i=1}^n$. According to Serfling [2009, Section 5.2.1, Lemma A],

$$\mathrm{Var}[\hat{U}_c^{(t)}] \leq \frac{4-c}{n}\mathrm{Var}[h_{\mathrm{sym}}(y_{\alpha_1},\dots,y_{\alpha_c},y_{\beta_1},\dots,y_{\beta_{2-c}},y_{\gamma_1},\dots,y_{\gamma_{2-c}})],$$

where the variance is taken with respect to the product measure $\otimes_{i=1}^{4-c}R_t$. Note that the variance on the RHS is bounded as, for each $c\in\{0,1,2\}$,

$$\begin{cases}\mathrm{Var}[h_{\mathrm{sym}}(y_1y_2,y_3,y_4)] \leq \mathbb{E}\left[\tilde{h}_t(y_1,y_2)^4\right] & (c=0),\\ \mathrm{Var}[h_{\mathrm{sym}}(y_1,y_2,y_3)] \leq \mathbb{E}\left[\tilde{h}_t(y_1,y_2)^4\right] & (c=1), \text{ and}\\ \mathrm{Var}[h_{\mathrm{sym}}(y_1,y_2)] \leq \mathbb{E}[\tilde{h}_t(y_1,y_2)^4] & (c=2),\end{cases}$$

where $\tilde{h}_t = h - \mathbb{E}_{R_t\otimes R_t}[h(y_1,y_2)]$. By the assumption $\limsup_{t\to\infty}\mathbb{E}_{(y_1,y_2)\sim R_t\otimes R_t}[h(y_1,y_2)^4] <$

∞, the above observation implies that for each $c \in \{0, 1, 2\}$,

$$\limsup_{t\to\infty} \mathrm{Var}_{\otimes_{i=1}^{4-c} R_t}[h_{\mathrm{sym}}(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\beta_1}, \ldots, y_{\beta_{2-c}}, y_{\gamma_1}, \ldots, y_{\gamma_{2-c}})] < \infty.$$

Therefore, taking appropriate n, t diminishes $\mathrm{Var}[\hat{U}_c^{(t)}]$. In consequence, as for the bias term, we can take $N_{v,\epsilon} \geq 1$ such that $\mathbb{E}[(v_{n,t}^J - \mathbb{E}[v_{n,t}^J])^2] < \epsilon$ for $n, t \geq N_{v,\epsilon}$. Thus, the bias and variance terms can be made arbitrarily small by taking $n, t \geq \max(N_{v,\epsilon}, N_{b,\epsilon})$. □

3.A Kernel Stein discrepancy for bounded domains

We detail the case where data take values in a bounded open set $\mathcal{X} \subset \mathbb{R}^D$. For a density p, we require that $p \in \mathcal{C}(\bar{\mathcal{X}}) \cap C^1(\mathcal{X})$ with $\bar{\mathcal{X}}$ the closure of $\mathcal{X}$. The kernel is assumed to satisfy the same conditions everywhere; i.e., $k(x, \cdot) \in \mathcal{C}(\bar{\mathcal{X}}) \cap C^1(\mathcal{X})$. Additionally, we require the following conditions : (a) The domain $\mathcal{X}$ is assumed to be sufficiently regular – we assume that $\mathcal{X}$ is convex or C^1; (b) for any boundary point $x \in \partial\mathcal{X}$, $p(x)k(x, \cdot) = 0$; and (c) $\mathbb{E}_{x\sim P}\|\mathbf{s}_p(x)\|_2 < \infty$, $\mathbb{E}_{x\sim P}[k(x, x)^{1/2}] < \infty$, and $\mathbb{E}_{x\sim P}[k_{12}(x, x)^{1/2}] < \infty$. The first assumption allows us to use the divergence theorem, and the third one ensures the P-integrability of $\langle \mathbf{s}_p, f\rangle$ and $\langle \nabla, f\rangle$ [Steinwart and Christmann, 2008, Corollary 4.36]. By the divergence theorem, the second condition on the kernel implies that any $f = (f_1, \ldots f_D)$ of the corresponding RKHS $\mathcal{F}$ satisfies $\mathbb{E}_{x\sim P}[\mathcal{A}_P f(x)] = 0$ [Gorham and Mackey, 2015, Proposition 1]. We can fulfill the required condition by choosing a kernel k_φ defined by $k_\varphi(x, x') = \varphi(x)\varphi(x')k(x, x')$ where $\varphi : \mathcal{X} \to [0, \infty)$ such that $\varphi \in \mathcal{C}(\bar{\mathcal{X}}) \cap C^1(\mathcal{X})$ and $\varphi(x) = 0$ for $x \in \mathbb{R}^D \setminus \mathcal{X}$. The KSD is then defined similarly as in Section 3.2.

3.B Numerically stable Stein operator

As we note in Section 3.2, the Stein operator of Yang et al. [2018] can induce numerically unstable functions, as it contains the reciprocal $1/p(x)$. This appendix provides a more stable alternative using the Zanella-Stein operator proposed by Hodgkinson et al. [2020]. We focus on the Barker-Stein operator of Shi et al. [2022] below, but other operators can be considered depending on the application. For instance, when the size L of the discrete domain is small (e.g., binary domains such as adjacent matrices of graphs), we may use the Gibbs-Stein operator [Bresler and Nagaraj, 2019, Reinert and Ross, 2019, Shi et al., 2022], as it is manageable to perform the required conditional expectation.

3.B.1 Univariate case

We first consider the univariate case $\mathcal{X} = \{0, \ldots, L-1\}$ with $L > 1$, as the multivariate case builds on the construction here. Let be a density p on $\mathcal{X}$. The Zanella-Stein operator is defined by

$$\mathcal{A}_p^{\mathrm{Uni}} f(x) = \sum_{y\in\mathcal{N}_x} a\left(\frac{p(y)}{p(x)}\right)\left(f(y) - f(x)\right),$$

where the function $a : [0, \infty) \to [0, \infty)$ is referred to as a balancing function, which is assumed to satisfy $a(0) = 0$, and $a(r) = ra(1/r)$ for $r > 0$. The symbol $\mathcal{N}_x \subset \mathcal{X}$ denotes the neighborhood of x. The operator is derived from the infinitesimal generator of a Markov jump process with the jump rate given by a as a function of $p(y)/p(x)$. In the following, we use the Barker balancing function $a(r) = r/(1 + r)$; this results in

$$a\left(\frac{p(y)}{p(x)}\right) = \frac{\frac{p(y)}{p(x)}}{1 + \frac{p(y)}{p(x)}} = \frac{p(y)}{p(x) + p(y)}.$$

This choice of the balancing function was proposed by Shi et al. [2022] and is referred to as the Barker-Stein operator. The above ratio takes values between 0 and 1 (it saturates to one as the ratio $p(y)/p(x)$ gets large), and is thus numerically stable even when $p(x)$ is close to zero. The neighborhood $\mathcal{N}_x$ is chosen such that the process can transit from any starting point to any other point, and admits p as the invariant distribution [see Example 2 of Hodgkinson et al., 2020, for details]. In the following, $\mathcal{N}_x$ is assumed to be the two adjacent points of x with respect to the cyclic difference. This construction is not limited to $D = 1$, but it can be challenging to compute the sum in a high-dimensional space when applied to the KSD, since the number of neighbors typically grows with the dimension.

3.B.2 Multivariate case

We next consider the multivariate case $\mathcal{X} = \{0, \dots, L-1\}^D$ with $D > 1$. We follow the product space construction of a Stein operator [Hodgkinson et al., 2020, Proposition 2]. We define an operator $\mathcal{A}_p$ that acts on a D-dimensional vector-valued function $f = (f_1, \dots, f_D) : \mathcal{X} \to \mathbb{R}^D$ by

$$\mathcal{A}_p f(x) = \sum_{d=1}^{D} \mathcal{A}^{\mathrm{Uni}}_{p_d(\cdot|x_{-d})} \mathcal{P}^x_d f_d(x),$$

where p_d is the distribution given by conditioning on all but the d-th coordinate, $x_{-d} = (x_1, \dots, x_{d-1}, x_{d+1}, \dots x_D)$, and $\mathcal{P}^x_d$ is the projection defining a function on the d-th coordinate by freezing all the other coordinates; i.e.,

$$\mathcal{P}^x_d f : y \in \{0, \dots, L-1\} \mapsto f(x_1, \dots x_{d-1}, y, x_{d+1}, \dots, x_D).$$

We can define the KSD using a vector-valued RKHS $\mathcal{F} = \prod_{d=1}^{D} \mathcal{F}_k$ with $\mathcal{F}_k$ the RKHS of a scalar kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. With this choice, we can define the KSD as in 3.2.1; in particular, for this operator, we have

$$\mathrm{KSD}^2\,(P\|R) = \mathbb{E}_{x,x' \sim R \otimes R}[h_p(x, x')]$$

where

$$h_p(x, y) = \sum_{d=1}^{D} \mathcal{A}^{\mathrm{Uni}}_{x, p_d(\cdot|x_{-d})} \otimes \mathcal{A}^{\mathrm{Uni}}_{y, p_d(\cdot|y_{-d})} k(x, y)$$

for any $x, y \in \mathcal{X}$, with $\mathcal{A}^{\mathrm{Uni}}_{*,p_d(\cdot|x_{-d})}$ acting on $*$. Specifically, the Stein kernel h_p is given by

$$h_p(x, y) = \sum_{d=1}^{D} \sum_{\nu \in \mathcal{N}_{x,d}} \sum_{\tilde{\nu} \in \mathcal{N}_{y,d}} a_\nu(x) a_\nu(y) \{k(\nu(x), \tilde{\nu}(y)) + k(x, y) - k(x, \tilde{\nu}(y)) - k(\nu(x), y)\}.$$

Here, $\mathcal{N}_{x,d}$ denotes the set of neighborhood points with respect to the dth coordinate; we identify this as a set of functions, each of which maps x to the corresponding neighboring point. The weight a_ν is defined by

$$a_\nu(x) = a\left(\frac{p(\nu(x))}{p(x)}\right).$$

From this expression, we can conclude that the KSD is possible to estimate as long as we can evaluate a_ν. Note that a single evaluation of the Stein kernel h_p requires $O(DN)$ with N being the maximum of the sizes of the neighborhood $\mathcal{N}_{x,d}$ (in our case, $N = 2$).

3.B.3 Application to latent variable models

The KSD defined above admits essentially the same treatment as in the main body. The Stein operator above requires evaluating the marginal density $p(x) = \int p(x|z) P_Z(\mathrm{d}z)$. Following the approach in Section 3.2.2, we can circumvent this issue as

$$\begin{aligned} a_\nu(x) &= \frac{p(\nu(x))}{p(x) + p(\nu(x))} \\ &= \int \frac{p(\nu(x)|z)}{p(x|z) + p(\nu(x)|z)} \frac{p(x|z) + p(\nu(x)|z)}{p(x) + p(\nu(x))} P_Z(\mathrm{d}z). \end{aligned}$$

The weight $a_\nu(x)$ can be estimated by sampling from the *modified* posterior

$$P_{Z,\nu}(\mathrm{d}z|x) \propto \{p(x|z) + p(\nu(x)|z)\} P_Z(\mathrm{d}z).$$

Given a sample $\mathbf{z} = \{z_i\}_{i=1}^m$ for simulating $P_{Z,\nu}(\mathrm{d}z)$, we can estimate $a_\nu(x)$ by

$$a_\nu(x|\mathbf{z}) := \frac{1}{m} \sum_{i=1}^{m} \frac{p(\nu(x)|z_i)}{p(x|z_i) + p(\nu(x)|z_i)},$$

which is possible to evaluate as long as the likelihood $p(x|z)$ is tractable. In contrast to the difference operator KSD, this approach requires simulating $O(D)$ posterior distributions, as each dimension generates distinct modified posterior distributions. Thus, while numerically stable, this approach is computationally more expensive. The user might want to consider this limitation when they choose between the operator defined here and the difference Stein operator considered in the main body.

3.C Integrability condition for the KSD expression

We give sufficient conditions for the integrability condition

$$\mathbb{E}_{(x,z),(x,z')\sim\tilde{R}\otimes\tilde{R}}|H_p[(x,z),(x',z')]| < \infty. \tag{3.16}$$

By the triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}_{(x,z),(x,z')\sim\tilde{R}\otimes\tilde{R}}|H_p[(x,z),(x',z')]| \\
&\leq \mathbb{E}_{(x,z),(x,z')\sim\tilde{R}\otimes\tilde{R}}\left\{\left|\mathbf{s}_p(x|z)^\top\mathbf{s}_p(x'|z')k(x,x')\right| + \left|\mathbf{s}_p(x|z)^\top k_1(x',x)\right|\right. \\
&\qquad\qquad\left. + \left|k_1(x,x')^\top\mathbf{s}_p(x'|z')\right| + \left|k_{12}(x,x')\right|\right\}. \\
&\leq \mathbb{E}_{(x,x')\sim R\otimes R}\left\{\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2\mathbb{E}_{z'|x'}\left\|\mathbf{s}_p(x'|z')\right\|_2 k(x,x') + \mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2\left\|k_1(x',x)\right\|_2\right. \\
&\qquad\qquad\left. + \left\|k_1(x,x')\right\|_2\mathbb{E}_{z'|x'}\left\|\mathbf{s}_p(x'|z')\right\|_2 + \left|k_{12}(x,x')\right|\right\}.
\end{aligned}$$

From this, the integrability condition is satisfied if

1. $\mathbb{E}_{(x,x')\sim R\otimes R}\left[\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2\mathbb{E}_{z'|x'}\left\|\mathbf{s}_p(x'|z')\right\|_2 k(x,x')\right] < \infty$,
2. $\mathbb{E}_{(x,x')\sim R\otimes R}\left[\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2\left\|k_1(x',x)\right\|_2\right] < \infty$, and
3. $\mathbb{E}_{(x,x')\sim R\otimes R}\left|k_{12}(x,x')\right| < \infty$.

Note that for a finite domain $\mathcal{X} = \{0,\cdots,L-1\}^D$, these conditions are trivial, as

$$\begin{aligned}
\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2 &= \int\frac{\|\Delta_x p(x,z)\|_2}{p(x,z)}\frac{p(x,z)}{p(x)}\mathrm{d}P_Z(z) \\
&\leq \frac{2}{p(x)} \leq \max_{x\in\mathcal{X}}\frac{2}{p(x)}.
\end{aligned}$$

In what follows, we consider continuous-valued x. As mentioned in the main body, the third condition is mild, and fulfilled by e.g., the exponentiated quadratic kernel. Unfortunately we do not have a handy test for the other requirements, and therefore deal with specific scenarios below. To this end, we clarify the growth of $\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2$ as a function of x so that the user can check the required conditions above.

Exponential families with bounded natural parameters. Let us first consider an exponential family likelihood $p(x|z)\propto\exp\left(\sum_{s=1}^S T_s(x)\eta_s(z)\right)$, $T_s:\mathbb{R}^D\to\mathbb{R}$, $\eta_s:\mathcal{Z}\to\mathbb{R}$ for $1\leq s\leq S$. For this likelihood, we have

$$\begin{aligned}
\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2 &= \mathbb{E}_{z|x}\left\|\sum_{s=1}^S\eta_s(z)\nabla_x T_s(x)\right\|_2 \\
&\leq \sum_{s=1}^S\left\|\nabla_x T_s(x)\right\|_2\mathbb{E}_{z|x}|\eta_s(z)|.
\end{aligned}$$

The conditions concerning the score function are satisfied provided that we have

1. $\mathbb{E}_{(x,x')\sim R\otimes R}\left[\|\nabla_x T_s(x)\|_2\, \mathbb{E}_{z|x}|\eta_s(z)|\, \|\nabla_x T_{s'}(x')\|\, \mathbb{E}_{z'|x'}|\eta_{s'}(z)| k(x,x')\right] < \infty$ for any $s, s' \in \{1, ..., S\}$.
2. $\mathbb{E}_{(x,x')\sim R\otimes R}\left[\|\nabla_x T_s(x)\|_2\, \mathbb{E}_{z|x}|\eta_s(z)|\, \|k_1(x', x)\|_2\right] < \infty$ for any $s \in \{1, ..., S\}$.

Let $a(x) := \|\nabla_x T_s(x)\|_2\, \mathbb{E}_{z|x}|\eta_s(z)|$. These conditions can be verified if both the kernel and its derivative decay faster than $a(x)$. This can be challenging in practice as we have a posterior expectation in $a(x)$ whose dependency on x may not be easily analyzed. If we restrict the likelihood to have bounded parameters (i.e., $\eta_s(z)$ is bounded), then the posterior expectation is bounded, so we only need to choose a kernel such as

$$k(x,x') = \frac{1}{\sqrt{1+\sum_{s=1}^S \|\nabla_x T_s(x)\|^2}} \frac{1}{\sqrt{1+\sum_{s=1}^S \|\nabla_x T_s(x')\|_2^2}} \kappa(x,x')$$

for a given kernel κ. We summarize this observation in the following lemma:

Lemma 3.9. *Consider a latent variable model with likelihood $p(x|z) \propto \exp\left(\sum_{s=1}^S T_s(x)\eta_s(z)\right)$, $T_s : \mathbb{R}^D \to \mathbb{R}$, $\eta_s : \mathcal{Z} \to \mathbb{R}$ for $1 \le s \le S$ and an arbitrary prior P_Z. If $\sup_{z\in\mathcal{Z}} n_s(z) < \infty$, then*

$$\mathbb{E}_{z|x} \|\mathbf{s}_p(x|z)\|_2 \le \max_s \sup_{z\in\mathcal{Z}} n_s(z) \cdot \sum_{s=1}^S \|\nabla_x T_s(x)\|_2 .$$

Furthermore, for a given bounded kernel κ, reweighting it by

$$k(x,x') = \left(1+\sum_{s=1}^S \|\nabla_x T_s(x)\|^2\right)^{-\delta} \left(1+\sum_{s=1}^S \|\nabla_x T_s(x')\|^2\right)^{-\delta} \kappa(x,x')$$

with some $\delta \ge 1/2$ ensures that the condition (3.16) holds.

The boundedness assumption on the natural parameter $\eta_s(z)$ may not be satisfied for some models. For instance, if we consider a normal mixture model with prior on the mean of the mixture component, the support of the prior could be unbounded.

Location-scale mixtures. Alternatively, we consider a location-scale family given by a radial-basis function $\psi : [0, \infty) \to (0, \infty)$,

$$p(x|z = (\mu, \sigma^2)) \propto \frac{1}{\sigma^D} \psi\left(\frac{\|x-\mu\|_2^2}{\sigma^2}\right),$$

with prior $P_{\mu,\sigma}$ placed on the parameters. Here, we make the following assumptions:

Assumption 3.10.

The function ψ is monotonically decreasing. The derivative-to-function ratio $|\psi'/\psi|$ is uniformly upper-bounded by a constant $M_\psi > 0$.

Assumption 3.11.

The prior satisfies $\int \|\mu\|_2^2/\sigma^4 \mathrm{d}P_{\mu,\sigma}(\mu,\sigma) < \infty$ and $\int \sigma^{-4}\mathrm{d}P_\sigma(\sigma) < \infty$. Furthermore, it satisfies

$$\int \left(\frac{\psi(\|x-\mu\|_2^2/\sigma^2)}{\sigma^D}\right)^2 \mathrm{d}P_{\mu,\sigma}(\mu,\sigma) < \infty.$$

Isotropic normal- and Student's t-densities satisfy Assumption 3.10. As ψ is monotonically decreasing, the third condition in Assumption 3.11 can be verified alternatively by

$$\int \frac{1}{\sigma^D}\mathrm{d}P_\sigma(\sigma) < \infty.$$

These assumptions effectively prevent the density from being *peaky* and thus control the growth of the score function.

Under these assumptions, we can quantify the growth of the score function as follows.

Lemma 3.12. *Consider a latent variable model having likelihood*

$$p(x|z=(\mu,\sigma^2)) \propto \frac{1}{\sigma^D}\psi\left(\frac{\|x-\mu\|_2^2}{\sigma^2}\right)$$

with radial-basis function $\psi : [0,\infty) \to (0,\infty)$ *and prior* $P_{\mu,\delta}$. *Under Assumptions 3.10, 3.11, in the limit of* $\|x\|_2 \to \infty$, *we have*

$$\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2 = O(\|x\|_2).$$

Furthermore, for a given bounded kernel κ, *reweighting it by*

$$k(x,x') = \left(1+\|x\|_2^2\right)^{-\delta}\left(1+\|x'\|_2^2\right)^{-\delta}\kappa(x,x')$$

with some $\delta \geq 1/2$ *ensures that the condition (3.16) holds.*

Proof. First, note that we have

$$\begin{aligned}
&\mathbb{E}_{z|x}\left\|\mathbf{s}_p(x|z)\right\|_2 \\
&= \int \frac{2}{\sigma^{D+2}}\|x-\mu\|_2\left|\frac{\psi'(\|x-\mu\|_2^2/\sigma^2)}{\psi(\|x-\mu\|_2^2/\sigma^2)}\right| \frac{\psi(\|x-\mu\|_2^2/\sigma^2)}{\int \frac{1}{\sigma^D}\psi\left(\|x-\mu\|_2^2/\sigma^2\right)\mathrm{d}P_{\mu,\sigma}(\mu,\sigma)}\mathrm{d}P_{\mu,\sigma}(\mu,\sigma) \\
&\leq 2M_\psi \underbrace{\frac{\int \frac{1}{\sigma^{D+2}}\|x-\mu\|_2\psi(\|x-\mu\|_2^2/\sigma^2)\mathrm{d}P_{\mu,\sigma}(\mu,\sigma)}{\int \frac{1}{\sigma^D}\psi\left(\|x-\mu\|_2^2/\sigma^2\right)\mathrm{d}P_{\mu,\sigma}(\mu,\sigma)}}_{g(x)}.
\end{aligned}$$

We therefore show that the growth of the function g is of the order of $\|x\|_2$ by examining the limit

$$\lim_{\|x\|_2\to\infty}\frac{g(x)}{\|x\|_2}.$$

Let $\tilde{p}_x(\mu,\sigma) = \psi(\|x-\mu\|_2^2/\sigma^2)/\sigma^D$. For the numerator of g, note that we have

$$\int \frac{\|x-\mu\|_2}{\sigma^2}\frac{\psi(\|x-\mu\|_2^2/\sigma^2)}{\sigma^D}\mathrm{d}P_{\mu,\sigma}(\mu,\sigma) \leq \sqrt{\int \left(\|x-\mu\|_2/\sigma^2\right)^2\mathrm{d}P_{\mu,\sigma}(\mu,\sigma)} \cdot \|\tilde{p}_x\|_{L_2(P_{\mu,\sigma})}$$

$$\overset{\text{Asm.3.11}}{<} \infty,$$

where the first inequality is given by the Cauchy-Schwartz inequality. Thus,

$$\begin{aligned}
\frac{g(x)}{\|x\|_2} &\leq \frac{\sqrt{\int \left(\|x-\mu\|_2/\sigma^2\right)^2 \mathrm{d}P_{\mu,\sigma}(\mu,\sigma)} \cdot \|\tilde{p}_x\|_{L_2(P_{\mu,\sigma})}}{\|x\|_2 \|\tilde{p}_x\|_{L_1(P_{\mu,\sigma})}} \\
&\leq \sqrt{\frac{\int \left(\|x-\mu\|_2/\sigma^2\right)^2 \mathrm{d}P_{\mu,\sigma}(\mu,\sigma)}{\|x\|_2^2}} \\
&\leq \sqrt{\int \frac{(\|x\|_2+\|\mu\|_2)^2}{\|x\|_2^2} \frac{1}{\sigma^4} \mathrm{d}P_{\mu,\sigma}(\mu,\sigma)}. \qquad (3.17)
\end{aligned}$$

The second line is obtained with the relation between the L_1- and L_2-norms: $\|\tilde{p}_x\|_{L_1(P_{\mu,\sigma})} \leq \|\tilde{p}_x\|_{L_2(P_{\mu,\sigma})}$. The function inside the integral in (3.17) is monotonically decreasing at each (μ, σ) as

$$\lim_{\|x\|_2 \to \infty} \left(1 + 2\frac{\|\mu\|_2}{\|x\|_2} + \frac{\|\mu\|_2^2}{\|x\|_2^2}\right) \frac{1}{\sigma^4} \searrow \frac{1}{\sigma^4},$$

and by Assumption 3.11, it is integrable when $\|x\|_2 = 1$. Thus, by the monotone convergence theorem, we have

$$\lim_{\|x\|_2 \to \infty} \frac{g(x)}{\|x\|_2} \leq \sqrt{\int \frac{1}{\sigma^4} \mathrm{d}P_\sigma(\sigma)},$$

where the upper-bound is finite by Assumption 3.11, indicating that $g(x) = O(\|x\|_2)$. Therefore, for the location-scale family, we have that the score part grows at the speed at most of $\|x\|_2$. Therefore, modifying kernel κ by

$$k(x, x') = \left(1 + \|x\|_2^2\right)^{-\delta} \left(1 + \|x'\|_2^2\right)^{-\delta} \kappa(x, x') \text{ with } \delta \geq \frac{1}{2}$$

ensures that the decay of the kernel is as fast as the score part. □

3.D Convergence assumption in the asymptotic normality proof

To apply Theorem 3.4 to the KSD estimate, we need to verify that the bias

$$\left| \mathbb{E}_{y,y' \sim \tilde{R}_t \otimes \tilde{R}_t} \bar{H}_p(y, y') - \mathbb{E}_{y,y' \sim \tilde{R} \otimes \tilde{R}} H_p(y, y') \right| \qquad (3.18)$$

decays at some rate. Here, $\tilde{R}_t(\mathrm{d}(x, \mathbf{z})) = P_Z^{(t)}(\mathrm{d}\mathbf{z}|x) R(\mathrm{d}x)$ (see the paragraph before Theorem 3.4 for the notation).

In what follows, we assume that the MCMC sampler satisfies

$$\left| \mathbb{E}_{\mathbf{z}^{(t)}|x} \bar{\mathbf{s}}_{p,d}(x|\mathbf{z}^{(t)}) - \mathbf{s}_{p,d}(x) \right| \leq r(t, x) := M(x) r(t), \text{ for } d = 1, \ldots, D, \qquad (3.19)$$

for some function $r(t) : \mathbb{N} \to (0, 1]$ decreasing to 0 in t and a positive function $M(x)$, where $\bar{\mathbf{s}}_{p,d}$ denotes the d-th element of the conditional score $\bar{\mathbf{s}}_p$ (the same rule applies to $\mathbf{s}_p$). There is usually dependency in $M(x)$ on the initial state of the chain, but this is suppressed here for simplicity. This assumption can be understood as the convergence of the t-step transition law of the Markov chain $P_Z^{(t)}(\mathrm{d}z|x)$ to the target $P_Z(\mathrm{d}z|x)$ in terms of the test functions $\mathbf{s}_{p,d}(x|\cdot)$. The convergence can then be checked with the assumption that the test functions belong to a certain class, and the upper-bound of (3.19) can be stated as a worst-case convergence rate in that class. For example, the convergence in total variation distance corresponds to the class of nonnegative measurable function bounded uniformly by 1.

A specification of the decay rate r and the function class relates the condition (3.19) to standard notions of ergodicity studied in the Markov chain literature [Roberts and Rosenthal, 2004, Meyn et al., 2009]. For instance, suppose that we have the geometric rate $r(t) = \rho^t$ for some $0 < \rho < 1$, and that the function class is such that all members are measurable and bounded uniformly by a function $V : \mathcal{Z} \to [1, \infty)$. Then, the corresponding convergence is known as V-uniform convergence (if $V \equiv 1$, this corresponds to the uniform ergodicity) [Meyn et al., 2009, Chapter 16].

We can reduce the convergence of the bias (3.18) to that of the score estimate (3.19). To see this assertion, we first note that

$$\begin{aligned}
&\left|\text{The first term of } \left(\mathbb{E}_{y,y'\sim\tilde{R}_t\otimes\tilde{R}_t}\bar{H}_p(y,y') - \mathbb{E}_{y,y'\sim\tilde{R}\otimes\tilde{R}}\bar{H}_p(y,y')\right)\right| \\
&\le \mathbb{E}_{x,x'}\left|k(x,x')\sum_{d=1}^{D}\left\{\mathbb{E}_{\mathbf{z}^{(t)}|x}\bar{\mathbf{s}}_{p,d}(x|\mathbf{z}^{(t)})\mathbb{E}_{\mathbf{z}'^{(t)}|x'}\bar{\mathbf{s}}_{p,d}(x'|\mathbf{z}'^{(t)}) - \mathbf{s}_{p,d}(x)\mathbf{s}_{p,d}(x')\right\}\right| \\
&\le \sum_{d=1}^{D}\mathbb{E}_{x,x'}\left[k(x,x')\left\{\left|\mathbb{E}_{\mathbf{z}^{(t)}|x}\bar{\mathbf{s}}_{p,d}(x|\mathbf{z}^{(t)})\right| r(t,x') + r(t,x)\left|\mathbf{s}_{p,d}(x')\right|\right\}\right] \\
&\le \sum_{d=1}^{D}\mathbb{E}_{x,x'}\left[k(x,x')\left\{r(t,x)r(t,x') + r(t,x')\left|\mathbf{s}_{p,d}(x)\right| + r(t,x)\left|\mathbf{s}_{p,d}(x')\right|\right\}\right] \\
&\le Dr(t)^2\mathbb{E}_{x,x'}[k(x,x')M(x)M(x')] + 2r(t)\sum_{d=1}^{D}\mathbb{E}_{x.x'}\left[k(x,x')M(x')\left|\mathbf{s}_{p,d}(x')\right|\right].
\end{aligned}$$

That is, if $\mathbb{E}_{x,x'}[k(x,x')M(x)M(x')] < \infty$ and $\sum_{d=1}^{D}\mathbb{E}_{x.x'}\left[k(x,x')M(x')\left|\mathbf{s}_{p,d}(x')\right|\right] < \infty$, the difference in the first terms can be bounded by $r(t)$. A similar argument can be applied to the second and third terms. The constant $M(x)$ in the bound in (3.19) often depends on certain properties of the target $P_Z(\mathrm{d}z|x)$. If those properties hold uniformly over x (i.e., $M(x)$ can be treated as a constant), then the validation of these conditions is straightforward. A concrete example of such properties is strong log-convexity and having a Lipschitz continuous gradient of the target (assuming the target is given by a density $p(z|x)$) [Dalalyan, 2017, Dwivedi et al., 2019, Bou-Rabee et al., 2020]. In this situation, the bias (3.18) is determined by the rate $r(t)$ at which the score function converges. Hence, t has to grow as a function of n such that $t(n) = O\{r^{-1}(n^{-s})\}$ with $s > 1/2$ to apply Theorem 3.4.

3.E The maximum mean discrepancy relative goodness-of-fit test

We provide the detail of the MMD relative goodness-of-fit test proposed by Bounliphone et al. [2016] and describe our modification to correct for underestimation of the variances required in the test (see Appendix 3.J.3).

Given a sample $\{z_i\}_{i=1}^{n_R} \overset{\text{i.i.d.}}{\sim} R$ and two competing models P, Q, the relative MMD test is defined as follows:

$$H_0 : \mathrm{MMD}(P, R) \le \mathrm{MMD}(Q, R) \text{ (null hypothesis)},$$
$$H_1 : \mathrm{MMD}(P, R) > \mathrm{MMD}(Q, R) \text{ (alternative)}.$$

Note that here we use a different symbol for the data in the main body (there, the data variable is x). Their procedure does not consider the case where the sample size n_R does not match the sizes n_P, n_Q of respective samples $\{x_i\}_{i=1}^{n_P}$, $\{y_i\}_{i=1}^{n_Q}$ from P and Q. Therefore, we provide the test procedure accommodating this case.

The test statistic is defined by the difference between estimates of the squared MMDs

$$\begin{aligned}&\widehat{\mathrm{MMD}^2}(P, R) - \widehat{\mathrm{MMD}^2}(Q, R)\\ &= \frac{1}{\binom{n_P}{2}} \frac{1}{\binom{n_Q}{2}} \frac{1}{\binom{n_R}{2}} \sum_{i_{x1}<i_{x2}} \sum_{i_{y1}<i_{y2}} \sum_{i_{z1}<i_{z2}} \ell_{\text{diff}}(x_{i_{x1}}, x_{i_{x2}}; y_{i_{y1}}, y_{i_{y2}}; z_{i_{z1}}, z_{i_{z2}}),\end{aligned}$$

where

$$\ell_{\text{diff}}(x, x'; y, y'; z, z') = \ell(x, x'; z, z') - \ell(y, y'; z, z'),$$

and ℓ is the kernel from Section 2.2. This statistic is a three-sample U-statistic; its asymptotic distribution is normal under the same assumptions on the relations between the models and data distribution as in the main body (the second paragraph of Section 3.3.1). Let $n_{\text{sum}} = n_P + n_Q + n_R$. Assume the following growth condition on the sample sizes

$$\frac{n_{\text{sum}}}{n_P} \to \rho_P^2, \ \frac{n_{\text{sum}}}{n_Q} \to \rho_Q^2, \text{ and } \frac{n_{\text{sum}}}{n_R} \to \rho_R^2$$

with finite constants ρ_P, ρ_Q, and ρ_R. Assume $\mathbb{E}[\ell_{\text{diff}}(x, x'; y, y'; z, z')^2] < \infty$. Then, according to Kowalski and Tu [2007, Theorem 3, p.168], the limit of (n_P, n_Q, n_R) gives

$$\begin{aligned}&\sqrt{n_{\text{sum}}} \left[\left\{\widehat{\mathrm{MMD}^2}(P, R) - \widehat{\mathrm{MMD}^2}(Q, R)\right\} - \{\mathrm{MMD}^2(P, R) - \mathrm{MMD}^2(Q, R)\}\right]\\ &\xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2_{P,Q,R}),\end{aligned}$$

where

$$\sigma^2_{P,Q,R} = 4 \left(\rho_P^2 \mathrm{Var}_x \left[\mathbb{E}[\ell_{\text{diff}}|x]\right] + \rho_Q^2 \mathrm{Var}_y \left[\mathbb{E}[\ell_{\text{diff}}|y]\right] + \rho_R^2 \mathrm{Var}_z \left[\mathbb{E}[\ell_{\text{diff}}|z]\right]\right),$$

with

$$\mathbb{E}[\ell_{\text{diff}}|x] = \mathbb{E}[\ell_{\text{diff}}(x, x'; y, y'; z, z')|x],$$

and the same notation applies to the conditional expectations of ℓ_{diff} of y and z.

With a consistent estimator $\hat{\sigma}_{P,Q,R}$, Bounliphone et al. [2016] proposed an asymptotically level-α test that rejects the null hypothesis if $\widehat{\text{MMD}}^2(P,R)-\widehat{\text{MMD}}^2(Q,R) \geq (\hat{\sigma}_{P,Q,R}/\sqrt{n_{\text{sum}}})\cdot \tau_{1-\alpha}$ with $\tau_{1-\alpha}$ the $(1-\alpha)$-quantile of the standard normal distribution. We found that the estimator given by Bounliphone et al. [2016] tends to underestimate the target variance, and that their test exceeded the nominal level in some problems where two models are close to each other. We therefore consider another estimator for our experiments, as described below.

Variance estimators Following are the variances required for $\sigma^2_{P,Q,R}$:

$$\begin{aligned}
\text{Var}_x\left[\mathbb{E}[\ell_{\text{diff}}|x]\right] &= \text{Var}_x\left[\mathbb{E}[\ell(x,x';z,z')|x]\right] \\
&= \text{Var}_x\left[\mathbb{E}_{x',z}[k(x,x')-k(z,x)|x]\right], \\
\text{Var}_y\left[\mathbb{E}[\ell_{\text{diff}}|y]\right] &= \text{Var}_y\left[\mathbb{E}[\ell(y,y';z,z')|y]\right] \\
&= \text{Var}_y\left[\mathbb{E}_{z,y'}[k(y,y')-k(z,y)|y]\right], \text{ and} \\
\text{Var}_z\left[\mathbb{E}[\ell_{\text{diff}}|z]\right] &= \text{Var}_z\left[\mathbb{E}[\ell(x,x';z,z')-\ell(y,y';z,z')|z]\right] \\
&= \text{Var}_z\left[\mathbb{E}_{x,y}[k(z,x)-k(z,y)|z]\right].
\end{aligned}$$

The first two quantities are symmetric in terms of x and y, and therefore we only need to consider one of them. An estimator for the first variance is given by

$$\text{Var}_x\left[\underbrace{\mathbb{E}_{x',z}[k(x,x')-k(z,x)|x]}_{f(x)}\right] \approx \frac{1}{n_P(n_P-1)}\sum_{i\neq j}\frac{\left(\hat{f}_i(x_i)-\hat{f}_j(x_j)\right)^2}{2}, \tag{3.20}$$

where $\hat{f}_i(x_i)$ is an approximation to $f(x_i)$ defined by

$$\hat{f}_i(x_i) = \frac{1}{(n_P-1)}\sum_{l\neq i}k(x_l,x_i) - \frac{1}{n_R}\sum_{l=1}^{n_R}k(z_l,x_i).$$

We similarly estimate the third variance using

$$\text{Var}_z\left[\underbrace{\mathbb{E}_{x,y}[k(z,x)-k(z,y)|z]}_{g(z)}\right] \approx \frac{1}{n_R(n_R-1)}\sum_{i\neq j}\frac{\left(\hat{g}_i(z_i)-\hat{g}_j(z_j)\right)^2}{2}, \tag{3.21}$$

where

$$\hat{g}_i(z_i) = \frac{1}{n_P}\sum_{l=1}^{n_P}k(x_l,z_i) - \frac{1}{n_Q}\sum_{l=1}^{n_Q}k(y_l,z_i)\ (\approx g(z_i)).$$

The estimators (3.20) and (3.21) are simple to compute and always nonnegative. The consistency of these estimators can be checked by expanding the expressions. The derivation is tedious, and therefore we only prove it for (3.20).

Lemma 3.13. *Assume* $\mathbb{E}_{x,x'\sim P\otimes P}[k(x,x')^2] < \infty$ *and* $\mathbb{E}_{x,x'\sim P\otimes R}[k(x,z)^2] < \infty$. *Then, Eq.*

(3.20) estimates

$$\mathrm{Var}_x\left[\mathbb{E}_{x',z}[k(x,x')-k(z,x)|x]\right]$$

consistently in the limit of (n_P, n_R) such that the ratio n_P/n_R converges to a finite constant.

Lemma 3.14. *Assume that the following quantities are finite:*

$$\mathbb{E}_{x,z\sim Q\otimes R}[k(x,z)^2],\ \mathbb{E}_{y,z\sim P\otimes R}[k(y,z)^2],\ \textit{and}\ \mathbb{E}_{z,z'\sim R\otimes R}[k(z,z')^2].$$

Then, Eq. (3.21) estimates

$$\mathrm{Var}_z\left[\mathbb{E}_{x,y}[k(z,x)-k(z,y)|z]\right]$$

consistently in the limit of (n_P, n_Q, n_R) such that the ratios n_P/n_R and n_Q/n_R converge to finite constants.

Proof. Note that the estimator (3.20) is the (approximate) sample variance

$$\frac{1}{n_P-1}\left\{\sum_{i=1}^{n_P}\hat{f}_i(x_i)^2 - n_P\left(\frac{1}{n_P}\sum_{i=1}^{n_P}\hat{f}_i(x_i)\right)^2\right\}.$$

Showing the consistency is equivalent to proving the following limits (the symbol $\xrightarrow{\mathrm{P}}$ denotes convergence in probability):

$$\frac{1}{n_P}\sum_{i=1}^{n_P}\hat{f}_i(x_i) \xrightarrow{\mathrm{P}} \mathbb{E}_x[f(x)] \text{ and } \frac{1}{n_P}\sum_{i=1}^{n_P}\hat{f}_i(x_i)^2 \xrightarrow{\mathrm{P}} \mathbb{E}_x[f(x)^2].$$

The first limit is immediate as

$$\frac{1}{n_P}\sum_{i=1}^{n_P}\hat{f}_i(x_i) = \frac{1}{n_P(n_P-1)}\sum_{l\neq i}k(x_l,x_i) - \frac{1}{n_P}\frac{1}{n_R}\sum_{l=1}^{n_R}k(z_l,x_i),$$

which is a U-statistic of $\mathbb{E}_x[f(x)]$.

For the second convergence claim, we expand the expressions as follows:

$$\begin{aligned}\hat{f}_i(x_i)^2 &= \frac{1}{(n_P-1)^2}\sum_{j\neq i}\sum_{l\neq i}k(x_j,x_i)k(x_l,x_i) + \frac{1}{n_R^2}\sum_{j,j'}k(z_j,x_i)k(z_{j'},x_i)\\ &\quad - 2\frac{1}{(n_P-1)n_R}\sum_{j\neq i}\sum_{l=1}^{n_R}k(x_j,x_i)k(z_l,x_i)\\ &= \frac{(n_P-2)}{(n_P-1)}\frac{1}{(n_P-1)(n_P-2)}\sum_{j\neq i}\sum_{l\neq i,j}k(x_j,x_i)k(x_l,x_i)\\ &\quad \underbrace{-\frac{1}{(n_P-1)^2}\sum_{j\neq i}k(x_j,x_i)^2}_{\mathrm{A}} + \frac{1}{n_R^2}\sum_{j\neq j'}k(z_j,x_i)k(z_{j'},x_i)\end{aligned}$$

$$+\underbrace{\frac{1}{n_R^2}\sum_j k(z_j,x_i)^2}_{\text{B}} -2\frac{1}{(n_P-1)n_R}\sum_{j\neq i}\sum_{l=1}^{n_R} k(x_j,x_i)k(z_l,x_i), \text{ and}$$

$$\mathbb{E}[f(x_i)^2] = \mathbb{E}_{x_i}\left[\mathbb{E}_{x'}[k(x_i,x')|x_i]^2 + \mathbb{E}_z[k(z,x_i)|x_i]^2 - 2\mathbb{E}_{x'}[k(x_i,x')|x_i]\mathbb{E}_z[k(z,x_i)|x_i]\right].$$

Note that the terms in $n_P^{-1}\sum_{i=1}^{n_P}\hat{f}_i(x_i)^2$ corresponding to A and B above vanish in the limit, since by the law of large numbers for U-statistics,

$$\frac{1}{n_P(n_P-1)}\sum_{i=1}^{n_P}\sum_{j\neq i} k(x_j,x_i)^2 \xrightarrow{\text{P}} \mathbb{E}_x[k(x,x_i)^2] < \infty \text{ and}$$
$$\frac{1}{n_P n_R}\sum_{i,j} k(z_j,x_i)^2 \xrightarrow{\text{P}} \mathbb{E}[k(x,z)^2] < \infty.$$

The other three terms are U-statistics (up to scaling negligible in the limit) of their counterparts in $\mathbb{E}_x[f(x)^2]$. Thus, by the same reasoning, the second limit holds. □

3.F MMD and KSD for Gaussian distributions

We provide explicit forms of MMD and KSD measured for Gaussian distributions. These results are used in constructing the PPCA experiment in Section 3.4 in the main body. In the process, we also obtain an understanding of the role of the reproducing kernel in the KSD.

3.F.1 MMD

This section provides an explicit formula for MMD between two normal distributions, defined by the exponentiated quadratic kernel

$$k(x,x') = \exp\left(\frac{-\|x-x'\|_2^2}{2\lambda^2}\right), \text{where } \lambda > 0.$$

The MMD expression in this setting has been shown [see e.g., Sriperumbudur et al., 2012, Example 3]. We use this formula to compute the difference of MMDs so that we can construct a problem as in Section 3.4.1.

Lemma 3.15. *Let $k(x,x') = \exp\left\{-\|x-x'\|_2^2/(2\lambda^2)\right\}$ with $\lambda > 0$. For three D-dimensional Gaussian distributions $P = \mathcal{N}(\mathbf{0},\Sigma_p)$, $Q = \mathcal{N}(\mathbf{0},\Sigma_q)$, and $R = \mathcal{N}(\mathbf{0},\Sigma_r)$ with full-rank covariance matrices, we have*

$$\mathrm{MMD}^2(P,R) - \mathrm{MMD}^2(Q,R) = \lambda^D\left\{\frac{1}{\sqrt{|2\Sigma_p+\lambda^2 I|}} - \frac{1}{\sqrt{2\Sigma_q+\lambda^2 I|}} - 2\left(\frac{1}{\sqrt{|\Sigma_p+\Sigma_r+\lambda^2 I)|}} - \frac{1}{\sqrt{|\Sigma_q+\Sigma_r+\lambda^2 I|}}\right)\right\},$$

where $|\cdot|$ *denotes the determinant.*

Note that we can numerically evaluate this expression given covariance matrices. For completeness, we provide a proof below.

Proof. Recall that the MMD between two distributions P, R is given by

$$\mathrm{MMD}^2(P,R) = \mathbb{E}_{(x,x')\sim P\otimes P}[k(x,x')] - 2\mathbb{E}_{x\sim P, x'\sim R}[k(x,x')] + \mathbb{E}_{(x,x')\sim R\otimes R}[k(x,x')].$$

Let $p(x) = \mathcal{N}(x; \mu_p, \Sigma_p)$, $r(x) = \mathcal{N}(x; \mu_r, \Sigma_r)$. Note that when properly scaled, the exponentiated quadratic kernel can be also seen as a Gaussian density function. Therefore, by convolution, we have

$$\mathbb{E}_{x'\sim P}[k(x,x')] = (2\pi\lambda^2)^{D/2}\mathcal{N}(x; \mu_p, \Sigma_p + \lambda^2 I).$$

Then, the first term is

$$\begin{aligned}
&\mathbb{E}_{x,x'\sim P\otimes P}[k(x,x')] = \int \mathbb{E}_{x'\sim P}[k(x,x')]\mathcal{N}(x;\mu_p,\Sigma_p)\mathrm{d}x \\
&= (2\pi\lambda^2)^{D/2}\int \mathcal{N}(x;\mu_p,\Sigma_p+\lambda^2)\mathcal{N}(x;\mu_p,\Sigma_p)\mathrm{d}x \\
&= \lambda^D \sqrt{\frac{|\tilde{\Sigma}_p|}{|(\Sigma_p+\lambda^2 I)||\Sigma_p|}}\exp\left(\frac{1}{2}\left\{\tilde{\mu}_p^\top\tilde{\Sigma}_p^{-1}\tilde{\mu}_p - \mu_p^\top(\Sigma_p+\lambda^2 I)^{-1}\mu_p - \mu_p^\top\Sigma_p^{-1}\mu_p\right\}\right),
\end{aligned}$$

where $|\cdot|$denotes the determinant, and

$$\tilde{\Sigma}_p = \big((\Sigma_p+\lambda^2 I)^{-1} + \Sigma_p^{-1}\big)^{-1}, \quad \tilde{\mu}_p = \tilde{\Sigma}_p\big((\Sigma_p+\lambda^2 I)^{-1}\mu_p + \Sigma_p^{-1}\mu_p\big).$$

The second term is

$$\begin{aligned}
&\mathbb{E}_{x\sim P,x'\sim R}[k(x,x')] \\
&= (2\pi\lambda^2)^{D/2}\int \mathcal{N}(x;\mu_p,\Sigma_p+\lambda^2 I)\mathcal{N}(x;\mu_r,\Sigma_r)\mathrm{d}x \\
&= \lambda^D \sqrt{\frac{|\Sigma_{p,r}|}{|(\Sigma_p+\lambda^2 I)||\Sigma_r|}}\exp\left(\frac{1}{2}\left\{\mu_{p,r}\Sigma_{p,r}^{-1}\mu_{p,r} - \mu_p^\top(\Sigma_p+\lambda^2 I)^{-1}\mu_p - \mu_r^\top\Sigma_r^{-1}\mu_r\right\}\right),
\end{aligned}$$

where

$$\Sigma_{p,r} = \big(\Sigma_r^{-1} + (\Sigma_p+\lambda^2 I)^{-1}\big)^{-1}, \quad \mu_{p,r} = \Sigma_{p,r}\left\{(\Sigma_p+\lambda^2 I)^{-1}\mu_p + \Sigma_r^{-1}\mu_r\right\}.$$

The third term is similarly obtained, but its form is not necessary for comparing models. We then impose the condition $\mu_p = \mu_r = \mathbf{0}$. In this case, we have

$$\mathrm{MMD}^2(P,R) = \lambda^D\left(\sqrt{\frac{|\tilde{\Sigma}_p|}{|(\Sigma_p+\lambda^2 I)||\Sigma_p|}} - 2\sqrt{\frac{|\Sigma_{p,r}|}{|(\Sigma_p+\lambda^2 I)||\Sigma_r|}}\right) + \mathbb{E}_{(x,x)'\sim R\otimes R}[k(x,x')]$$

$$= \lambda^D \left(\frac{1}{\sqrt{|2\Sigma_p + \lambda^2 I|}} - 2\frac{1}{\sqrt{|(\Sigma_p + \Sigma_r + \lambda^2 I|}} \right) + \mathbb{E}_{(x,x)' \sim R \otimes R}[k(x, x')]$$

Thus, substituting three Gaussian distributions $P = \mathcal{N}(\mathbf{0}, \Sigma_p)$, $Q = \mathcal{N}(\mathbf{0}, \Sigma_q)$, and $R = \mathcal{N}(\mathbf{0}, \Sigma_r)$, we obtain the desired equality. □

3.F.2 KSD

The KSD can be equivalently written in terms of the difference of score functions [Liu et al., 2016, Definition 3.2]:

$$\mathrm{KSD}^2\left(P \| R\right) = \mathbb{E}_{x_1, x_2 \sim R \otimes R}\left[(\mathbf{s}_p(x_1) - \mathbf{s}_r(x_1))^\top (\mathbf{s}_p(x_2) - \mathbf{s}_r(x_2)) k(x_1, x_2) \right].$$

For two Gaussian densities $p(x) = \mathcal{N}(x; \mathbf{0}, \Sigma_p)$, $r(x) = \mathcal{N}(x; \mathbf{0}, \Sigma_r)$, the difference between their score functions is

$$\mathbf{s}_p(x) - \mathbf{s}_r(x) = -(\Sigma_p^{-1} - \Sigma_r^{-1})x.$$

Therefore,

$$\begin{aligned} \mathrm{KSD}^2\left(P \| R\right) &= \mathbb{E}_{x_1, x_2 \sim R \otimes R}\left[(\mathbf{s}_p(x_1) - \mathbf{s}_r(x_1))^\top (\mathbf{s}_p(x_2) - \mathbf{s}_r(x_2)) k(x_1, x_2) \right] \\ &= \mathbb{E}_{x_1, x_2 \sim R \otimes R}\left[\left\langle \left(\Sigma_p^{-1} - \Sigma_r^{-1} \right)^2, k(x_1, x_2) x_1 x_2^\top \right\rangle \right] \\ &= \left\langle \left(\Sigma_p^{-1} - \Sigma_r^{-1} \right)^2, \underbrace{\mathbb{E}_{x_1, x_2 \sim R \otimes R}[k(x_1, x_2) x_1 {x_2}^\top]}_{M_{k,R}} \right\rangle, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product, and $M_{k,R}$ is a matrix depending on the kernel k and the data distribution R. Therefore, it can be informally understood that the KSD depends on the difference between the covariances Σ_p and Σ_r; if Σ_p is given by additive perturbation as $\Sigma_r + E$, the difference depends on the perturbation matrix E. Note that in the PPCA experiments, the perturbation matrix is an increasing function of δ, element-wise.

3.F.3 Kernel choice and KSD: Gaussian models and data

We illustrate how kernel choice affects the sensitivity of the KSD. We consider the following setting:

$$P \sim \mathcal{N}(0, \mathrm{diag}(1, , 1, \ldots, 1)), R \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \ldots, 1))$$

for some positive $\sigma_1 \neq 1$. The model P misestimates the variance of the first coordinate of the data. Let us consider the effect of a parameter choice for the IMQ kernel. We specifically compare the following kernels:

$$k_{\mathrm{IMQ}}(x,y)=\frac{1}{\sqrt{1+\|x-y\|_2^2}},\ k_{\mathrm{IMQ}}^{\mathrm{scale}}(x,y)=\frac{1}{\sqrt{1+\sum_{i>1}(x^i-y^i)^2+\sigma_1^{-2}(x^1-y^1)^2}},$$

where x^1 denotes the first coordinate of x. The latter kernel can be considered as the (preconditioned) IMQ kernel with dimension-wise scaling $\Lambda=\mathrm{diag}(\sigma_1^2,1,\ldots,1)$ where the scale is determined by the dimension-wise variance of the data. The KSDs corresponding to these kernel choices are given as follows:

$$\mathrm{KSD}\left(P\|R;k_{\mathrm{IMQ}}\right)^2=(\sigma_1-1)^2\underbrace{\mathbb{E}\left[\frac{X^1Y^1}{\sqrt{1+\sigma_1^2(X^1-Y^1)^2+\sum_{i>1}(X^i-Y^i)^2}}\right]}_{\mathbb{E}\left[k_{\mathrm{IMQ}}(X,Y)X^1Y^1\right]},$$

$$\mathrm{KSD}\left(P\|R;k_{\mathrm{IMQ}}^{\mathrm{scale}}(x,y)\right)^2=(\sigma_1-1)^2\underbrace{\mathbb{E}\left[\frac{X^1Y^1}{\sqrt{1+(X^1-Y^1)^2+\sum_{i>1}(X^i-Y^i)^2}}\right]}_{\mathbb{E}\left[k_{\mathrm{IMQ}}^{\mathrm{scale}}(X,Y)X^1Y^1\right]},$$

where the expectations are taken with respect to independent standard Gaussian random variables X,Y. For the KSD to be sensitive to this deviation in variance, the expectations on the RHS have to be large. In this regard, the key difference between these expressions is that the variance σ_1^2 appears in the coefficient of $(X^1-Y^1)^2$. When $\sigma_1\gg 1$, the non-scaled IMQ kernel k_{IMQ} pays more attention to the first coordinate than the scaled counterpart $k_{\mathrm{IMQ}}^{\mathrm{scale}}$, and we can therefore expect a higher expectation value for k_{IMQ}. On the other hand, when $\sigma_1\ll 1$, the relation flips as σ_1 reduces the contribution of the first coordinate. Note that this relation holds more starkly in high dimensions, as the rest of the coordinates have greater influence on the kernel output. These considerations show that the ability to choose an effective kernel depends on the problem (not surprisingly). In particular, it shows that dimension-wise scaling (or covariance preconditioning) could hurt the performance in some problems.

Finally, if we instead use the exponentiated quadratic (EQ) kernel

$$k_{\mathrm{EQ}}(x,y)=\exp\bigl(-\|x-y\|_2^2\bigr),$$

then we have

$$\mathrm{KSD}\left(P\|R\right)^2=(\sigma_1-1)^2\underbrace{\mathbb{E}_{X,Y}\left[\exp\Bigl(-\sum_{i>1}(X^i-Y^i)^2\Bigr)\exp\bigl(-\sigma_1^2(X^1-Y^1)^2\bigr)X^1Y^1\right]}_{\mathbb{E}[k_{\mathrm{EQ}}(X,Y)X^1Y^1]}.$$

When $\sigma_1\gg 1$, the EQ has higher selectivity in the first coordinate than the IMQ because of its exponential decay; the KSD with the EQ kernel could be more useful in revealing this perturbation. However, in practice, we do not know the discrepancy of our models a priori. At least in terms of local sensitivity such as the example above, the IMQ kernel could be considered more robust against poor specification of input scaling than the EQ, as the effect of input scaling is less significant.

3.G The score formula for Dirichlet process models

We provide an explicit formula for the score formula (3.3) for Dirichlet process mixture models, mentioned in 3.4. We first note that the density $p(x|\mathcal{D})$ is given by

$$\begin{aligned} p(x|\mathcal{D}) &= \mathbb{E}_F\left[\int \psi(x|z)\mathrm{d}F(z)\middle|\mathcal{D}\right] \\ &= \int\int \psi(x|z)\mathrm{d}\bar{F}_{\mathcal{D}}(z), \end{aligned}$$

where $\bar{F}_{\mathcal{D}}$ is the mean measure of the posterior distribution of F given $\mathcal{D}$. Note that $\bar{F}_{\mathcal{D}}$ is the mean of a mixture of Dirichlet processes with the mixing distribution given by the distribution of the latents of the training data $\{\tilde{z}_i\}_{i=1}^{n_{\mathrm{tr}}}$ conditioned on $\mathcal{D}$ [see Ghosal and van der Vaart, 2017, Remark 5.4]. We can interchange the inside integral and the expectation of F, which results in the second line. This expression immediately gives

$$\begin{aligned} \mathbf{s}_p(x) &= \frac{\int\int \nabla_x\psi(x|z)\mathrm{d}\bar{F}_{\mathcal{D}}(z)}{p(x|\mathcal{D})} \\ &= \int\int \mathbf{s}_\psi(x|z,\phi)\frac{\psi(x|z)}{p(x|\mathcal{D})}\mathrm{d}\bar{F}_{\mathcal{D}}(z). \end{aligned}$$

We discuss how to evaluate the expectation with MCMC. Our target distribution is

$$\frac{\psi(x|z)}{p(x|\mathcal{D})}\bar{F}_{\mathcal{D}}(\mathrm{d}z).$$

Note that this distribution is a mixture of two distributions written as

$$\begin{aligned} \frac{\psi(x|z)}{p(x|\mathcal{D})}\bar{F}_{\mathcal{D}}(\mathrm{d}z) &= \frac{1}{p(x|\mathcal{D})}\left\{\frac{C_a}{n+1}\frac{\psi(x|z)}{C_a}\mathrm{d}a(z) + \frac{nC_b}{n+1}\frac{\psi(x|z)}{C_b}\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_i \delta_{\tilde{Z}_i}(\mathrm{d}z)\middle|\mathcal{D}\right]}_{b(\mathrm{d}z)}\right\} \\ &= \pi_a\frac{\psi(x|z)}{C_a}\mathrm{d}a(z) + \pi_b\frac{\psi(x|z)}{C_b}\mathrm{d}b(z), \end{aligned}$$

where $C_\alpha = \int\psi(x|z)\mathrm{d}a(z)$, $C_b = \int\psi(x|z)\mathrm{d}b(z)$, $\pi_a = C_a/(C_a + nC_b)$, and $\pi_b = 1 - \pi_a$. For the Gaussian DPM model in 3.4.2, we can sample from the posterior $\psi(x|z,\phi)/C_a\mathrm{d}a(z)$, and it can be used for initializing the Markov chain. The distribution in the second term is not given in closed form as the mean measure b is unknown, but we can sample from b (and so from $\bar{F}_{\mathcal{D}}$) by Gibbs sampling [Ghosal and van der Vaart, 2017, Theorem 5.3]. Assuming that we can generate samples from $\bar{F}_{\mathcal{D}}$, we can use the random walk Metropolis algorithm where the acceptance probability of the transition from z to z' is given by

$$\min\left(1, \frac{\psi(x|z')}{\psi(x|z)}\right)$$

with the proposal distribution $\bar{F}_{\mathcal{D}}$. However, sampling from $\bar{F}_D$ cannot be performed exactly, and therefore we use Gibbs sampling after sufficient burn-in. Consequently, the use of Gibbs and Metropolis samplers allows us to sample from $\psi(x|z)/p(x|\mathcal{D})\bar{F}_{\mathcal{D}}$ approximately.

3.H Invariance properties of kernel Stein discrepancy

3.H.1 Model invariance

In some applications, models are designed to be invariant to certain transformations. When comparing a class of models invariant under a transformation, model comparison should be made so that the ranks of the models remain unaffected under the transformation of the data. We show that essentially for rotational transformations, we can make the KSD invariant by choosing a rotationally invariant kernel. In the following, for a map $T : \mathcal{X} \to \mathcal{X}$ and a distribution P, we denote T-push-forward of P by $T_{\#}P$; i.e., $T_{\#}P$ is defined as the distribution of a random variable Tx with $x \sim P$.

Lemma 3.16. *Assume that for an orthogonal matrix O, the following conditions hold: (a) P has a density p such that $p(Ox) = p(x)$ for any $x \in \mathbb{R}^D$, and (b) kernel k satisfies $k(Ox, Oy) = k(x, y)$ for any $x, y \in \mathbb{R}^D$. Then, we have* $\mathrm{KSD}\left(P \| O_{\#}R\right) = \mathrm{KSD}\left(P \| R\right)$.

Proof. When we push forward the data distribution by O, the KSD becomes

$$\mathrm{KSD}\left(P \| O_{\#}R\right)^2 = \mathbb{E}_{x,x' \sim R\otimes R}\big[h_p(Ox, Ox')\big].$$

The assumption $p(Ox) = p(x)$ implies that

$$\begin{aligned}\mathbf{s}_p(Ox) &= \frac{1}{p(x)}\Big(\partial_h p(x + hO^{-1}e_d)|_{h=0}\Big)_{d=1}^{D}\\ &= (O^{-1})^\top \frac{\nabla p(x)}{p(x)} = O\mathbf{s}_p(x),\end{aligned}$$

where $\{e_1, \dots, e_D\}$ denotes the standard basis of $\mathbb{R}^D$. For the kernel derivatives k_1 and k_{12}, we obtain

$$k_1(Ox, Ox') = Ok_1(x, x') \text{ and } k_{12}(Ox, Ox') = k_{12}(x, x').$$

Thus,

$$\begin{aligned}h_p(Ox, Ox') &= \mathbf{s}_p(Ox)^\top \mathbf{s}_p(Ox')k(Ox, Ox') + \mathbf{s}_p(Ox)^\top k_1(Ox', Ox)\\ &\quad + k_1(Ox, Ox')^\top \mathbf{s}_p(Ox') + k_{12}(Ox, Ox')\\ &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x')k(x, x') + \mathbf{s}_p(x)^\top k_1(x', x)\end{aligned}$$

$$+ k_1(x,x')^\top \mathbf{s}_p(x') + k_{12}(x,x') = h_p(x,x'),$$

and therefore $\mathrm{KSD}\left(P \| O_\# R\right) = \mathrm{KSD}\left(P \| R\right)$. □

An analogous result holds for the KSD for discrete observations.

Lemma 3.17. *Let $\sigma : \{1,\ldots,D\} \to \{1,\ldots,D\}$ be a permutation represented by a permutation matrix O_σ. Assume that P is invariant to O_σ; i.e., $(O_\sigma)_\# P = P$. Assume that kernel k satisfies $k(O_\sigma x, O_\sigma y) = k(x,y)$ for any $x, y \in \{0,\ldots,L-1\}^D$. Then,* $\mathrm{KSD}\left(P \| (O_\sigma)_\# P\right) = \mathrm{KSD}\left(P \| R\right)$.

Proof. The proof proceeds as in the previous lemma. Note that taking the cyclic forward difference with respect to the i-th coordinate gives

$$\begin{aligned}\Delta_i p(O_\sigma x) &= p(x^{\sigma(1)},\ldots,\tilde{x}^i,\ldots x^{\sigma(D)}) - p(x^{\sigma(1)},\ldots,x^{\sigma(i)},\ldots x^{\sigma(D)})\\ &= p(x^1,\ldots,\tilde{x}^{\sigma^{-1}(i)},\ldots x^D) - p(x)\\ &= \Delta_{\sigma^{-1}(i)} p(x),\end{aligned}$$

where $\tilde{x} = x + 1 \mod L$. Thereby,

$$\mathbf{s}_p(O_\sigma x) = O_\sigma^{-1}\mathbf{s}_p(x) = O_\sigma^\top \mathbf{s}_p(x).$$

Similarly, for the kernel *derivatives* k_1 and k_{12}, we have

$$\begin{aligned}k_1(O_\sigma x, O_\sigma x') &= O_\sigma^\top k_1(x,x')\\ k_{12}(O_\sigma x, O_\sigma x') &= k_{12}(x,x').\end{aligned}$$

Thus,

$$\begin{aligned}h_p(O_\sigma x, O_\sigma x') &= \mathbf{s}_p(O_\sigma x)^\top \mathbf{s}_p(O_\sigma x') k(O_\sigma x, O_\sigma x') + \mathbf{s}_p(O_\sigma x)^\top k_1(O_\sigma x', O_\sigma x)\\ &\quad + k_1(O_\sigma x, O_\sigma x')^\top \mathbf{s}_p(O_\sigma x') + k_{12}(O_\sigma x, O_\sigma x')\\ &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x,x') + \mathbf{s}_p(x)^\top k_1(x',x)\\ &\quad + k_1(x,x')^\top \mathbf{s}_p(x') + k_{12}(x,x') = h_p(x,x'),\end{aligned}$$

and therefore $\mathrm{KSD}\left(P \| (O_\sigma)_\# R\right) = \mathrm{KSD}\left(P \| R\right)$. □

3.H.2 Coordinate-choice independence

In general, the KSD is not invariant to a change of coordinates, and the KSD may be affected if we transform both the model and the data distribution. Precisely, for some one-to-one map $T : \mathcal{X} \to \mathcal{X}$, we might have $\mathrm{KSD}\left(T_\# P \| T_\# R\right) \neq \mathrm{KSD}\left(P \| R\right)$. The following result is essentially the same as the previous lemmas except that here we do not have the invariance assumptions for the model; it shows that the KSD can be made invariant at least under rotation and translation.

Lemma 3.18. *Let T be an affine transform such that $Tx = Ox + b$ where O is an orthogonal matrix and b is a vector. Let P, R be probability distributions. Let $k = k_R$ be a* data-dependent *kernel k_R such that $k_{T_\# R}(x, x') = k_R(T^{-1}x, T^{-1}x')$ for any $x, x' \in \mathbb{R}^D$. Then the KSD between P and R is invariant under T; that is,* $\mathrm{KSD}\left(T_\# P \| T_\# R\right) = \mathrm{KSD}\left(P \| R\right)$.

Proof. Let us denote the density of $T_\# P$ by $p_T(x) = p(T^{-1}x)$. Then, its score function satisfies

$$\mathbf{s}_{p_T}(x) = O\mathbf{s}_p(T^{-1}x).$$

Similarly, for the kernel derivatives k_1 and k_{12}, we have

$$\begin{aligned} k_{T_\# R,1}(x, y) &= \nabla_{\tilde{x}} k_R(T^{-1}\tilde{x}, T^{-1}y)|_{\tilde{x}=x} \\ &= Ok_{R,1}(T^{-1}x, T^{-1}y), \text{ and} \\ k_{T_\# R,12}(x, y) &= \nabla_{\tilde{x}}^\top \nabla_{\tilde{y}} k_R(T^{-1}\tilde{x}, T^{-1}y)|_{\tilde{x}=x,\tilde{y}=y} \\ &= (\nabla_{\tilde{x}}^\top O^{-1} O^\top \nabla_{\tilde{y}}) k_R(\tilde{x}, \tilde{y})|_{\tilde{x}=x,\tilde{y}=y} = k_{R,12}(x, y). \end{aligned}$$

These relations imply that the Stein kernel satisfies

$$\begin{aligned} h_{p_T}(Tx, Tx') &= \mathbf{s}_{p_T}(Tx)^\top \mathbf{s}_{p_T}(Tx') k_{T_\# R}(Tx, Tx') + \mathbf{s}_{p_T}(x)^\top k_{T_\# R,1}(Tx', Tx) \\ &\quad + k_{T_\# R,1}(Tx, Tx')^\top \mathbf{s}_{p_T}(x') + k_{T_\# R,12}(Tx, Tx') \\ &= \mathbf{s}_p(x) O^\top O \mathbf{s}_p(x') k_R(x, x') + \mathbf{s}_p(x)^\top O^\top O k_{R,1}(x', x) \\ &\quad + k_{R,1}(x, x')^\top O^{-1} O \mathbf{s}_p(x') + (\nabla_{\tilde{x}}^\top O^{-1} O \nabla_{\tilde{y}}) k_{R,12}(T^{-1}\tilde{x}, T^{-1}\tilde{y})|_{\tilde{x}=Tx,\tilde{y}=Tx'} \\ &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k_R(x, x') + \mathbf{s}_p(x)^\top k_{R,1}(x', x) \\ &\quad + k_{R,1}(x, x')^\top \mathbf{s}_p(x') + k_{R,12}(x, x') = h_p(x, x'). \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{KSD}\left(T_\# P \| T_\# R\right)^2 &= \mathbb{E}_{x,x' \sim R \otimes R}\Big[h_{p_T}(Tx, Tx')\Big] \\ &= \mathbb{E}_{x,x' \sim R \otimes R}\Big[h_p(x, x')\Big] = \mathrm{KSD}\left(P \| R\right)^2. \end{aligned}$$

□

An example of the data-dependent kernel is a covariance-preconditioned kernel

$$k_R^{\text{precond}}(x, y) = \phi\big((x - y)^\top \hat{\Sigma}_R^{-1}(x - y)\big)$$

where $\hat{\Sigma}_R$ is the sample covariance matrix of R and ϕ is some positive-definite function. Another example is the median-scaled kernel

$$k_R^{\text{med}}(x, y) = \phi\big(\|x - y\|_2^2 / \sigma_{R,\text{med}}^2\big),$$

where $\sigma_{R,\text{med}}$ is the sample median: $\text{median}\{\|x_i - x_j\|_2 | 1 < i < j < n\}$. In fact, radial basis kernels with data-independent scaling also satisfy the required condition since $k_{T_\# R}(x, y) =$

$k_R(x, y) = k_R(T^{-1}x, T^{-1}y)$.

3.I Arvesen's formula for the jackknife variance estimator

The formula in [Arvesen, 1969, Eq. 25] has minor errors. For completeness, we provide a proof for the decomposition (3.13).

Lemma 3.19. *For an i.i.d. sample $\mathcal{D}_n = \{y_i\}_{i=1}^n$ from some distribution, define a U-statistic with symmetric kernel $f : \mathcal{Y}^s \to \mathbb{R}$,*

$$U_n = \binom{n}{s}^{-1} \sum_{C_{n,s}} f(y_{i_1}, \dots, y_{i_s}),$$

where $C_{n.s}$ denotes the set of s combinations of integers chosen from $\{1, \dots, n\}$ with $n \geq s \geq 1$. Then, we have

$$\begin{aligned} v_n^J &:= (n-1)\sum_{i=1}^n (U_{n,-i} - U_n)^2 \\ &= \sum_{c=0}^s a_{n,c}\hat{U}_c, \end{aligned}$$

with $U_{n,-i}$ the U-statistic computed with $\mathcal{D}_n \setminus \{y_i\}$,

$$a_{n,c} = \frac{n-1}{n}\binom{n-1}{s}^{-2}\{n\mathbb{I}(c>0) - s^2\}\binom{n}{c}\binom{n-c}{s-c}\binom{n-s}{s-c},$$

and $\hat{U}_c$ is a U-statistic given in Eq. (3.27) in the proof.

Proof. For a combination in $C_{n,s}$, we fix a order of integers (say, sorted in ascending order) and evaluate f. We may assume that the statistic is centered so that $\mathbb{E}[U_n] = 0$. The jackknife estimator is expressed as

$$\begin{aligned} v_n^J &= (n-1)\sum_{i=1}^n (U_{n,-i} - U_n)^2 \\ &= (n-1)\Big\{\sum_{i=1}^n (U_{n,-i})^2 - n(U_n)^2\Big\} \\ &= (n-1)\binom{n-1}{s}^{-2}\Bigg[\sum_{i=1}^n \sum_{\alpha\in C_{n,s-1}^i} f(y_{\alpha_1}, \dots, f_{\alpha_s}) \sum_{\beta\in C_{n,s-1}^i} f(y_{\beta_1}, \dots, f_{\beta_s}) \\ &\qquad - \frac{(n-s)^2}{n}\sum_{\alpha\in C_{n,s}} f(y_{\alpha_1}, \dots, f_{\alpha_s}) \sum_{\beta\in C_{n,s}} f(y_{\beta_1}, \dots, f_{\beta_s})\Bigg], \end{aligned} \tag{3.22}$$

where $C_{n,s-1}^i$ denotes the set of all s combinations of integers chosen from $\{1, \dots, i-1, i+1, \dots, n\}$, and $C_{n,s}$ denotes the set of all s combinations of integers chosen from $\{1, \dots, n\}$.

The first sum can be expressed as

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\alpha \in C^i_{n,s-1}} f(y_{\alpha_1}, \dots, f_{\alpha_s}) \sum_{\beta \in C^i_{n,s-1}} f(y_{\beta_1}, \dots, f_{\beta_s}) \\
&= \sum_{i=1}^{n} \Bigg\{ \Big(\sum_{\alpha \in C_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}) - \sum_{\alpha \in C^i_{n,s-1}} f(y_i, y_{\alpha_1} \dots, f_{\alpha_{s-1}}) \Big) \\
&\qquad \cdot \Big(\sum_{\beta \in C_{n,s}} f(y_{\beta_1} \dots, f_{\beta_s}) - \sum_{\beta \in C^i_{n,s-1}} f(y_i, y_{\beta_1} \dots, f_{\beta_{s-1}}) \Big) \Bigg\}.
\end{aligned}$$

Note that we have

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{\alpha \in C^i_{n,s-1}} f(y_i, y_{\alpha_1} \dots, f_{\alpha_{s-1}}) &= \frac{1}{(s-1)!} \sum_{i=1}^{n} \sum_{\alpha \in P^i_{n,s-1}} f(y_i, y_{\alpha_1} \dots, f_{\alpha_{s-1}}) \\
&= \frac{1}{(s-1)!} \sum_{\alpha \in P_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}) \\
&= \frac{s!}{(s-1)!} \sum_{\alpha \in C_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}) \\
&= s \sum_{\alpha \in C_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}),
\end{aligned}$$

where $P^i_{n,s-1}$ is the set of all ordered $(s-1)$-tuples of integers chosen from $\{1, \dots i-1, i+1, \dots, n\}$, and $P_{n,s}$ denotes the set of all ordered s-tuple of integers chosen from $\{1, \dots, n\}$. The first and third lines are due to the symmetry of f. The second lines holds as the indices on the RHS on the first line runs all over $P_{n,s}$. Moreover, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\alpha \in C^i_{n,s-1}} f(y_i, y_{\alpha_1} \dots, f_{\alpha_{s-1}}) \sum_{\beta \in C^i_{n,s-1}} f(y_i, y_{\beta_1} \dots, f_{\beta_{s-1}}) \\
&= \sum_{\substack{\alpha, \beta \in C_{n,s} \\ |\alpha \cap \beta| \geq 1}} f(y_{\alpha_1}, \dots, f_{\alpha_s}) f(y_{\beta_1} \dots, f_{\beta_s}),
\end{aligned}$$

where $|\alpha \cap \beta|$ expresses the number of common elements between two sets of integers α, β. Thus, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\alpha \in C^i_{n,s-1}} f(y_{\alpha^i_1}, \dots, f_{\alpha^i_s}) \sum_{\beta \in C^i_{n,s-1}} f(y_{\beta^i_1}, \dots, f_{\beta^i_s}) \\
&= n \sum_{\alpha \in C_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}) \sum_{\beta \in C_{n,s}} f(y_{\beta_1} \dots, f_{\beta_s}) - 2s \sum_{\alpha \in C_{n,s}} f(y_{\alpha_1} \dots, f_{\alpha_s}) \\
&\quad + \sum_{\substack{\alpha, \beta \in C_{n,s} \\ |\alpha \cap \beta| \geq 1}} f(y_{\alpha_1}, \dots, f_{\alpha_s}) f(y_{\beta_1} \dots, f_{\beta_s}).
\end{aligned}$$

Now, the expression (3.22) can be summarized as

$$
\begin{aligned}
&\left(\frac{n-1}{n}\right)^{-1}\binom{n-1}{s}^{2} \times (3.22)\\
&=\Bigg[\sum_{c=0}^{s}\{n^2-2sn+n\mathbb{I}(c>0)\}\sum_{\substack{\alpha,\beta\in C_{n,s}\\ |\alpha\cap\beta|=c}} f(y_{\alpha_1},\dots,f_{\alpha_s})f(y_{\beta_1},\dots,f_{\beta_s}) \qquad (3.23)\\
&\qquad -(n-s)^2\sum_{c=0}^{s}\sum_{\substack{\alpha,\beta\in C_{n,s}\\ |\alpha\cap\beta|=c}} f(y_{\alpha_1},\dots,f_{\alpha_s})f(y_{\beta_1},\dots,f_{\beta_s})\Bigg]\\
&=\sum_{c=0}^{s}\{n\mathbb{I}(c>0)-s^2\}\sum_{\substack{\alpha,\beta\in C_{n,s}\\ |\alpha\cap\beta|=c}} f(y_{\alpha_1},\dots,f_{\alpha_s})f(y_{\beta_1},\dots,f_{\beta_s}). \qquad (3.24)
\end{aligned}
$$

Finally, we show that each term of the RHS (3.24) is written by a U-statistic $\hat{U}_c$ with kernel

$$
\begin{aligned}
&f_{\text{sym}}(y_{\alpha_1},\dots,y_{\alpha_c},y_{\beta_1},\dots,y_{\beta_{s-c}},y_{\gamma_1},\dots,y_{\gamma_{s-c}})\\
&:=\sum_{\sigma}\frac{f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\beta_1)},\dots,y_{\sigma(\beta_{s-c})})f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\gamma_1)},\dots,y_{\sigma(\gamma_{s-c})})}{(2s-c)!},
\end{aligned}
$$

where the sum is over $\Sigma(\alpha,\beta,\gamma)$, the set of all permutations of given integers (α,β,γ) with $\alpha =(\alpha_1,\dots,\alpha_c)$, $\beta=(\beta_1,\dots,\beta_{s-c})$, and $\gamma=(\gamma_1,\dots,\gamma_{s-c})$. The reasoning is as follows:

$$
\begin{aligned}
&\hat{U}_c\\
&:=\binom{n}{2s-c}^{-1}\sum_{C_{n,2s-c}} f_{\text{sym}}(y_{\alpha_1},\dots,y_{\alpha_c},y_{\beta_1},\dots,y_{\beta_{s-c}},y_{\gamma_1},\dots,y_{\gamma_{s-c}})\\
&=\frac{(n-2s+c)!}{n!}\sum_{P_{n,2s-c}} f_{\text{sym}}(y_{\alpha_1},\dots,y_{\alpha_c},y_{\beta_1},\dots,y_{\beta_{s-c}},y_{\gamma_1},\dots,y_{\gamma_{s-c}})\\
&=\frac{(n-2s+c)!}{n!}\frac{1}{(2s-c)!}\\
&\quad\cdot\sum_{P_{n,2s-c}}\sum_{\sigma\in\Sigma(\alpha,\beta,\gamma)}\Big(f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\beta_1)},\dots,y_{\sigma(\beta_{s-c})})\\
&\qquad\qquad f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\gamma_1)},\dots,y_{\sigma(\gamma_{s-c})})\Big) \qquad (3.25)\\
&=\frac{(n-2s+c)!}{n!}\frac{1}{(2s-c)!}\\
&\quad\cdot\sum_{\sigma}\sum_{\sigma(\alpha)\in P_{n,c}}\sum_{\substack{\sigma(\beta)\in P_{n,s-c}\\ \sigma(\alpha)\cap\sigma(\beta)=\emptyset}}\sum_{\substack{\sigma(\gamma)\in P_{n,s-c}\\ \sigma(\gamma)\cap\{\sigma(\alpha)\cup\sigma(\beta)\}=\emptyset}}\Big(f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\beta_1)},\dots,y_{\sigma(\beta_{s-c})})\\
&\qquad\qquad\cdot f(y_{\sigma(\alpha_1)},\dots,y_{\sigma(\alpha_c)},y_{\sigma(\gamma_1)},\dots,y_{\sigma(\gamma_{s-c})})\Big)\\
&\qquad (3.26)
\end{aligned}
$$

The second line follows from the permutation invariance of f_{sym}. The third line is obtained by inserting the definition. The fourth line is a result of exchanging the sums. Continuing to manipulate the expression, we obtain

$$\begin{aligned}\hat{U}_c &= \frac{(n-2s+c)!}{n!}(s-c)!(s-c)!c! \\ &\quad \cdot \sum_{\alpha \in C_{n,c}} \sum_{\substack{\beta \in C_{n,s-c} \\ \alpha \cap \beta = \emptyset}} \sum_{\substack{\gamma \in C_{n,s-c} \\ \gamma \cap \{\alpha \cup \beta\} = \emptyset}} f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\beta_1}, \ldots, y_{\beta_{s-c}}) f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\gamma_1}, \ldots, y_{\gamma_{s-c}}). \\ &= \frac{(n-2s+c)!(s-c)!}{(n-s)!} \frac{(n-s)!(s-c)!}{(n-c)!} \frac{c!(n-c)!}{n!} \\ &\quad \cdot \sum_{\alpha \in C_{n,c}} \sum_{\substack{\beta \in C_{n,s-c} \\ \alpha \cap \beta = \emptyset}} \sum_{\substack{\gamma \in C_{n,s-c} \\ \gamma \cap \{\alpha \cup \beta\} = \emptyset}} f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\beta_1}, \ldots, y_{\beta_{s-c}}) f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\gamma_1}, \ldots, y_{\gamma_{s-c}}) \\ &= \binom{n-s}{s-c}^{-1} \binom{n-c}{s-c}^{-1} \binom{n}{c}^{-1} \\ &\quad \cdot \sum_{\alpha \in C_{n,c}} \sum_{\substack{\beta \in C_{n,s-c} \\ \alpha \cap \beta = \emptyset}} \sum_{\substack{\gamma \in C_{n,s-c} \\ \gamma \cap \{\alpha \cup \beta\} = \emptyset}} f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\beta_1}, \ldots, y_{\beta_{s-c}}) f(y_{\alpha_1}, \ldots, y_{\alpha_c}, y_{\gamma_1}, \ldots, y_{\gamma_{-c}}).\end{aligned} \tag{3.27}$$

The first line holds because the inner sum (3.26) in Σ_σ has the same value for each σ and because of the permutation invariance of f. Note that the sum in the final expression is

$$\sum_{\substack{\alpha, \beta \in C_{n,s} \\ |\alpha \cap \beta| = c}} f(y_{\alpha_1}, \ldots, f_{\alpha_s}) f(y_{\beta_1}, \ldots, f_{\beta_s}).$$

Therefore,

$$\begin{aligned} v_n^J &= (n-1) \sum_{i=1}^{n} (U_{n,-i} - U_n)^2 \\ &= \sum_{c=0}^{s} a_{n,c} \hat{U}_c, \end{aligned}$$

with

$$a_{n,c} = \frac{n-1}{n} \binom{n-1}{s}^{-2} \{n\mathbb{I}(c > 0) - s^2\} \binom{n}{c} \binom{n-c}{s-c} \binom{n-s}{s-c}.$$

□

3.J Additional experiments

3.J.1 PPCA: type-I errors and test power

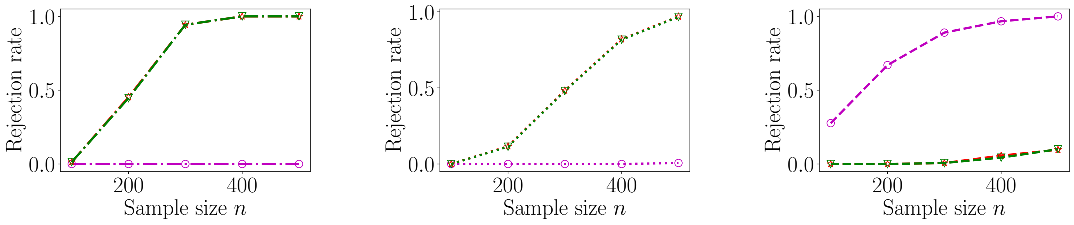
We provide results supplementary to the results in Section 3.4.3.1.

Problem 1 (null) Table 3.J.1 summarizes the result for the experiment with $(\delta_P, \delta_Q) = (1, 1 + 10^{-5})$ in (3.J.1), where all the tests use the IMQ kernel with covariance preconditioning. The result for $\alpha = 0.01$ is omitted as none of the examined tests rejected the hypotheses.

Table 3.J.1: Type-I errors the MMD test of Bounliphone et al. [2016], the proposed LKSD test, and the KSD test with the covariance preconditioner in PPCA Problem 1. Rejection rates are computed on 300 trials for significance level $\alpha = 0.05$.

Sample size n	Rejection rates		
	MMD	KSD	LKSD
100	0.000	0.017	0.010
200	0.013	0.000	0.003
300	0.020	0.000	0.007
400	0.017	0.000	0.007
500	0.013	0.000	0.007

Problem 2 (alternative) We present the result of the same experiment with $\alpha = 0.01$ (Figure 3.J.1) to show power decay due to the conservatism.



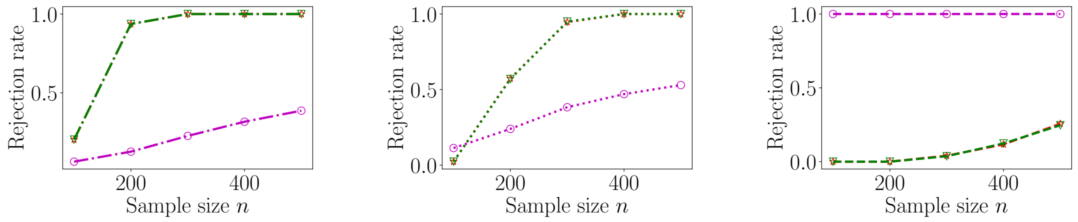
(a) (a): PPCA $\delta_P = 2$, $\delta_Q = 1$. EQ kernel with median scaling.

(b) PPCA $\delta_P = 2$, $\delta_Q = 1$. IMQ kernel with median scaling.

(c) PPCA $\delta_P = 2$, $\delta_Q = 1$. IMQ kernel with covariance preconditioning.

Figure 3.J.1: Power curves of the MMD test of Bounliphone et al. [2016], the proposed LKSD test, and the KSD test with the exact score function in PPCA Problem 2. The perturbation parameters are set as $(\delta_P, \delta_Q = 2, 1)$. Each result is computed on 300 trials. The significance level $\alpha = 0.01$. Markers: ▽ (the LKSD test); ☆ (the KSD test); ◯ (the relative MMD test).

Next, we provide the result for the same power experiment with a different choice of perturbation parameters $(\delta_P, \delta_Q) = (3, 1)$. Figure 3.J.2 shows the power curves of the tests. The MMD test with the covariance pre-conditioner achieves power 1 at $n = 100$.



(a) (a): PPCA $\delta_P = 3$, $\delta_Q = 1$. EQ kernel with median scaling.

(b) PPCA $\delta_P = 3$, $\delta_Q = 1$. IMQ kernel with median scaling.

(c) PPCA $\delta_P = 3$, $\delta_Q = 1$. IMQ kernel with covariance preconditioning.

Figure 3.J.2: Power curves of the MMD test of Bounliphone et al. [2016], the proposed LKSD test, and the KSD test with the exact score function in PPCA Problem 2. The perturbation parameters are set as $(\delta_P, \delta_Q = 3, 1)$. Each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: $\bigtriangledown$ (the LKSD test); ☆ (the KSD test); ◯ (the relative MMD test).

3.J.2 LDA

3.J.2.1 Type-I error and test power

LDA models with more sparse topics We run the same experiment as in Section 3.4.3 except that the topics b are made sparse by sampling from the Dirichlet distribution with all the concentration parameters 0.1. The results are summarized in Tables 3.J.2 and 3.J.3. The LKSD test underperforms the MMD test in this case. As the topics are more sparse, generated documents tend to have words from a particular topic; this trend escalates as we increase the concentration parameter of the topic proportion prior. Models thus observe compositions of words that they would not generate, resulting in a high-variance test statistic for the same reason as in Section 3.4.3.2 (In fact, the topics have probabilities as low as 10^{-40}, which comes close to violating the assumption on the density).

Table 3.J.2: Type-I error for LDA Problem 1 $(\delta_P, \delta_Q) = (0.5, 0.6)$

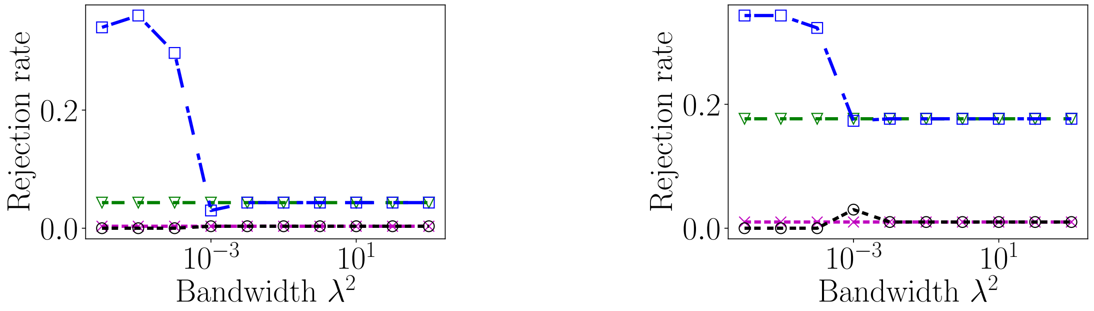
Sample size n	Rejection rates			
	EQ BoW		IMQ BoW	
	MMD	LKSD	MMD	LKSD
100	0.007	0.013	0.007	0.013
200	0.007	0.007	0.003	0.007
300	0.007	0.017	0.000	0.017
400	0.013	0.017	0.003	0.020
500	0.020	0.010	0.003	0.010

Table 3.J.3: Power estimates for LDA Problem 2 $(\delta_P, \delta_Q) = (1.1, 0.6)$

Sample size n	Rejection rates							
	$\alpha = 0.01$				$\alpha = 0.05$			
	EQ BoW		IMQ BoW		EQ BoW		IMQ BoW	
	MMD	LKSD	MMD	LKSD	MMD	LKSD	MMD	LKSD
100	0.000	0.003	0.023	0.003	0.013	0.040	0.083	0.037
200	0.000	0.000	0.030	0.000	0.017	0.040	0.170	0.043
300	0.003	0.013	0.047	0.013	0.010	0.053	0.290	0.053
400	0.003	0.007	0.093	0.007	0.013	0.080	0.373	0.077
500	0.000	0.000	0.146	0.000	0.000	0.050	0.477	0.050

3.J.2.2 Kernel parameter

As in the PPCA experiment, we investigate the performance dependence on the kernel choice. Using LDA problem 2 , we examine how the test power is affected by the scaling parameter. We use the EQ and IMQ BoW kernels as above, and choose their scaling parameter λ^2 from $\{10^{-6}, 10^{-5}, \ldots, 10^{3}\}$. For each $n \in \{100, 300\}$ we run 300 trials and estimate the test power of the LKSD and MMD tests. Figure 3.2 plots the power curves of the tests. We can see that the MMD test fails for any choice of the kernel. For the LKSD test, the IMQ kernel has a flat curve, indicating its independence from the bandwidth (at least in this candidate range), whereas the EQ kernel benefits from a small bandwidth value.



(a) $n = 100$.

(b) $n = 300$.

Figure 3.J.3: Power curves of the proposed LKSD test and the MMD test in LDA Problem 2. The perturbation parameters are set as $(\delta_P, \delta_Q = 2, 1)$. Each result is computed on 300 trials. The significance level $\alpha = 0.05$. Markers: $\bigtriangledown$ (LKSD test with IMQ kernel); $\square$ (LKSD test with EQ kernel); $\bigcirc$ (MMD test with IMQ kernel); $\times$ (MMD test with EQ kernel).

3.J.3 Experiment: close models and type-I errors

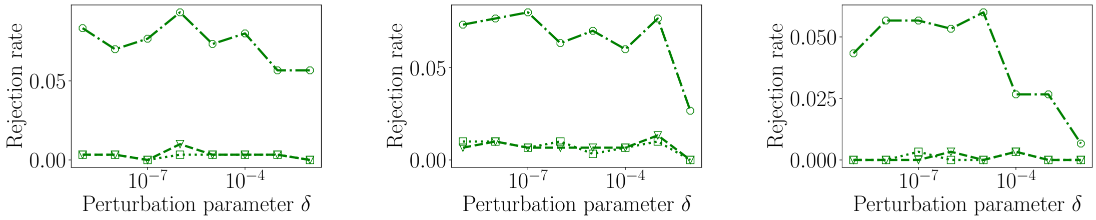
We investigate the behavior of the LKSD test when two models are close to each other. In this case, the difference of the U-statistic kernels defined by the models is small, which could therefore lead to the degeneracy of the U-statistic; i.e., the normal approximation of the test statistic is not appropriate. In the following, we investigate the three variants of the LKSD test defined by different choices of the variance estimator. We compare the jackknife estimator (3.11) with the following two estimators: (i) a U-statistic variance estimator where ζ_1 and ζ_2 in

(3.10) are estimated by U-statistics, and (ii) a V-statistic variance estimator where ζ_1 is estimated by a V-statistic. The U-statistic estimation was considered by Bounliphone et al. [2016] and Jitkrittum et al. [2018]. The issue with the U-statistic estimator is that it underestimates the actual variance. In fact, we observed that the variance estimator sometimes returns **negative** values. This can occur since the statistic is given as a difference between unbiased estimates of quantities close to zero (this issue applies to the V-statistic estimator). We made the (arbitrary) choice to accept the null hypothesis when the variance estimate was negative to avoid false rejections. For this reason, we recommend against using the U-statistic estimator.

3.J.3.1 PPCA

Our first experiment concerns PPCA models. Specifically, we choose a PPCA model for the data distribution as in Section 3.4.1 with $D = 50$. The difference is that we fix the perturbation parameter δ_P for P at 1, and vary the parameter δ_Q by choosing it from $\{10^{-i} : i \in \{2, 3, \dots, 9\}\}$ (this choice yields null H_0 scenarios). We set the significance level $\alpha = 0.05$. For each $n \in \{100, 200, 300\}$, we run the tests for 300 trials and examine the behavior of the tests under the null.

Figure 3.J.4 shows the tests' rejection rates. We first note that as the perturbation parameter decays (the models get closer to each other), the test with the U-statistic estimator rejects more and has higher type-I errors than the nominal level $\alpha = 0.05$. These plots demonstrate that the jackknife and V-statistic versions of the test are more robust in this setting.



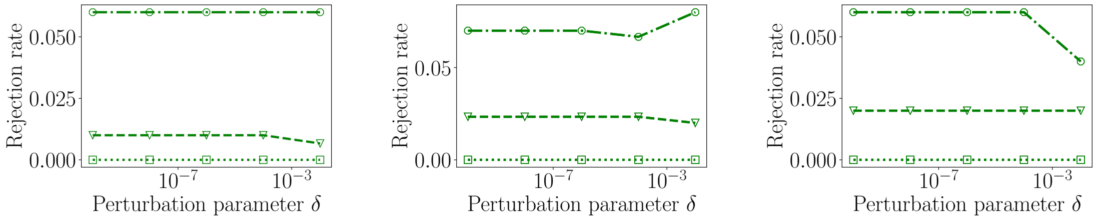
(a) $n = 100$ (b) $n = 200$ (c) $n = 300$

Figure 3.J.4: The behaviors of the two LKSD tests under the null. The nominal level α is set to 0.05. The test with the U-statistic variance estimator has higher type-I errors as the models get closer to each other. Markers: ▽ (LKSD test with the jackknife variance estimator); ○ (LKSD test with the U-statistic variance estimator); □ (LKSD test with the V-statistic variance estimator).

3.J.3.2 LDA

We conduct a similar experiment with LDA models. The problem setup is the same as in Section 3.4.3, except that the vocabulary size $L = 100$. We perturb the sparsity parameter of the Dirichlet prior of an LDA model. We set $\delta_P = 1$ and $\delta_Q = 1 + \delta$, where δ is chosen from $\{10^{-2i} : i \in \{1, \dots, 5\}\}$. For each $n \in \{100, 200, 300\}$, we run the tests for 300 trials with significance level $\alpha = 0.05$.

Figure 3.J.5 shows the rejection rate of each test. The test with the V-statistic estimator is more conservative than the other tests. The U-statistic variance appears to underestimate the actual variance.

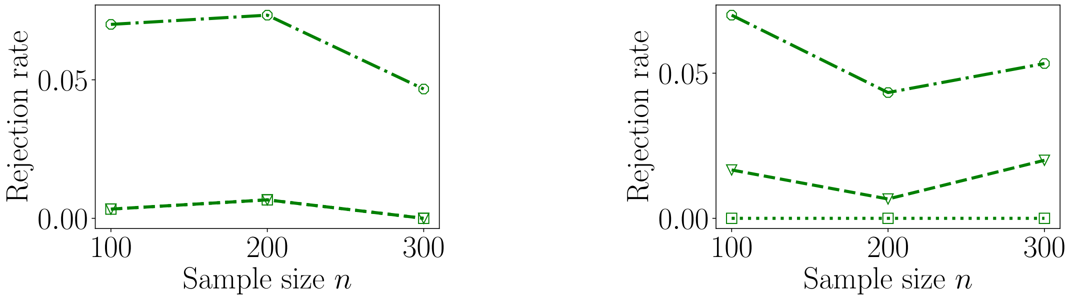


(a) $n = 100$ (b) $n = 200$ (c) $n = 300$

Figure 3.J.5: The behaviors of the two LKSD tests under the null. The nominal level α is set to 0.05. The test with the U-statistic variance estimator has higher type-I errors as the models get closer to each other. Markers: ▽ (LKSD test with the jackknife variance estimator); ○ (LKSD test with the U-statistic variance estimator); □ (LKSD test with the V-statistic variance estimator).

3.J.4 Experiment: identical models

We look into the behaviors of the LKSD test when the models are identical. Our test procedure provides no guarantee in this case, as the asymptotic distribution would deviate from the normal distribution. As in the previous section, we compare the performance of the tests with the three proposed variance estimators. In the following, we fix the significance level α at 0.05. As in Sections 3.4.1, 3.4.3, we choose perturbation parameters for the candidate models, and run 300 trials for a differing sample size $n \in \{100, 200, 300\}$. For both PPCA and LDA models, we choose $\delta_P = \delta_Q = 1$. Figure 3.J.6 shows the plot for each problem. We note that the U-statistic test has higher type-I error rates in this setting, although they are closed to the design level. Notwithstanding that our test assumptions are violated, the jackknife and V-statistic approaches reject the null at a rate well below 0.05, and remain conservative in this example.



(a) Identical PPCA models. (b) Identical LDA models.

Figure 3.J.6: Plots of type-I errors when two models are identical. Markers: ▽ (LKSD test with the jackknife variance estimator); ○ (LKSD test with the U-statistic variance estimator); □ (LKSD test with the V-statistic variance estimator). The LKSD test with the U-statistic variance estimator has higher errors than the nominal level $\alpha = 0.05$.

Chapter 4

Comparing latent variable models using the kernel Stein discrepancy – Part 2

This chapter addresses the same problem as in the previous chapter. We consider an alternative approach to the score function estimation discussed in Section 3.2.2. This short chapter examines the relative merit of the MCMC approach over this alternative. While the approach taken in this chapter is theoretically sound and straightforward to implement, it results in an extremely conservative test, and is therefore not useful in practice.

4.1 Introduction

For a distribution R and a latent variable model P defined by a density function $p(x) = \int p(x|z)\mathrm{d}P_Z(z)$, the previous chapter considered the estimation of KSD $(P\|R)$ using Fisher's identity and MCMC. Use of MCMC has its own disadvantages. One major shortcoming is that it requires the chain to converge to the stationarity; this may result in a long computational time, and it may not be simple to diagnose the convergence.

This chapter considers a more straightforward alternative. Instead of using the Fisher's identity, we directly estimate the numerator and the denominator of the score function $\mathbf{s}_p = p(x)^{-1}\nabla p(x)$. Specifically, assume that we have an i.i.d. latent sample $\{z_j\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} P_Z$, and we simply replace the marginal $p(x)$ with an empirical estimate,

$$p_m(x) := \frac{1}{m}\sum_{j=1}^{m} p(x|z_j),$$

yielding a plug-in score estimator

$$\mathbf{s}_{p_m}(x) = \frac{\nabla_x p_m(x)}{p_m(x)}.$$

As with the MCMC estimator, this score estimator is also biased; i.e., $\mathbb{E}_{Z_m}[\mathbf{s}_{p_m}(x)] \neq \mathbf{s}_p(x)$ for each $x \in \mathcal{X}$, where $Z_m = \{z_j\}_{j=1}^m$.

Given a sample $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$ (independent of the latent sample), this score estimate allows us to consider an approximate U-statistic of $\mathrm{KSD}^2\left(P\|R\right)$,

$$U_{n,m}(P) = \frac{1}{n(n-1)} \sum_{i \neq j} h_{p_m}(x_i, x_j), \tag{4.1}$$

where h_{p_m} is the Stein kernel introduced in Section 3.2.1 of the previous chapter, defined by the approximate density p_m. As in the previous chapter, we can consider the same U-statistic $U_{n,m}(Q)$ for another model Q and take the difference $U_{n,m}(P)$-$U_{n,m}(Q)$ to draw inferences about the KSD difference $\mathrm{KSD}^2\left(P\|R\right) - \mathrm{KSD}^2\left(Q\|R\right)$. Below, we show that the estimator of (4.1) has two issues: slow bias decay, and challenges in characterizing asymptotic variance of the null. For simplicity, the following discussion only deals with the continuous observation case. We follow the same notation as in Chapter 3.

4.2 U-statistics with random kernels

The technique to construct a model comparison test closely follows that of Chapter 3. We obtain an analogous result of asymptotic normality, where it is crucial to ensure the decay of the bias induced by the approximation. To facilitate our discussion on the estimator (4.1), we first present the advertised result. As in Section 3.3.1 of Chapter 3, we take $(\Omega, \mathcal{S}, \Pi)$ as an underlying probability space, and assume that all random variables are measurable functions from this space.

Theorem 4.1 (Asymptotic normality)**.** *Let $H_m : \Omega \to L^3(\mathcal{X} \times \mathcal{X}, R \otimes R)$ be a measurable random element independent of the data $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$. Let $U_{n,m}$ be a U-statistic defined by H_m and U_n be a U-statistic defined by a fixed U-statistic kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We assume that the random kernel H_m satisfies the convergences $\sigma^2_{H_m} \xrightarrow{\text{P}} \sigma^2 > 0$ and $\nu^3(H_m) \xrightarrow{\text{P}} \nu^3 < \infty$, where $\sigma^2_{H_m} = \mathrm{Var}_{x\sim R}\left[\mathbb{E}_{x'\sim R}[H_m(x,x')]\right]$ and $\nu^3(H_m) = \mathbb{E}_{x,x'\sim R\otimes R}\big|H_m(x,x') - \mathbb{E}_{x,x'\sim R\otimes R}[H_m(x,x')]\big|^3$. Let $\theta = \mathbb{E}_{x,x'\sim R\otimes R}[h(x,x')]$. Let $\theta(H_m) = \mathbb{E}_{x,x'\sim R\otimes R}[H_m(x,x')]$ the expectation of the random kernel with respect to the data distribution. Suppose that the two kernels H_m and h are related by the following condition: $Y_m := \sqrt{m}\big(\theta(H_m) - \theta\big)$ converges to a random variable Y in distribution. With $r_{n,m} = n/m \to r \in [0, \infty)$, we have*

$$\lim_{n,m\to\infty} \Pi\left[\sqrt{n}(U_{n,m} - \theta) < t\right] = \mathbb{E}_Y\left[\Phi\left(\frac{t - \sqrt{r}Y}{\sigma}\right)\right], \tag{4.2}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0,1)$.

Theorem 4.1 allows us to construct a hypothesis test for model comparison. However, there are two challenges associated with the use of the approximate U-statistic (4.1) and Theorem 4.1. First, the convergence $Y_m \xrightarrow{\text{d}} Y$ effectively requires that the bias $\mathbb{E}[U_{n,m}|H_m] - \theta$ induced by the random kernel H_m decays at a $1/\sqrt{m}$-rate. Second, Theorem 4.1 needs the existence of the random variable Y for which the RHS of (4.2) can be computed.

As discussed in Appendix 3.D, the bias decay depends on how fast the approximate score function $\mathbf{s}_{p_m}(x)$ converges to the target $\mathbf{s}_p(x)$. Establishing fast convergence is a challenging task for this ratio estimator, however. For simplicity, we illustrate the challenge in the case $D = 1$. By the law of large numbers, under suitable assumptions, we have almost-sure converges $p'_m(x) \to p'(x)$ and $p_m(x) \to p(x)$ for each $x \in \mathcal{X}$. This being said, the convergence rate may be slow, particularly when $p(x)$ is small. To see this, decompose the score difference as

$$\mathbf{s}_{p_m}(x) - \mathbf{s}_p(x) = \frac{p'_m(x) - p(x) + p(x)}{p(x)} \left(1 - \frac{p(x)}{p_m(x)}\right) + \left(\frac{p'_m(x) - p'(x)}{p(x)}\right).$$

We can see that both terms decay as $p'_m(x) \to p'(x)$ and $p_m(x) \to p(x)$. However, the division by $p(x)$ amplifies the errors; a smaller $p(x)$ makes it more challenging to attain a certain precision. This situation is likely to occur when the support of the model p substantially deviates from that of the data. Hence, the naive ratio estimator could result in a slow bias decay. In contrast, the MCMC approach taken in the previous chapter successfully avoids directly estimating the ratio.

Deriving a random variable Y with a tractable expectation (4.2) requires substantial effort. The bias mentioned above may be solved by a more sophisticated ratio estimator, but a complex estimator might render the task of finding Y more demanding. Fortunately for our ratio estimator, we can show that the scaled bias of Y_m follows a normal distribution asymptotically.

Lemma 4.2. *Let $p(x) = \int p(x|z)\mathrm{d}P_Z(z)$. Let $p_m(x) := \frac{1}{m}\sum_{j=1}^m p(x|z_j)$ with $\{z_j\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} P_Z$. Assume $\mathcal{X}$ is bounded and open. Assume that the set of likelihood functions $\mathcal{L} = \{p(x|\cdot)|x \in \mathcal{X}\}$ and their partial derivatives $\mathcal{PD}_d = \{\partial_d p(x|\cdot)|x \in \mathcal{X}\}$, $(d = 1, \dots, D)$ belong to P_Z-Donsker classes [van der Vaart, 2000, Section 19.2], where ∂_d denotes the partial derivative operator with respect to the d-th coordinate. Assume $\partial_d p \in L^2(\mathcal{X}, R)$ for each $d = 1, \dots, D$. Assume $\inf_{x\in\mathcal{X}} p(x) > 0$. Then,*

$$\sqrt{m}\big(\mathbb{E}_{x,x'\sim R\otimes R}[h_{p_m}(x,x')] - \mathbb{E}_{x,x'\sim R\otimes R}[h_p(x,x')]\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \gamma_p^2),$$

where the convergence is in the sense of outer integrals [van der Vaart, 2000, Section 18.2]. Here the variance γ_p^2 is defined by

$$\gamma_p^2 = \mathrm{Var}_{f\sim G}\left[\left(\sum_{d=1}^{D}(L_{1,d}f_d + L_{2,d}f_{D+1})\right)\left(\sum_{d'=1}^{D}(L_{1,d'}f_{d'} + L_{2,d'}f_{D+1})\right)\right],$$

where $L_{1,d}, L_{2,d}$ are linear operators defined in the proof (see Equation 4.4), $f = (f_1, \dots, f_{D+1}) \sim G$ with $G = \{G_x\}_{x\in\mathcal{X}}$ a zero-mean multivariate Gaussian process defined by the covariance function

$$\mathrm{Cov}\Big[G_{x,d}, G_{x',d'}\Big] = \mathbb{E}_{Z\sim P_Z}\Big[V_{p,d}(x|Z)V_{p,d'}(x'|Z)\Big] - \mathbb{E}_{Z\sim P_Z}\Big[V_{p,d}(x|Z)\Big]\mathbb{E}_{Z\sim P_Z}\Big[V_{p,d'}(x'|Z)\Big],$$

where $V_{p,d}(x|z)$ denotes the d-th component of $V_p(x|z) = (\partial_{x_1}p(x|z), \dots, \partial_{x_D}p(x|z), p(x|z))$.

This result makes additional assumptions on the model p. The most stringent is $\inf_{x\in\mathcal{X}} p(x) >$

0, which rules out distributions supported on the entire space $\mathbb{R}^D$. In contrast, the MCMC approach can handle this situation, as showcased in the experiment with PPCA models. The P_Z-Donsker assumption states that a uniform central limit theorem holds for $\mathcal{L}$ and $\mathcal{PD}_d$ $(d = 1, \ldots, D)$; i.e. the empirical process with index sets $\mathcal{L}$ and $\mathcal{PD}_d$ converges to a Gaussian process. This condition requires regularity assumptions on the likelihood and the partial derivatives, such as a variant of Lipschitz-continuity for bounded $\mathcal{X}$ [van der Vaart, 2000, Example 19.7]. For details, we refer the reader to van der Vaart [2000, Chapter 19].

Unfortunately, the normal approximation of Lemma 4.2 may be inadequate. The variance γ_p^2 in Lemma 4.2 involves an integral with respect to z, where the integrand depends on the reciprocal of the likelihood $1/p(x|z)$. The variance could therefore be enormous due to the reciprocal dependence. Moreover, estimating this quantity is challenging, as the reciprocal can easily blow up the estimator (we observed this trend in our preliminary experiment). The following result illustrates the role of the variance γ_p^2 in constructing a test.

Corollary 4.3. *Assume the latent samples for P and Q are independent. Let $U_{n,m}(P, Q) = U_{n,m}(P) - U_{n,m}(Q)$. Let $h_{p,q}(x, x') = h_p(x, x') - h_q(x, x')$. Assume that the kernels h_{p_m,q_m} and $h_{p,q}$ satisfy the conditions in Theorem 4.1. Assume $n/m \to r \in [0, \infty)$. Then, under the conditions given in Lemma 4.2, we have $\lim_{n,m\to\infty} \sqrt{n}(U_{n,m}(P, Q) - \mu_{P,Q}) \to \mathcal{N}(0, c^2)$ with $c = \sigma_{h_{p,q}}\sqrt{1 + r\rho^2}$, $\rho^2 = (\gamma_p^2 + \gamma_q^2)/\sigma^2_{h_{p,q}}$, $\mu_{P,Q} = \mathrm{KSD}\,(P\|R)^2 - \mathrm{KSD}\,(Q\|R)^2$, and $\sigma^2_{h_{p,q}} = \mathrm{Var}_{x\sim R}\big[\mathbb{E}_{x'\sim R}[h_{p,q}(x, x')]\big]$.*

When $r > 0$, the variance $\gamma_p^2 + \gamma_q^2$ coming from the normal approximation (Lemma 4.2) amplifies the asymptotic variance $\sigma^2_{h_{p,q}}$ of the U-statistic without score approximations. Therefore, the utility of the test critically depends on the additional variance $\gamma_p^2 + \gamma_q^2$. A large value of this variance therefore makes the test conservative.

4.3 Conclusion

In this section, we have considered a ratio estimator of the score function derived from the empirical approximation of the marginal density. The resulting KSD U-statistic provides a similar test based on the asymptotic normality of the statistic. However, we have shown that the test has the following shortcomings compared to the MCMC approach in Chapter 3. First, the bias of the KSD estimate may decay undesirably slowly, thereby requiring a large number of latent sample points for the approximation. Second, the resulting test applies in theory only to densities bounded away from zero over the domain. Finally, even in bounded domains, the test threshold is challenging to estimate and may be overly conservative. These observations support the test proposed in Chapter 3.

4.A Proof of asymptotic normality of random kernel U-statistics

Theorem 4.1 (Asymptotic normality)**.** *Let $H_m : \Omega \to L^3(\mathcal{X} \times \mathcal{X}, R \otimes R)$ be a measurable random element independent of the data $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$. Let $U_{n,m}$ be a U-statistic defined by H_m and U_n be a U-statistic defined by a fixed U-statistic kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We assume that the random kernel H_m satisfies the convergences $\sigma^2_{H_m} \xrightarrow{\text{p}} \sigma^2 > 0$ and $\nu^3(H_m) \xrightarrow{\text{p}} \nu^3 < \infty$, where $\sigma^2_{H_m} = \operatorname{Var}_{x\sim R}\left[\mathbb{E}_{x'\sim R}[H_m(x,x')]\right]$ and $\nu^3(H_m) = \mathbb{E}_{x,x'\sim R\otimes R}\big|H_m(x,x') - \mathbb{E}_{x,x'\sim R\otimes R}[H_m(x,x')]\big|^3$. Let $\theta = \mathbb{E}_{x,x'\sim R\otimes R}[h(x,x')]$. Let $\theta(H_m) = \mathbb{E}_{x,x'\sim R\otimes R}[H_m(x,x')]$ the expectation of the random kernel with respect to the data distribution. Suppose that the two kernels H_m and h are related by the following condition: $Y_m := \sqrt{m}\big(\theta(H_m) - \theta\big)$ converges to a random variable Y in distribution. With $r_{n,m} = n/m \to r \in [0,\infty)$, we have*

$$\lim_{n,m\to\infty} \Pi\left[\sqrt{n}(U_{n,m} - \theta) < t\right] = \mathbb{E}_Y\left[\Phi\left(\frac{t - \sqrt{r}Y}{\sigma}\right)\right], \tag{4.2}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0,1)$.

Proof. Recall that $(\Omega, \mathcal{S}, \Pi)$ is the underlying probability space, and H_m, $U_{n,m}$, and U_n are random variables on it. The probability on the LHS can be expressed as

Proof.

$$\begin{aligned}
&\Pi\left[\sqrt{n}(U_{n,m} - \theta) < t\right] \\
&= \mathbb{E}_{H_m}\left[\Pi\left(\sqrt{n}\left(\frac{U_{n,m} - \theta(H_m)}{\sigma_{H_m}}\right) < -\sqrt{\frac{n}{m}}\frac{Y_m}{\sigma_{H_m}} + \frac{t}{\sigma_{H_m}}\,\middle|\, H_m\right)\right] \\
&= \mathbb{E}_{H_m}\left[F_{n|H_m}\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right)\right],
\end{aligned}$$

where $F_{n|H_m}$ denotes the CDF of $\sqrt{n}\{U_{n,m} - \theta(H_m)\}/\sigma_{H_m}$ conditioned on H_m. The difference between the two quantities in Equation (4.2) is bounded as follows:

$$\begin{aligned}
&\lim_{n,m\to\infty}\left|\mathbb{E}_{H_m}F_{n|H_m}\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right) - \mathbb{E}_Y\Phi\left(\frac{t - \sqrt{r}Y}{\sigma_h}\right)\right| \\
&\le \lim_{n,m\to\infty}\mathbb{E}_{H_m}\left|F_{n|H_m}\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right) - \Phi\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right)\right| \\
&\quad + \lim_{n,m\to\infty}\left|\mathbb{E}_{H_m}\Phi\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right) - \mathbb{E}_Y\Phi\left(\frac{t - \sqrt{r}Y}{\sigma_h}\right)\right|.
\end{aligned} \tag{4.3}$$

□

For the second term, by Slutsky's theorem [van der Vaart, 2000, p.11], we have

$$\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}} \xrightarrow{\text{d}} \frac{t - \sqrt{r}Y}{\sigma_h}$$

as $m \to \infty$. Therefore, the second term in Equation (4.3) also converges to zero by the fact the CDF of a Gaussian distribution is bounded and continuous and by the definition of weak convergence.

The first term is dealt with as follows. For some $\delta > 0$, let A_m be the event $\{|\nu^3(H_m) - \nu(h)| < \delta\}$ and $\bar{A}_m$ denote its complement. Similarly, let B_m be the event $\{|\sigma_{H_m} - \sigma| < \epsilon\}$ for $0 < \epsilon < \sigma_h$. Then, the first term is bounded by

$$\begin{aligned}
&\lim_{n,m\to\infty} \mathbb{E}_{H_m} \left| F_{n|H_m}\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right) - \Phi\left(\frac{t - \sqrt{r_{n,m}}Y_m}{\sigma_{H_m}}\right)\right| \cdot 1_{A_m\cap B_m}(H_m)\\
&\leq \lim_{n,m\to\infty} \sup_{u\in\mathbb{R}} \mathbb{E}_{H_m}\big|F_{n|H_m}(u) - \Phi(u)\big| 1_{A_m\cap B_m}(H_m),
\end{aligned}$$

where we have the fact that the integrand is bounded and $\lim_{m\to\infty}\Pi(\bar{A}_m \cup \bar{B}_m) = 0$. By the Berry-Esseen bound for U-statistics [Callaert and Janssen, 1978], the expectation on the RHS is then

$$\begin{aligned}
\mathbb{E}_{H_m}\left|F_{n|H_m}(u) - \Phi(u)\right| 1_{A_m\cap B_m}(H_m) &\leq Cn^{-\frac{1}{2}}\mathbb{E}_{H_m}\left[\frac{\nu_3(H_m)}{\sigma^3_{H_m}} \cdot 1_{A_m\cap B_m}(H_m)\right]\\
&< \frac{C(\nu^3+\delta)}{(\sigma-\epsilon)^3}\cdot n^{-\frac{1}{2}},
\end{aligned}$$

where C is a universal constant. The RHS thus goes to zero as $n \to \infty$, which concludes the proof.

□

4.B Details of the delta method

We prove two preliminary lemmas.

Lemma 4.4. *Let $\ell^\infty_{>0}(\mathcal{X}) = \{f \in \ell^\infty(\mathcal{X}) : \inf_{x\in\mathcal{X}} f(x) > 0\}$ be a subset of the space $\ell^\infty(\mathcal{X})$ of bounded functions on $\mathcal{X}$ equipped with the norm $\|f\|_\infty = \sup_{x\in\mathcal{X}}|f(x)|$. Assume $p' \in L^2(\mathcal{X}, R)$. Assume $\delta = \inf_{x\in\mathcal{X}} p(x) > 0$. Then, the map $s : L^2(\mathcal{X}, R) \times \ell^\infty_{>0}(\mathcal{X}) \to L^2(\mathcal{X}, R)$ defined by $s(f, g)(x) = f(x)/g(x)$ is Hadamard differentiable at $\theta = (p', p)$ tangentially to $\ell^\infty(\mathcal{X}) \times \ell^\infty(\mathcal{X})$; Its Hadamard derivative is given by $s'_\theta(h_1,h_2) = h_1/p - h_2p'/p^2$. In particular, $s'_\theta(\ell^\infty \times \ell^\infty) \subset L^2(\mathcal{X}, R)$.*

Proof. By the assumption on p' and p,we have the following: (a) $\theta = (p, p')$ is included in the domain of s, and (b) the range of s'_θ is included in $L^2(\mathcal{X}, R)$. By the definition of the domain of s, we have the range is included in $L^2(\mathcal{X}, R)$. In the following, we verify the Hadamard differentiability and derive the derivative. For a converging sequence $t_n \to 0$, take arbitrary sequences $\{h_{1,n}\}^\infty_{n=1} \subset L^2(\mathcal{X}, R)$ and $\{h_{2,n}\}^\infty_{n=1} \subset \ell^\infty(\mathcal{X})$ such that $h_{1,n} \to h_1 \in L^2(\mathcal{X}, R)$ and $h_{2,n} \to h_2 \in \ell^\infty(\mathcal{X})$, respectively, as $n \to \infty$. We prove

$$\lim_{n\to\infty}\left\|\frac{s(p' + t_nh_{1,n}, p + t_nh_{2,n}) - s(p', p)}{t_n} - s'_\theta(h_1,h_2)\right\|_2 = 0.$$

As we have

$$\frac{s(p' + t_n h_{1,n}, p + t_n h_{2,n}) - s(p', p)}{t_n} = \frac{h_{1,n}}{p + t_n h_{2,n}} + p' \left(\frac{-h_{2,n}}{p(p + t_n h_{2,n})} \right),$$
$$s'_\theta(h_1, h_2) = \frac{h_1}{p} - p' \frac{h_2}{p^2},$$

it suffices to show

$$\left\| \frac{h_{1,n}}{p + t_n h_{2,n}} - \frac{h_1}{p} \right\|_2 \to 0, \left\| \frac{p'}{p} \left(\frac{h_{2,n}}{p + t_n h_{2,n}} - \frac{h_2}{p} \right) \right\|_2 \to 0 \ (n \to \infty).$$

Note that $p + t_n h_{2,n} \in \ell^\infty_{>0}(\mathcal{X})$ is guaranteed if we take sufficiently large n. For any $\varepsilon_1 > 0$, there exists $n_{1,0} \in \mathbb{N}$ such that $\|h_{1,n} - h_1\|_2 < \varepsilon_1$ for $n \geq n_{1,0}$. Similarly, for any $\varepsilon_2 > 0$, we can take $n_{2,0} \in \mathbb{N}$ such that $\|h_{2,n} - h_2\|_\infty < \varepsilon_2$ for $n \geq n_{2,0}$. Let $n_0 \geq 1$ such that $t_n \leq (\delta - b)/(\varepsilon_2 + M_2) \wedge \varepsilon_1 \wedge \varepsilon_2$ for $n \geq n_0$, where $M_2 = \|h_2\|_\infty$ and $0 < b < \delta$. Then, taking $n \geq \max(n_{2,0}, n_0)$ gives $p(x) + t_n h_{2,n}(x) \geq b$ for all $x \in \mathcal{X}$. Thus, for $n \geq \max(n_0, n_{1,0}, n_{2,0})$, with $M_1 = \|h_1\|_\infty$, we have

$$\begin{aligned}
\left\| \frac{h_{1,n}}{p + t_n h_{2,n}} - \frac{h_1}{p} \right\|_2 &\leq \left\| \frac{(h_{1,n} - h_1)}{p + t_n h_{2,n}} \right\|_2 + t_n \left\| \frac{h_1(h_{2,n} - h_2)}{p(p + th_{2,n})} \right\|_2 + t_n \left\| \frac{h_1 h_2}{p(p + th_{2,n})} \right\|_2 \\
&\leq \frac{\epsilon_1}{b} + \frac{t_n M_1}{b\delta} (\epsilon_2 + M_2) \\
&\leq \frac{\epsilon_1}{b} + \frac{M_1}{b\delta} (\varepsilon_2 + M_2)(\varepsilon_1 \wedge \varepsilon_2)
\end{aligned}$$

and

$$\begin{aligned}
&\left\| \frac{p'}{p} \left(\frac{h_{2,n}}{p + t_n h_{2,n}} - \frac{h_2}{p} \right) \right\|_2 \\
&\leq \left\| \frac{p'}{p} \left(\frac{h_{2,n} - h_2}{p + t_n h_{2,n}} \right) \right\|_2 + t_n \left\| \frac{p'}{p} \frac{h_2(h_{2,n} - h_2)}{p(p + t_n h_{2,n})} \right\| + t_n \left\| \frac{p'}{p} \frac{h_2^2}{p(p + t_n h_{2,n})} \right\|_2 \\
&\leq \frac{\|p'\|_2}{\delta b} \left(\epsilon_2 + \frac{t_n M_2}{\delta} (\epsilon_2 + M_2) \right) \\
&\leq \frac{\|p'\|_2}{\delta b} \left(\epsilon_2 + \frac{M_2}{\delta} (\epsilon_2 + M_2)(\varepsilon_1 \wedge \varepsilon_2) \right).
\end{aligned}$$

Thus, the convergence of the two terms have been proved. □

Lemma 4.5. *For a differentiable kernel $k \in C^{(1,1)} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define its partial derivative $k_{2,d}(x, x') = \partial_{x'_d} k(x, x')$ with respect to the d-th coordinate of the second argument. Let*

$$\psi_d(f) := \langle f, T_k f \rangle_{L^2(R)} + 2 \left\langle f, T_{k_{2,d}} 1 \right\rangle_{L^2(R)},$$

where $T_k f(\cdot) = \mathbb{E}_{x' \sim R}[k(\cdot, x') f(x')]$ and $T_{k_{2,d}} 1(\cdot) = \mathbb{E}_{x' \sim R}[k_{2,d}(\cdot, x')]$. Assume $k \in L^2(\mathcal{X} \times \mathcal{X}, R \otimes R)$ and $\mathbb{E}_{x \sim R}[k_{2,d}(\cdot, x')] \in L^2(\mathcal{X}, R)$. Then, $\psi_d : L^2(\mathcal{X}, R) \to L^2(\mathcal{X}, R)$ is Hadamard *differentiable at any $f \in L^2(\mathcal{X}, R)$, and its derivative $\psi'_{d,f} : L^2(\mathcal{X}, R) \to \mathbb{R}$ is given by*

$$\psi'_{d,f}(h) = 2 \langle h, T_k f \rangle_{L^2(R)} + 2 \left\langle h, T_{k_{2,d}} 1 \right\rangle_{L^2(R)} .$$

Proof. Take any $h \in L^2(\mathcal{X}, R)$. As we have

$$\frac{\psi_d(f + t_n h_n) - \psi_d(f)}{t_n} = 2 \langle h_n, T_k f \rangle_{L^2(R)} + 2 \left\langle h_n, T_{k_{2,d}} 1 \right\rangle_{L^2(R)} + t_n \langle h_n, T_k h_n \rangle_{L^2(R)} ,$$

for any sequence $\{h_n\}_{n \geq 1} \subset L^2(\mathcal{X}, R)$ such that $h_n \to h$ in $L^2(\mathcal{X}, R)$ and $t_n > 0$. Thus, if $t_n \to 0$, we have

$$\begin{aligned} \left| \frac{\psi_d(f + t_n h_n) - \psi_d(f)}{t_n} - \psi'_{d,f}(h) \right| &\leq 2 \|h_n - h\|_{L^2(\mathcal{X},R)} \Big(\|T_k f\|_{L^2(\mathcal{X},R)} + \|T_{k_{2,d}} 1\|_{L^2(\mathcal{X},R)} \Big) \\ &\quad + t_n \sqrt{\mathbb{E}_{x,x' \sim R \otimes R}[k(x,x')^2]} \|h_n\|^2_{L^2(\mathcal{X},R)} \\ &\to 0 \ (n \to \infty). \end{aligned}$$

□

Proof of Lemma 4.2

Proof. We use the functional delta method [van der Vaart, 2000, p.291]. We first rewrite the KSD $\mathrm{KSD}^2 (P \| R) = \mathbb{E}_{x,x' \sim R \otimes R}[h_p(x, x')]$ as a functional of the density p and its partial derivatives. Let $\ell^\infty(\mathcal{X})$ be the set of all bounded functions equipped with the supremum norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Let $\ell^\infty_{>0}$ be its subset of functions having positive minimum values. Define $s : L^2(\mathcal{X}, R) \times \ell^\infty_{>0}(\mathcal{X}) \to L^2(\mathcal{X}, R)$ by $s(f, g)(x) := f(x)/g(x)$ and $\psi_d : L^2(R) \to \mathbb{R}$ by

$$\begin{aligned} \psi_d(h) &:= \mathbb{E}_{x,x'} \left\{ f(x) f(x') k(x, x') + f(x) k_{1,d}(x', x) + f(x') k_{1,d}(x, x') \right\} \\ &= \langle h, T_k h \rangle_{L^2(R)} + 2 \left\langle h, T_{k_{2,d}} 1 \right\rangle_{L^2(R)} , \end{aligned}$$

where we have used $k_{1,d}(a, b) = k_{2,d}(b, a)$. Here, $T_k : L^2(R) \to L^2(R)$ is the integral operator given by $T_k f(\cdot) = \mathbb{E}_{x \sim R}\big[k(\cdot, x) f(x)\big]$ [Steinwart and Christmann, 2008, Theorem 4.27], and $\langle f, g \rangle_{L^2(R)} = \mathbb{E}_{x \sim R}[f(x) g(x)]$. Then, the expectation $\mathbb{E}_{x,x' \sim R \otimes R}[h_p(x, x')]$ is written as a functional as follows:

$$\mathbb{E}_{x,x' \sim R \otimes R}[h_p(x, x')] = \sum_{d=1}^{D} \psi_d(s(\partial_d p, p)) + \text{const.},$$

where the second term involves second derivatives of the kernel and is constant in p and $\partial_d p$. By Lemmas 4.4, 4.5 and the chain rule of Hadamard differentiation [van der Vaart, 2000, Theorem 20.9], the composite functional $\psi_d \circ s : l^\infty(\mathcal{X}) \times D_0 \to \mathbb{R}$ is Hadamard differentiable at $\theta_d = (\partial_d p, p)$ tangentially to $l^\infty(\mathcal{X}) \times l^\infty(\mathcal{X})$ for $d = 1, \ldots, D$. The derivative is given by

$$(\psi_d \circ s)'_{\theta_d}(t_d, t_{D+1}) = 2 \left\langle s'_{\theta_d}(t_d, t_{D+1}), T_k s(\theta_d) \right\rangle_{L^2(R)} + 2 \left\langle s'_{\theta_d}(t_d, t_{D+1}), T_{k_{2,d}} 1 \right\rangle_{L^2(R)}$$

$$\begin{aligned}&= 2\left\langle \frac{1}{p}t_d, T_k s(\theta_d) + T_{k_{2,d}}1 \right\rangle_{L^2(R)} - 2\left\langle \frac{\partial_d p}{p^2} t_{D+1}, T_k s(\theta_d) + T_{k_{2,d}}1 \right\rangle_{L^2(R)} \qquad (4.4)\\ &=: L_{1,d}t_d + L_{2,d}t_{D+1}.\end{aligned}$$

Therefore, the derivative of $\Psi := \sum_d \psi_d\{s(\partial_d p, p)\}$ at $\theta = (\partial_1 p, \dots, \partial_D p, p)$ is given as $\Psi'_\theta(t_1, \dots, t_{D+1}) = \sum_d (\psi_d \circ s)'_{\theta_d}(t_d, t_{D+1})$ for $(t_1, \dots, t_{D+1}) \in \prod_{d=1}^{D+1} \ell^\infty(\mathcal{X})$. Let $V_p(x) = (\partial_1 p(x), \dots, \partial_D p(x), p(x))$ denote the concatenation of the density p and its partial derivatives. Note that we can write

$$V_{p_m}(x) = \frac{1}{m}\sum_{j=1}^m p(x|z_j),\ V_p(x) = \mathbb{E}_Z\big[p(x|Z)\big],$$

which defines an empirical process $\Big\{\sqrt{m}\big(V_{p_m}(x) - V_p(x)\big)\Big\}_{x\in\mathcal{X}}$. In the limit of $m \to \infty$, by the definition of P_Z-Donsker class [van der Vaart, 2000, Section 19.2], we have the convergence

$$\big\{\sqrt{m}\big(V_{p_m}(x) - V_p(x)\big)\big\}_{x\in\mathcal{X}} \xrightarrow{\text{d}} G,$$

where G is the zero-mean multivariate P_Z- brownian bridge process with covariance function

$$\operatorname{Cov}\Big[G_{x,d}, G_{x',d'}\Big] = \mathbb{E}_{Z\sim P_Z}\Big[V_{p,d}(x|Z)V_{p,d'}(x'|Z)\Big] - \mathbb{E}_{Z\sim P_Z}\Big[V_{p,d}(x|Z)\Big]\mathbb{E}_{Z\sim P_Z}\Big[V_{p,d'}(x'|Z)\Big].$$

In consequence, by the functional delta method [van der Vaart, 2000, Theorem 20.8], we have

$$\begin{aligned}&\sqrt{m}\big(\mathbb{E}_{x,x'\sim R\otimes R}[h_{p_m}(x,x')] - \mathbb{E}_{x,x'\sim R\otimes R}[h_p(x,x')]\big)\\ &= \sqrt{m}\left(\Psi(V_{p_m}) - \Psi(V_p)\right) \xrightarrow{\text{d}} \mathcal{N}(0, \operatorname{Var}[\Psi'_\theta(G)]).\end{aligned}$$

Specifically, with $f = (f_1, \dots, f_{D+1}) \sim G$, the variance is given by

$$\operatorname{Var}[\Psi'_\theta(G)] = \operatorname{Var}_{f\sim G}\left[\left(\sum_{d=1}^D (L_{1,d}f_d + L_{2,d}f_{D+1})\right)\left(\sum_{d'=1}^D (L_{1,d'}f_{d'} + L_{2,d'}f_{D+1})\right)\right]. \qquad (4.5)$$

□

4.C Proof of the asymptotic normality of the test statistic

Proof of Corollary 4.3

Proof. Apply Theorem 4.1 with $H_m(x,x') = h_{p_m,q_m}(x,x')$ and $h = h_{p,q}$. In this case, Y_m is given by

$$Y_m = \underbrace{\sqrt{m}\left(\mathbb{E}\big[U_{n,m}(P)|h_{p_m}\big] - \mathrm{KSD}\,(P\|R)^2\right)}_{Y_m^{(p)}} - \underbrace{\sqrt{m}\left(\mathbb{E}\big[U_{n,m}(Q)|h_{q_m}\big] - \mathrm{KSD}\,(Q\|R)^2\right)}_{Y_m^{(q)}}.$$

By Lemma 4.2, $Y_m^{(p)}$ and $Y_m^{(q)}$ converge to $\mathcal{N}(0, \gamma_p^2)$ and $\mathcal{N}(0, \gamma_q^2)$, respectively. The two variables are independent of each other, so the difference between them converges to a normal variable $Y \sim \mathcal{N}(0, \gamma_p^2 + \gamma_q^2)$. As $-Y \sim \mathcal{N}(0, \gamma_p^2 + \gamma_q^2)$, we have

$$\lim_{n,m\to\infty} \Pi\left[\sqrt{n}(U_{n,m}(P,Q) - \mu_{P,Q}) < t\right] = \mathbb{E}_Y\left[\Phi\left(-\frac{\sqrt{r}Y}{\sigma_{h_{p,q}}} + \frac{t}{\sigma_{h_{p,q}}}\right)\right]$$
$$= \frac{1}{\sqrt{2\pi}} \int \Phi\left(\sqrt{r}\rho y + \frac{t}{\sigma_{h_{p,q}}}\right) e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi\left(\frac{t}{\sigma_{h_{p,q}}\sqrt{1 + r\rho^2}}\right).$$

□

Chapter 5

Interpretable features for model comparison

Summary Given two candidate models, and a set of target observations, we address the problem of measuring the relative goodness of fit of the two models. We propose two new statistical tests which are nonparametric, computationally efficient (runtime complexity is linear in the sample size), and interpretable. As a unique advantage, our tests can produce a set of examples (informative features) indicating the regions in the data domain where one model fits significantly better than the other. In a real-world problem of comparing GAN models, the test power of our new test matches that of the state-of-the-art test of relative goodness of fit, while being one order of magnitude faster.

5.1 Introduction

One of the most fruitful areas in recent machine learning research has been the development of effective generative models for very complex and high dimensional data. Chief among these have been the generative adversarial networks [Goodfellow et al., 2014, Nowozin et al., 2016, Arjovsky et al., 2017], where samples may be generated without an explicit generative model or likelihood function. A related thread has emerged in the statistics community with the advent of Approximate Bayesian Computation, where simulation-based models without closed-form likelihoods are widely applied in bioinformatics applications [see Lintusaari et al., 2017, for a review]. In these cases, we might have several competing models, and wish to evaluate which is the better fit for the data.

The problem of model criticism is traditionally defined as follows: how well does a model Q fit a given sample $Z_n := \{z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$? This task can be addressed in two ways: by comparing samples $Y_n := \{y_i\}_{i=1}^n$ from the model Q and data samples, or by directly evaluating the goodness of fit of the model itself. In both of these cases, the tests have a null hypothesis (that the model agrees with the data), which they will reject given sufficient evidence. Two-sample tests fall into the first category: there are numerous nonparametric tests which may be used [Alba Fernández et al., 2008, Friedman and Rafsky, 1979, Gretton et al., 2012a, Székely and Rizzo, 2004, Rosenbaum, 2005, Harchaoui et al., 2008, Hall and Tajvidi, 2002, Jitkrittum et al., 2016], and recent work in applying two-sample tests to the problem of model criticism [Lloyd and Ghahramani, 2015]. A second approach requires the model density q explicitly. In the case of simple models for which normalization is not an issue (e.g., checking for Gaussianity), several tests exist [Baringhaus and Henze, 1988, Székely and Rizzo, 2005]; when a model density is known only up to a normalization constant, tests of goodness of fit have been developed using a Stein-based divergence [Chwialkowski et al., 2016, Liu et al., 2016, Jitkrittum et al., 2017b].

An issue with the above notion of model criticism, particularly in the case of modern generative models, is that *any* hypothetical model Q that we design is likely a poor fit to the data. Indeed, as noted in Yamada et al. [2019, Section 5.5], comparing samples from various Generative Adversarial Network (GAN) models [Goodfellow et al., 2014] to the reference sample Z_n by a variant of the Maximum Mean Discrepancy (MMD) test [Gretton et al., 2012a] leads to the trivial conclusion that all models are wrong [Box, 1976], i.e., $H_0\colon Q = R$ is rejected by the test in all cases. A more relevant question in practice is thus: "Given two models P and Q, which is closer to R, and **in what ways**?" This is the problem we tackle in this chapter.

To our knowledge, the only nonparametric statistical test of *relative* goodness of fit is the $\mathrm{Rel\text{-}MMD}$ test of Bounliphone et al. [2016], based on the maximum mean discrepancy [MMD, Gretton et al., 2012a]. While shown to be practical (e.g., for comparing network architectures of generative networks), two issues remain to be addressed. Firstly, its runtime complexity is quadratic in the sample size n, meaning that it can be applied only to problems of moderate size. Secondly and more importantly, it does not give an indication of where one model is better than the other. This is essential for model comparison: in practical settings, it is highly unlikely that one model will be uniformly better than another in all respects: for instance, in hand-written digit generation, one model might produce better "3"s, and the other better "6"s. The ability to produce a few examples which indicate regions (in the data domain) in which one model fits better than the other will be a valuable tool for model comparison. This type of interpretability is useful especially in learning generative models with GANs, where the "mode collapse" problem is widespread [Salimans et al., 2016, Srivastava et al., 2017]. The idea of generating such distinguishing examples (so called *test locations*) was explored in Jitkrittum et al. [2016, 2017b] in the context of model criticism and two-sample testing.

In this chapter, we propose two new linear-time tests for relative goodness-of-fit. In the first test, the two models P, Q are represented by their two respective samples X_n and Y_n, and the test generalizes that of Jitkrittum et al. [2016]. In the second, the test has access to the probability density functions p, q of the two respective candidate models P, Q (which need only be known up to normalization), and is a three-way analogue of the test of Jitkrittum et al. [2017b]. In both cases, the tests return locations indicating where one model outperforms the other. We emphasize that the practitioner must choose the model ordering, since as noted earlier, this will determine the locations that the test prioritizes. We further note that the two tests complement each other, as both address different aspects of the model comparison problem. The first test simply finds the location where the better model produces mass closest to the test sample: a worse model can produce too much mass, or too little. The second test does not address the overall probability mass, but rather the shape of the model density: specifically, it penalizes the model whose derivative log density differs most from the target (the interpretation is illustrated in our experiments). In the experiment on comparing two GAN models, we find that the performance of our new test matches that of Rel-MMD while being one order of magnitude faster. Further, unlike the popular Fréchet Inception Distance (FID) [Heusel et al., 2017] which can give a wrong conclusion when two GANs have equal goodness of fit, our proposed method has a well-calibrated threshold, allowing the user to flexibly control the false positive rate.

5.2 Measures of Goodness of Fit

In the proposed tests, we test the relative goodness of fit by comparing the relative magnitudes of two distances, following Bounliphone et al. [2016]. More specifically, let $d(P, R)$ be a discrepancy measure between P and R. Then, the problem can be formulated as a hypothesis test proposing $H_0 \colon d(P, R) \leq d(Q, R)$ against $H_1 \colon d(P, R) > d(Q, R)$. This is the approach taken by Bounliphone et al. who use the MMD as d, resulting in the relative MMD test (Rel-MMD); we have considered the kernel Stein discrepancy [Chwialkowski et al., 2016, Liu et al., 2016] as d in Chapter 3. The proposed Rel-UME and Rel-FSSD tests are based on two recently proposed discrepancy measures for d: the Unnormalized Mean Embeddings (UME) statistic [Chwialkowski et al., 2015, Jitkrittum et al., 2016], and the Finite-Set Stein Discrepancy (FSSD) [Jitkrittum et al., 2017b], for the sample-based and density-based settings, respectively. We first review UME and FSSD. We will extend these two measures to construct two new relative goodness-of-fit tests in Section 5.3. We assume throughout that the probability measures P, Q, R have a common support $\mathcal{X} \subseteq \mathbb{R}^D$.

The Unnormalized Mean Embeddings (UME) statistic. UME is a (random) distance between two probability distributions [Chwialkowski et al., 2015] originally proposed for two-sample testing for $H_0 \colon Q = R$ and $H_1 \colon Q \neq R$. Let $k_Y \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Let μ_Q be the mean embedding of Q, and is defined such that $\mu_Q(w) := \mathbb{E}_{y \sim Q}[k_Y(y, w)]$ (assumed to exist) [Smola et al., 2007]. Gretton et al. [2012a] shows that when k_Y is characteristic [Sriperumbudur et al., 2011], the Maximum Mean Discrepancy (MMD) *witness function* $\mathrm{wit}_{Q,R}(w) := \mu_Q(w) - \mu_R(w)$ is a zero function if and only if $Q = R$. Based on this fact, the UME statistic evaluates the squared witness function at J_q test locations $W := \{w_j\}_{j=1}^{J_q} \subset \mathcal{X}$ to determine whether it is zero. Formally, the population squared UME statistic is defined as

$$U^2(Q, R) := \frac{1}{J} \sum_{j=1}^{J} (\mu_Q(w_j) - \mu_R(w_j))^2.$$

For our purpose, it will be useful to rewrite the UME statistic as follows. Define the feature function

$$\psi_W(y) := \frac{1}{\sqrt{J_q}} \left(k_Y(y, w_1), \ldots, k_Y(y, w_{J_q}) \right)^\top \in \mathbb{R}^{J_q}.$$

Define the feature expectation $\psi_W^Q := \mathbb{E}_{y \sim Q}[\psi_W(y)]$ with respect to Q, and its empirical estimate $\hat{\psi}_W^Q := n^{-1} \sum_{i=1}^n \psi_W(y_i)$. The squared population UME statistic is equivalent to $U^2(Q, R) := \|\psi_W^Q - \psi_W^R\|_2^2$. For $W \sim \eta$ where η is a distribution with a density, Theorem 2 of

Chwialkowski et al. [2015] states that if k_Y is real analytic, integrable, and characteristic, then η-almost surely $\|\psi_W^Q - \psi_W^R\|_2^2 = 0$ if and only if $Q = R$. In words, under the stated conditions, $U(Q, R) := U_Q$ defines a distance between Q and R (almost surely).[1] A consistent unbiased estimator is

$$\widehat{U_Q^2} = \frac{1}{n(n-1)}\left[\left\|\sum_{i=1}^{n}\{\psi_W(y_i) - \psi_W(z_i)\}\right\|_2^2 - \sum_{i=1}^{n}\|\psi_W(y_i) - \psi_W(z_i)\|_2^2\right],$$

which clearly can be computed in $\mathcal{O}(n)$ time. Jitkrittum et al. [2016] proposed optimizing the test locations W and k_Y so as to maximize the test power (i.e., the probability of rejecting H_0 when it is false) of the two-sample test with the normalized version of the UME statistic. It was shown that the optimized locations give an interpretable indication of where Q and R differ in the input domain $\mathcal{X}$.

The Finite-Set Stein Discrepancy (FSSD). FSSD is a discrepancy between two density functions q and r. Let $\mathcal{X} \subseteq \mathbb{R}^D$ be a connected open set. Assume that Q, R have probability density functions denoted by q, r respectively. Given a positive definite kernel k_Y, the *Stein witness function* [Chwialkowski et al., 2016, Liu et al., 2016] $g^{q,r} : \mathcal{X} \to \mathbb{R}^D$ between q and r is defined as $g^{q,r}(w) := \mathbb{E}_{z\sim R}[\xi^q(z, w)] = (g_1^{q,r}(w), \ldots, g_D^{q,r}(w))^\top$, where $\xi^q(z, w) := k_Y(z, w)\nabla_z \log q(z) + \nabla_z k_Y(z, w)$. Under appropriate conditions (see Chwialkowski et al. 2016, Theorem 2.2, Liu et al. 2016, Proposition 3.3, and Barp et al. 2019, Proposition 1), it can be shown that $g^{q,r} = \mathbf{0}$ (i.e., the zero function) if and only if $q = r$. An implication of this result is that the deviation of $g^{q,r}$ from the zero function can be used as a measure of mismatch between q and r. Different ways to characterize such deviation have led to different measures of goodness of fit.

The FSSD characterizes such deviation from $\mathbf{0}$ by evaluating $g^{q,r}$ at J_q test locations. Formally, given a set of test locations $W = \{w_j\}_{j=1}^{J_q}$, the squared FSSD is defined as follows [Jitkrittum et al., 2017b]:

$$\mathrm{FSSD}_q^2(r) := \frac{1}{dJ_q}\sum_{j=1}^{J_q}\|g^{q,r}(w_j)\|_2^2 := F_q^2$$

Under appropriate conditions, it is known that almost surely $F_q^2 = 0$ if and only if $q = r$. Using the notations as in Jitkrittum et al. [2017b], one can write $F_q^2 = \mathbb{E}_{(z,z')\sim R\otimes R}[\Delta_q(z, z')]$ where $\Delta_q(z, z') := \boldsymbol{\tau}_q^\top(z)\boldsymbol{\tau}_q(z')$, $\boldsymbol{\tau}_q(z) := \mathrm{vec}(\boldsymbol{\Xi}^q(z)) \in \mathbb{R}^{DJ_q}$ with $\mathrm{vec}(\mathbf{M})$ denoting a column vector of concatenated columns of $\mathbf{M}$, and $\boldsymbol{\Xi}^q(\boldsymbol{z}) \in \mathbb{R}^{D\times J_q}$ is defined such that $[\boldsymbol{\Xi}^q(z)]_{d,j} := \xi_d^q(z, w_j)/\sqrt{DJ_q}$ for $d = 1, \ldots, D$ and $j = 1, \ldots, J_q$. Equivalently, $F_q^2 = \|\boldsymbol{\mu}_q\|_2^2$ where $\boldsymbol{\mu}_q := \mathbb{E}_{z\sim R}[\boldsymbol{\tau}_q(z)]$. Similar to the UME statistic described previously, given a sample $Z_n = \{z_i\}_{i=1}^n \sim R$, an unbiased estimator of F_q^2, denoted by $\widehat{F_q^2}$ can be straightforwardly written as a second-order U-statistic, which can be computed in $\mathcal{O}(J_q n)$ time. It was shown in Jitkrittum et al. [2017b] that the test locations W can be chosen by maximizing the test power of the

[1] In this chapter, since the distance is always measured relative to the data generating distribution R, we write U_Q instead of $U(Q, R)$ to avoid cluttering the notation.

goodness-of-fit test proposing $H_0 : Q = R$ against $H_1 : Q \neq R$, using $\widehat{F_q^2}$ as the statistic. We note that, unlike UME, $\widehat{F_q^2}$ requires access to the density q. Another way to characterize the deviation of $g^{q,r}$ from the zero function is to use the norm in the reproducing kernel Hilbert space (RKHS) that contains $g^{q,r}$. This measure corresponds to the kernel Stein discrepancy having a runtime complexity of $\mathcal{O}(n^2)$ [Chwialkowski et al., 2016, Liu et al., 2016].

5.3 Proposal: Rel-UME and Rel-FSSD Tests

Relative UME (Rel-UME**)** Our first proposed relative goodness-of-fit test is based on UME and tests

$$H_0 \colon U^2(P,R) \leq U^2(Q,R) \text{ versus } H_1 \colon U^2(P,R) > U^2(Q,R).$$

The test uses $\sqrt{n}\hat{S}_n^U = \sqrt{n}(\widehat{U_P^2} - \widehat{U_Q^2})$ as the statistic, and rejects H_0 when it is larger than the threshold T_α. Here, we assume that the model sample sizes are equal to the data sample size. The threshold is given by the $(1-\alpha)$-quantile of the asymptotic distribution of $\sqrt{n}\hat{S}_n^U$ when H_0 holds i.e., the null distribution, and the pre-chosen α is the significance level. It is well-known that this choice for the threshold asymptotically controls the false rejection rate to be bounded above by α yielding a level-α test [Lehmann and Romano, 2005, Definition 11.1.1]. In the full generality of Rel-UME, two sets of test locations can be used: $V = \{v_j\}_{j=1}^{J_p}$ for computing $\widehat{U_P^2}$, and $W = \{w_j\}_{j=1}^{J_q}$ for $\widehat{U_Q^2}$. The feature function for $\widehat{U_P^2}$ is denoted by $\psi_V(x) := J_p^{-1/2}\left(k_X(x,v_1),\ldots,k_X(x,v_{J_p})\right)^\top \in \mathbb{R}^{J_p}$, for some kernel k_X which can be different from k_Y used in ψ_W. The asymptotic distribution of the statistic is stated in Theorem 5.1.

Theorem 5.1 (Asymptotic distribution of $\hat{S}_n^U$)**.** *Define* $C_W^Q := \mathrm{cov}_{y\sim Q}[\psi_W(y),\psi_W(y)]$, $C_V^P := \mathrm{cov}_{x\sim P}[\psi_V(x),\psi_V(x)]$, *and* $C_{VW}^R := \mathrm{cov}_{z\sim R}[\psi_V(z),\psi_W(z)] \in \mathbb{R}^{J_p\times J_q}$. *Let* $S^U := U_P^2 - U_Q^2$, *and* $\mathbf{M} := \begin{pmatrix} \psi_V^P - \psi_V^R & \mathbf{0} \\ \mathbf{0} & \psi_W^Q - \psi_W^R \end{pmatrix} \in \mathbb{R}^{(J_p+J_q)\times 2}$. *Assume that* $\underline{1)}$ P, Q *and* R *are all distinct,* $\underline{2)}$ (k_X, V) *are chosen such that* $U_P^2 > 0$, *and* (k_Y, W) *are chosen such that* $U_Q^2 > 0$, $\underline{3)}$ $\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} := \mathbf{M}^\top \begin{pmatrix} C_V^P + C_V^R & C_{VW}^R \\ (C_{VW}^R)^\top & C_W^Q + C_W^R \end{pmatrix} \mathbf{M}$ *is positive definite. Then, as* $n \to \infty$,

$$\sqrt{n}\left(\widehat{S}_n^U - S^U\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right).$$

A proof of Theorem 5.1 can be found in Section 5.C (appendix). Let $\nu := 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)$. Theorem 5.1 states that the asymptotic distribution of $\hat{S}_n^U$ is normal with the mean given by $S^U := U_P^2 - U_Q^2$. It follows that under H_0, $S^U \leq 0$ and the $(1-\alpha)$-quantile is $S^U + \sqrt{\nu}\tau_{1-\alpha}$ where $\tau_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. Since S^U is unknown in practice, we therefore adjust it to be $\sqrt{\nu}\Phi^{-1}(1-\alpha)$, and use it as the test threshold T_α. The adjusted threshold can be estimated easily by replacing ν with $\hat{\nu}_n$, a consistent estimate based on samples. It can be shown that the test with the adjusted threshold is still level-α (more conservative in rejecting H_0). We note that the same approach of adjusting the threshold is used in Rel-MMD [Bounliphone et al., 2016] and in the KSD test in Chapter 3.

Better Fit of Q in Terms of W. When specifying V and W, the model comparison is performed by comparing the goodness of fit of P (to R) as measured in the regions specified by V to the goodness of fit of Q as measured in the regions specified by W. By specifying V and setting $W = V$, testing with Rel-UME is equivalent to posing the question "*Does Q fit to the data better than P does, as measured in the regions of V?*" The regions might represent predictive points, or objects intended to generate. For instance, the observed sample from R might contain smiling and non-smiling faces, and P, Q are candidate generative models for face images. If we are interested in checking the relative fit in the regions of smiling faces, V can be a set of smiling faces. In the following, we will assume $V = W$ and $k := k_X = k_Y$ for interpretability. Investigating the general case without these constraints will be an interesting topic of future study. Importantly we emphasize that test results are always conditioned on the specified V. To be precise, let U_V^2 be the squared UME statistic defined by V. It is entirely realistic that the test rejects H_0 in favor of $H_1\colon U_{V_1}^2(P,R) > U_{V_1}^2(Q,R)$ (i.e., Q fits better) for some V_1, and also rejects H_0 in favor of the opposite alternative $H_1\colon U_{V_2}^2(Q,R) > U_{V_2}^2(P,R)$ (i.e., P fits better) for another setting of V_2. This is because the regions in which the model comparison takes place are different in the two cases. Although not discussed in Bounliphone et al. [2016], the same behavior can be observed for Rel-MMD i.e., test results are conditioned on the choice of kernel.

In some cases, it is not known in advance what features are better represented by one model versus another, and it becomes necessary to learn these features from the model outputs. In this case, we propose setting V to contain the locations which maximize the probability that the test can detect the better fit of Q, as measured at the locations. Following the same principle as in Gretton et al. [2012b], Sutherland et al. [2016], Jitkrittum et al. [2016, 2017a,b], this goal can be achieved by finding (k, V) which maximize the test power, while ensuring that the test is level-α. By Theorem 5.1, for large n the test power $\Pr\big(\sqrt{n}\hat{S}_n^U > T_\alpha\big)$ is approximately

$$\Phi\left(\frac{\sqrt{n}S^U - T_\alpha}{\sqrt{\nu}}\right) = \Phi\left(\sqrt{n}\frac{S^U}{\sqrt{\nu}} - \sqrt{\frac{\hat{\nu}_n}{\nu}}\Phi^{-1}(1-\alpha)\right),$$

where Φ is the cumulative distribution function of the standard normal distribution. Under H_1, we have $S^U > 0$. For large n, $\Phi^{-1}(1-\alpha)\sqrt{\hat{\nu}_n}/\sqrt{\nu}$ approaches a constant, and $\sqrt{n}S^U/\sqrt{\nu}$ dominates. It follows that, for large n,

$$(k^*, V^*) = \arg\max_{(k,V)} \Pr\left(\sqrt{n}\hat{S}_n^U > T_\alpha\right) \approx \arg\max_{(k,V)} S^U/\sqrt{\nu}.$$

We can thus use $\hat{S}_n^U/(\gamma + \sqrt{\hat{\nu}_n})$ as an estimate of the *power criterion* objective $S^U/\sqrt{\nu}$ for the test power, where $\gamma > 0$ is a small regularization parameter added to promote numerical stability following Jitkrittum et al. [2017b, p. 5]. To control the false rejection rate, the maximization is carried out on held-out training data which are independent of the data used for testing. In the experiments (Section 5.4), we hold out 20% of the data for the optimization. A unique consequence of this procedure is that we obtain optimized V^* which indicates where Q fits significantly better than P. We note that this interpretation only holds if the test, using the optimized hyperparameters (k^*, V^*), decides to reject H_0. The optimized locations may not be interpretable if the test fails to reject H_0.

Relative FSSD (Rel-FSSD). The proposed Rel-FSSD tests

$$H_0\colon F_p^2 \le F_q^2 \text{ versus } H_1\colon F_p^2 > F_q^2.$$

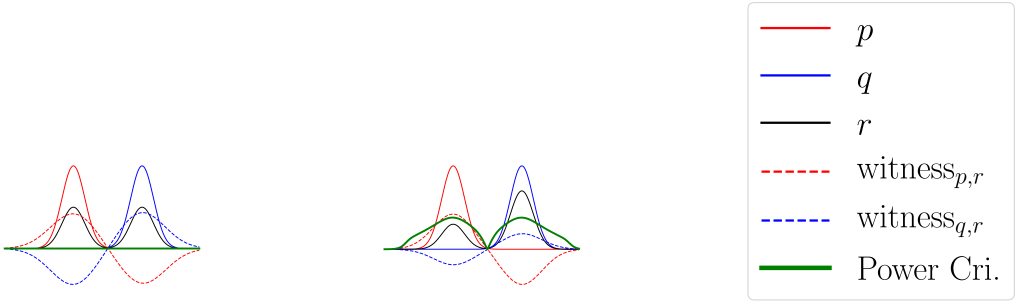
The test statistic is $\sqrt{n}\hat{S}_n^F := \sqrt{n}(\widehat{F_p^2} - \widehat{F_q^2})$. We note that the feature functions $\boldsymbol{\tau}_p$ (for F_p^2) and $\boldsymbol{\tau}_q$ (for F_q^2) depend on (k_X, V) and (k_Y, W) respectively, and play the same role as the feature functions ψ_V and ψ_W of the UME statistic. We only state the salient facts of Rel-FSSD, as the rest of the derivations closely follow Rel-UME. These include the interpretation that the relative fit is measured at the specified locations given in V and W, and the derivation of Rel-FSSD's power criterion (which can be derived using the asymptotic distribution of $\hat{S}_n^F$ given in Theorem 5.2, following the same line of reasoning as in the case of Rel-UME). A major difference is that Rel-FSSD requires explicit (gradients of the log) density functions of the two models, allowing it to gain structural information of the models that may not be as easily observed in finite samples. We next state the asymptotic distribution of the statistic (Theorem 5.2), which is needed for obtaining the threshold and for deriving the power criterion. The proof closely follows the proof of Theorem 5.1, and is omitted.

Theorem 5.2 (Asymptotic distribution of $\hat{S}_n^F$)**.** *Let* $\boldsymbol{\Sigma}^{ss'} := \mathrm{cov}_{\boldsymbol{z}\sim r}[\boldsymbol{\tau}_s(\boldsymbol{z}), \boldsymbol{\tau}_{s'}(\boldsymbol{z})]$ *for* $s, s' \in \{p, q\}$ *so that* $\boldsymbol{\Sigma}^{pq} \in \mathbb{R}^{DJ_p \times DJ_q}$, $\boldsymbol{\Sigma}^{qp} := (\boldsymbol{\Sigma}^{pq})^\top$, $\boldsymbol{\Sigma}^{pp} = \boldsymbol{\Sigma}^p \in \mathbb{R}^{DJ_p \times DJ_p}$, *and* $\boldsymbol{\Sigma}^{qq} = \boldsymbol{\Sigma}^q \in \mathbb{R}^{DJ_q \times DJ_q}$. *Define* $S^F := F_p^2 - F_q^2$. *Assume that* <u>*1)*</u> *p, q, and r are all distinct,* <u>*2)*</u> *(k_X, V) are chosen such that $F_p^2 > 0$, and (k_Y, W) are chosen such that $F_q^2 > 0$,* <u>*3)*</u> $\begin{pmatrix} \sigma_p^2 & \sigma_{pq} \\ \sigma_{pq} & \sigma_q^2 \end{pmatrix} := \begin{pmatrix} \boldsymbol{\mu}_p^\top \boldsymbol{\Sigma}^p \boldsymbol{\mu}_p & \boldsymbol{\mu}_p^\top \boldsymbol{\Sigma}^{pq} \boldsymbol{\mu}_q \\ \boldsymbol{\mu}_p^\top \boldsymbol{\Sigma}^{pq} \boldsymbol{\mu}_q & \boldsymbol{\mu}_q^\top \boldsymbol{\Sigma}^q \boldsymbol{\mu}_q \end{pmatrix}$ *is positive definite. Then, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{S}_n^F - S^F\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 4(\sigma_p^2 - 2\sigma_{pq} + \sigma_q^2)\right).$$

5.4 Experiments

In this section, we demonstrate the two proposed tests on both toy and real problems. We start with an illustration of the behaviors of Rel-UME and Rel-FSSD's power criteria using simple one-dimensional problems. In the second experiment, we examine the test powers of the two proposed tests using three toy problems. In the third experiment, we compare two hypothetical generative models on the CIFAR-10 dataset [Krizhevsky and Hinton, 2009] and demonstrate that the learned test locations (images) can clearly indicate the types of images that are better modeled by one of the two candidate models. In the last two experiments, we consider the problem of determining the relative goodness of fit of two given Generative Adversarial Networks (GANs) [Goodfellow et al., 2014]. Code to reproduce all the results is available at `https://github.com/wittawatj/kernel-mod`.



(a) Even mixture proportion

(b) Uneven mixing proportion

Figure 1: One-dimensional plots (in green) of Rel-UME's power criteria. The dashed lines indicate MMD's witness functions.

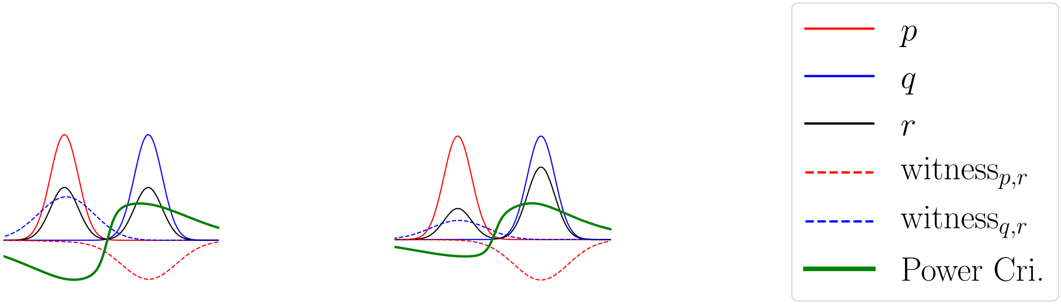
5.4.1 Illustration of Rel-UME and Rel-FSSD Power Criteria

We consider $k = k_X = k_Y$ to be an exponentiated quadratic, and set $V = W = \{v\}$ (one test location). The power criterion of Rel-UME as a function of $\boldsymbol{v}$ can be written as

$$\frac{1}{2} \frac{\mathrm{wit}^2_{P,R}(v) - \mathrm{wit}^2_{Q,R}(v)}{(\zeta^2_P(v) - 2\zeta_{PQ}(v) + \zeta^2_Q(v))^{1/2}},$$

where $\mathrm{wit}(\cdot)$ is the MMD witness function (see Section 5.2), and we explicitly indicate the dependency on v. To illustrate, we consider two Gaussian models p, q with different means but the same variance, and set r to be a mixture of p and q. Figure 1a shows that when each component in r has the same mixing proportion, the power criterion of Rel-UME is a zero function indicating that p and q have the same goodness of fit to r everywhere. To understand this, notice that at the left mode of r, p has excessive probability mass (compared to r), while q has almost no mass at all. Both models are thus wrong at the left mode of r. However, since the extra probability mass of p is equal to the missing mass of q, Rel-UME considers p and q as having the same goodness of fit. In Figure 1b, the left mode of r now has a mixing proportion of only 30%, and r more closely matches q. The power criterion is thus positive at the left mode indicating that q has a better fit.

The power criterion of Rel-FSSD indicates that q fits better at the right mode of r in the case of equal mixing proportion (see Figure 2a). In one dimension, the Stein witness function $g^{q,r}$ (defined in Section 5.2) can be written as $g^{q,r}(w) = \mathbb{E}_{z\sim r}\left[k_Y(z,w)\nabla_z\left(\log q(z) - \log r(z)\right)\right]$, which is the expectation under r of the difference in the derivative log of q and r, weighted by the kernel k_Y. The Stein witness thus only captures the matching of the shapes of the two densities (as given by the derivative log). Unlike the MMD witness, the Stein witness is insensitive to the mismatch of probability masses i.e., it is independent of the normalizer of q (this property will be revisited in Chapter 6). In Figure 2a, since the shape of q and the shape of the right mode of r match, the Stein witness $g^{q,r}$ (dashed blue curve) vanishes at the right mode of r, indicating a good fit of q in the region. The mismatch between the shape of q and the shape of r at the left mode of r is what creates the peak of $g^{q,r}$. The same reasoning holds for the Stein witness $g^{p,r}$.



(a) Even mixing proportion

(b) Uneven mixing proportion

Figure 2: One-dimensional plots (in green) of Rel-FSSD's power criteria. The dashed lines indicate FSSD's Stein witness functions.

The power criterion of Rel-FSSD, which is given by

$$\frac{1}{2}\frac{g^{p,r}(w)^2 - g^{q,r}(w)^2}{(\sigma_p^2(w) - 2\sigma_{pq}(w) + \sigma_q^2(w))^{1/2}},$$

is thus positive at the right mode of r (shapes of q and r matched there), and negative at the left mode of r (shapes of p and r matched there). To summarize, Rel-UME measures the relative fit by checking the probability mass, while Rel-FSSD does so by matching the shapes of the densities.

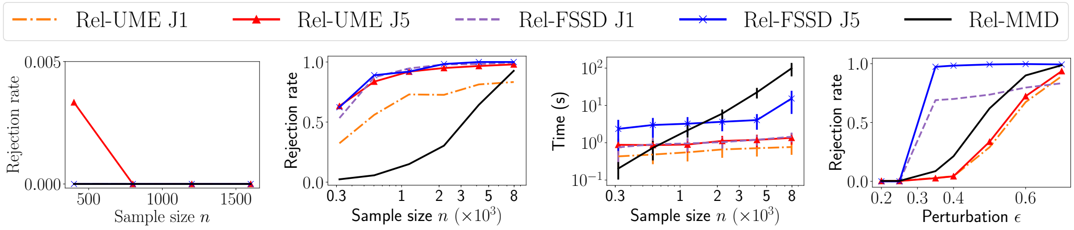
5.4.2 Test Powers on Toy Problems

The goal of this experiment is to investigate the rejection rates of several variations of the two proposed tests. To this end, we study three toy problems, each having its own characteristics. All the three distributions in each problem have density functions to allow comparison with Rel-FSSD.

1. *Mean shift*: All the three distributions are isotropic multivariate normal distributions: $p = \mathcal{N}([0.5, 0, \ldots, 0], I), q = \mathcal{N}([1, 0, \ldots 0], I)$, and $r = \mathcal{N}(\mathbf{0}, I)$, defined on $\mathbb{R}^{50}$. The two candidates models p and q differ in the mean of the first dimension. In this problem, the null hypothesis H_0 is true since p is closer to r.

2. *Blobs*: Each distribution is given by a mixture of four Gaussian distributions organized in a grid in $\mathbb{R}^2$. Samples from p, q and r are shown in Figure 4. In this problem, q is closer to r than p is i.e., H_1 is true. One characteristic of this problem is that the difference between p and q takes place in a small scale relative to the global structure of the data. This problem was studied in [Gretton et al., 2012b, Chwialkowski et al., 2015].

3. *RBM*: Each of the three distributions is given by a Gaussian Bernoulli Restricted Boltzmann Machine (RBM) model with density function $p'_{B,b,c}(x) = \sum_h p'_{B,b,c}(x, h)$, where $p'_{B,b,c}(x, h) = \exp\left(x^\top Bh + b^\top x + c^\top h - \frac{1}{2}\|x\|^2\right)/Z$, $h \in \{-1, 1\}^{D_h}$ is a latent vector, Z is the normalizer, and B, b, c are model parameters. Let $r(x) := p'_{B,b,c}(x), p(x) := p'_{B^p,b,c}(x)$, and $q(x) := p'_{B^q,b,c}(x)$. Following a similar setting as in Liu et al. [2016],



(a) Mean shift. $D = 50$. (b) Blobs. $D = 2$. (c) Blobs (Runtime) (d) RBM. $D = 20$

Figure 3: (a), (b), (d) Rejection rates (estimated from 300 trials) of the five tests with $\alpha = 0.05$. In the RBM problem, $n = 2000$. (c) Runtime in seconds for one trial in the Blobs problem.

Jitkrittum et al. [2017b], we set the parameters of the data generating density r by uniformly randomly setting entries of B to be from $\{-1, 1\}$, and drawing entries of b and c from the standard normal distribution. Let $E_{1,1}$ be a matrix of the same size as B such that the $(1, 1)$-entry is one, and all other entries are 0. We set $B^q = B + 0.3E_{1,1}$ and $B^p = B + \epsilon E_{1,1}$, where the perturbation constant ϵ is varied. We fix the sample size n to 2,000. Perturbing only one entry of B creates a problem in which the difference of distributions can be difficult to detect. This serves as a challenging benchmark to measure the sensitivity of statistical tests [Jitkrittum et al., 2017b]. We set $D = 20$ and $D_h = 5$.

We compare three kernel-based tests: Rel-UME, Rel-FSSD, and Rel-MMD (the relative MMD test of Bounliphone et al. 2016), all using an exponentiated quadratic. For Rel-UME and Rel-FSSD we set $k_X = k_Y = k$, where the Gaussian width of k, and the test locations are chosen by maximizing their respective power criteria described in Section 5.3 on 20% of the data. The optimization procedure is described in Section 5.A (appendix). Following Bounliphone et al. [2016], the Gaussian width of Rel-MMD is chosen by the median heuristic as implemented in the code by the authors. In the RBM problem, all problem parameters $\boldsymbol{B}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are drawn only once and fixed. Only the samples vary across trials.

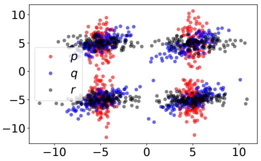
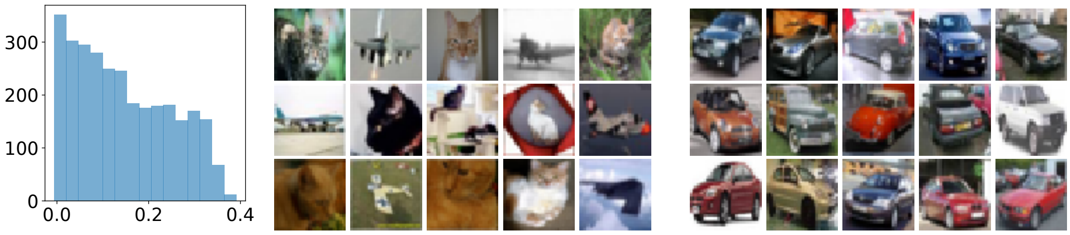


Figure 4: Blobs problem samples: p, q, r.

Figure 3 shows the test powers of all the tests. When H_0 holds, all tests have false rejection rates (type-I errors) bounded above by $\alpha = 0.05$ (Figure 3a). In the Blobs problem (Figure 3b), it can be seen that Rel-UME achieves larger power at all sample sizes, compared to Rel-MMD. Since the relative goodness of fit of p and q must be compared locally, the optimized test locations of Rel-UME are suitable for detecting such local differences. The poor performance of Rel-MMD is caused by unsuitable choices of the kernel bandwidth. The bandwidth chosen by the median heuristic is only appropriate for capturing the global length scale of the problem. It is thus too large to capture small-scale differences. No existing work has proposed a kernel selection procedure for Rel-MMD. Regarding the number J of test locations, we observe that changing J from 1 to 5 drastically increases the test power of Rel-UME, since more regions characterizing the differences can be pinpointed. Rel-MMD exhibits a quadratic-time profile (Figure 3c) as a function of n.

Figure 3d shows the rejection rates against the perturbation strength ϵ in p in the RBM problem. When $\epsilon \leq 0.3$, p is closer to r than q is (i.e., H_0 holds). We observe that all the tests have well-controlled false rejection rates in this case. At $\epsilon = 0.35$, while q is closer (i.e., H_1



(a) Power Criterion

(b) Sorted in ascending order

(c) Sorted in descending order

Figure 5: P = {airplane, cat}, Q = {automobile, cat}, and R = {automobile, cat}. (a) Histogram of Rel-UME power criterion values. (b), (c) Images as sorted by the criterion values in ascending and descending orders, respectively.

holds), the relative amount by which q is closer to r is so small that a significant difference cannot be detected when p and q are represented by samples of size $n = 2,000$, hence the low powers of Rel-UME and Rel-MMD. Structural information provided by the density functions allows Rel-FSSD (both $J = 1$ and $J = 5$) to detect the difference even at $\epsilon = 0.35$, as can be seen from the high test powers. The fact that Rel-MMD has higher power than Rel-UME, and the fact that changing J from 1 to 5 increases the power only slightly suggest that the differences may be spatially diffuse (rather than local).

5.4.3 Informative Power Objective

In this part, we demonstrate that test locations having positive (negative) values of the power criterion correctly indicate the regions in which Q has a better (worse) fit. We consider image samples from three categories of the CIFAR-10 dataset [Krizhevsky and Hinton, 2009]: airplane, automobile, and cat. We partition the images, and assume that the sample from P consists of 2,000 airplane, 1,500 cat images, the sample from Q consists of 2,000 automobile, 1,500 cat images, and the reference sample from R consists of 2,000 automobile, 1,500 cat images. All samples are independent. We consider a held-out random sample consisting of 1,000 images from each category, serving as a pool of test location candidates. We set the kernel to be the exponentiated quadratic on 2,048 features extracted by the Inception-v3 network at the pool3 layer [Szegedy et al., 2016]. We evaluate the power criterion of Rel-UME at each of the test locations in the pool individually. The histogram of the criterion values is shown in Figure 5a. We observe that all the power criterion values are non-negative, confirming that Q is better than P everywhere. Figure 5b shows the top 15 test locations as sorted in ascending order by the criterion, consisting of automobile images. These indicate the regions in the data domain where Q fits better. Notice that cat images do not have high positive criterion values because they can be modeled equally well by P and Q, and thus have scores close to zero as shown in Figure 5b.

5.4.4 Testing GAN Models

In this experiment, we apply the proposed Rel-UME test to comparing two generative adversarial networks (GANs) [Goodfellow et al., 2014]. We consider the CelebA dataset [Liu et al., 2015][2] in which each data point is an image of a celebrity with 40 binary attributes annotated

[2] CelebA dataset: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

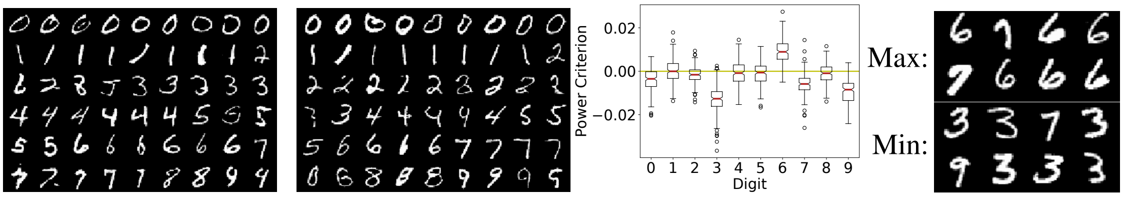
Table 1: Rejection rates of the proposed $\mathrm{Rel\text{-}UME}$, $\mathrm{Rel\text{-}MMD}$, KID and FID, in the GAN model comparison problem. "FID diff." refers to the average of $\mathrm{FID}(P, R) - \mathrm{FID}(Q, R)$ estimated in each trial. Significance level $\alpha = 0.01$ (for $\mathrm{Rel\text{-}UME}$, $\mathrm{Rel\text{-}MMD}$, and KID). S and N stand for Smile and Non-smile. The prefix R indicates that real data points are used.

	P	Q	R	Rel-UME			Rel-MMD	KID	FID	FID diff.
				J10	J20	J40				
1.	S	S	RS	0.0	0.0	0.0	0.0	0.0	0.53	-0.045 ± 0.52
2.	RS	RS	RS	0.0	0.0	0.0	0.03	0.02	0.7	0.04 ± 0.19
3.	S	N	RS	0.0	0.0	0.0	0.0	0.0	0.0	-15.22 ± 0.83
4.	S	N	RN	0.57	0.97	1.0	1.0	1.0	1.0	5.25 ± 0.75
5.	S	N	RM	0.0	0.0	0.0	0.0	0.0	0.0	-4.55 ± 0.82

e.g., pointy nose, smiling, mustache, etc. We create a partition of the images on the *smiling* attribute, thereby creating two disjoint subsets of *smiling* and *non-smiling* images. A set of 30,000 images from each subset is held out for subsequent relative goodness-of-fit testing, and the rest are used for training two GAN models: a model for smiling images, and a model for non-smiling images. Generated samples and details of the trained models can be found in Section 5.B (appendix). The two models are trained once and fixed throughout.

In addition to $\mathrm{Rel\text{-}MMD}$, we compare the proposed $\mathrm{Rel\text{-}UME}$ to Kernel Inception Distance (KID) [Bińkowski et al., 2018], and Fréchet Inception Distance (FID) [Heusel et al., 2017], which are distances between two samples (originally proposed for comparing a sample of generated images, and a reference sample). All images are represented by 2,048 features extracted from the Inception-v3 network [Szegedy et al., 2016] at the pool3 layer following Bińkowski et al. [2018]. When adapted for three samples, KID is in fact a variant of $\mathrm{Rel\text{-}MMD}$ in which a third-order polynomial kernel is used instead of an exponentiated quadratic (on top of the pool3 features). Following Bińkowski et al. [2018], we construct a bootstrap estimator for FID (10 subsamples with 1,000 points in each). For the proposed $\mathrm{Rel\text{-}UME}$, the $J \in \{10, 20, 40\}$ test locations are randomly set to contain $J/2$ smiling images, and $J/2$ non-smiling images drawn from a held-out set of real images. We create problem variations by setting $P, Q, R \in \{$S, N, RS, RN, RM$\}$ where S denotes generated smiling images (from the trained model), N denotes generated non-smiling images, M denotes an equal mixture of smiling and non-smiling images, and the prefix R indicates that real images are used (as opposed to generated ones). The sample size is $n = 2,000$, and each problem variation is repeated for 10 trials for FID (due to its high complexity) and 100 trials for other methods. The rejection rates from all the methods are shown in Table 1. Here, the test result for FID in each trial is considered "reject H_0" if $\mathrm{FID}(P, R) > \mathrm{FID}(Q, R)$. Heusel et al. [2017] did not propose FID as a statistical test. That said, there is a generic way of constructing a relative goodness-of-fit test based on repeated permutation of samples of P and Q to simulate from the null distribution. However, FID requires computing the square root of the feature covariance matrix (of size 2048×2048), and is computationally too expensive for permutation testing.

Overall, we observe that the proposed test does at least equally well as existing approaches, in identifying the better model in each case. In Problems 1 and 2, P and Q have the same goodness of fit, by design. In these cases, all the tests correctly yield low rejection rates, staying



(a) Sample from P = LSGAN trained for 15 epochs.

(b) Sample from Q = LSGAN trained for 17 epochs.

(c) Power criterion

(d) Greedily optimized test locations

Figure 6: Examining the training of an LSGAN model with Rel-UME. (a), (b) Samples from the two models P, Q trained on MNIST. (c) Distributions of power criterion values computed over 200 trials. Each distribution is formed by randomly selecting $J = 40$ test locations from real images of a digit type. (d) Test locations showing where Q is better (maximization of the power criterion), and test locations showing where P is better (minimization).

roughly at the design level ($\alpha = 0.01$). Without a properly chosen threshold, the (false) rejection rates of FID fluctuate around the expected value of 0.5. This means that simply comparing FIDs (or other distances) to the reference sample without a calibrated threshold can lead to a wrong conclusion on the relative goodness of fit. The FID is further complicated by the fact that its estimator suffers from bias in ways that are hard to model and correct for (see Bińkowski et al. [2018, Section D.1]). Problem 4 is a case where the model Q is better. We notice that increasing the number of test locations of Rel-UME helps detect the better fit of Q. In problem 5, the reference sample is bimodal, and each model can capture only one of the two modes (analogous to the synthetic problem in Figure 1a). All the tests correctly indicate that no model is better than another.

5.4.5 Examining GAN Training

In the final experiment, we show that the power criterion of Rel-UME can be used to examine the relative change of the distribution of a GAN model after training further for a few epochs. To illustrate, we consider training an LSGAN model [Mao et al., 2017] on MNIST, a dataset in which each data point is an image of a handwritten digit. We set P and Q to be LSGAN models after 15 epochs and 17 epochs of training, respectively. Details regarding the network architecture, training, and the kernel (chosen to be an exponentiated quadratic on features extracted from a convolutional network) can be found in Section 5.D. Samples from P and Q are shown in Figures 6a and 6b (see Figure 5.D.2 in the appendix for more samples).

We set the test locations V to be the set V_i containing $J = 40$ randomly selected real images of digit i, for $i \in \{0, \ldots, 9\}$. We then draw $n = 2,000$ points from P, Q and the real data (R), and use $V = V_i$ to compute the power criterion for $i \in \{0, \ldots, 9\}$. The procedure is repeated for 200 trials where V and the samples are redrawn each time. The results are shown in Figure 6c. We observe that when $V = V_3$ (i.e., box plot at the digit 3) or V_9, the power criterion values are mostly negative, indicating that P is better than Q, as measured in the regions indicated by real images of the digits 3 or 9. By contrast, when $V = V_6$, the large mass of the box plot in the positive orthant shows that Q is better in the regions of the digit 6. For other digits, the criterion values spread around zero, showing that there is no difference between P and Q, on average. We further confirm that the class proportions of the generated digits from both models are roughly correct (i.e., uniform distribution), meaning that the difference between P and Q in these cases is not due to the mismatch in class proportions (see Section 5.D). These observations imply that after the 15th epoch, training this particular LSGAN model two epochs further improves generation of the digit 6, and degrades generation of digits 3 and 9. A non-monotonic improvement during training is not uncommon since at the 15th epoch the training has not converged. More experimental results from comparing different GAN variants on MNIST can be found in Section 5.E in the appendix.

We note that the set V does not need to contain test locations of the same digit. In fact, the notion of class labels may not even exist in general. It is up to the user to define V to contain examples which capture the relevant concept of interest. For instance, to compare the ability of models to generate straight strokes, one might include digits 1 and 7 in the set V. An alternative to manual specification of V is to optimize the power criterion to find the locations that best distinguish the two models (as done in experiment 2). To illustrate, we consider greedily optimizing the power criterion by iteratively selecting a test location (from real images) which best improves the objective. Maximizing the objective yields locations that indicate the better fit of Q, whereas minimization gives locations which show the better fit of P (recall from Figure 1). The optimized locations are shown in Figure 6d. The results largely agree with our previous observations, and do not require manually specifying V. This optimization procedure is applicable to any models which can be sampled.

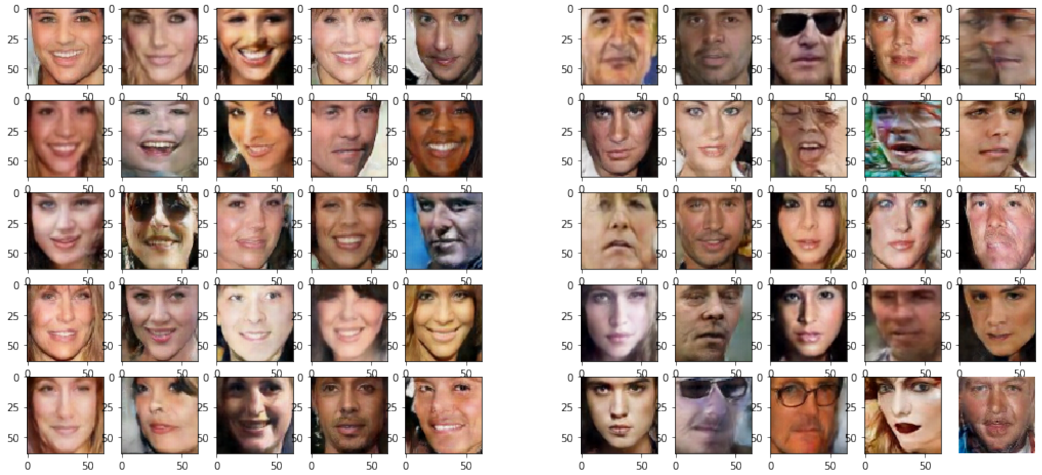
5.5 Conclusion

We have developed two interpretable test of relative goodness of fit. The proposed tests are based on two distributional discrepancies with interpretable features and come with the following two advantages. First, the user can formulate hypotheses based on features of their choice and interpret the test's result accordingly. Second, these tests allow the user to discover features distinguishing two models by optimizing their power proxies. Regarding the second benefit, our real-data experiment focused on discrete optimization, where we specified a finite set of test locations. Extending this procedure to continuous optimization is an interesting direction to explore, as considered in Jitkrittum et al. [2016].

5.A Optimization of Test Locations in Rel-UME and Rel-FSSD

This section describes the optimization procedure we use to select the test locations V and the bandwidth of the exponentiated quadratic kernel in the experiment “Test Powers on Toy Problems.” Since the two sets V, W of test locations are constrained to be the same i.e., $V = W$ consisting of $J = J_p = J_q$ locations, in total, we have $Jd + 1$ parameters. We follow a similar implementation of the optimization procedure for finding the test locations in FSSD.[3] All the parameters are optimized jointly by gradient ascent. We initialize the test locations by randomly picking J points from the training set. The Gaussian width is initialized (for gradient ascent) to the square of the mean of $\mathrm{med}_{X^{tr} \cup Z^{tr}}$ and $\mathrm{med}_{Y^{tr} \cup Z^{tr}}$, where $\mathrm{med}_A := \mathrm{median}\left(\{\|\boldsymbol{x} - \boldsymbol{x}'\|_2\}_{\boldsymbol{x}, \boldsymbol{x}' \in A}\right)$. This is a similar heuristic used in Bounliphone et al. [2016] to set the bandwidth of the exponentiated quadratic for Rel-MMD.

5.B Trained Models for Generating Smiling and Non-Smiling Images



(a) Samples from the smiling model

(b) Samples from the non-smiling model

Figure 5.B.1: Samples from the two trained models (smiling, and non-smiling) used in "Testing GAN Models" experiment in Section 5.4.

This section describes the details of the two GAN models (smiling, and non-smiling models) we use in the "Testing GAN Models" experiment in Section 5.4. We use the CelebA dataset [Liu et al., 2015] in which each data point is an image of a celebrity with 40 binary attributes annotated e.g., pointy nose, smiling, mustache, etc. We create a partition of the images on the *smiling* attribute, thereby creating two disjoint subsets of *smiling* and *non-smiling* images. To reduce confounding factors that are not related to smiling (e.g., sunglasses, background), each image is cropped to be 64x64 pixels, so that only the face remains. Cropping and image alignment with eyes and lips are done with the software described in Amos et al. [2016]. We

[3] Code for FSSD released by the authors: `https://github.com/wittawatj/kernel-gof`.

use DCGAN architecture [Radford et al., 2015] (for both generator and discriminator) for both smiling and non-smiling models, coded in PyTorch. Subsampling was performed so that the training sizes for the two models are equal. Each model is trained on 84,822 images (i.e., 84822 smiling faces, and 84822 non-smiling faces) for 50 epochs. The training time was roughly three hours using an Nvidia Titan X graphics card with Pascal architecture. We use Adam optimizer [Kingma and Ba, 2014] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to 10^{-3} (for both discriminator and generator in the two models). Some samples generated from the two trained models are shown in Figure 5.B.1.

5.C Proof of Theorem 5.1

Let all the notations be defined as in Section 5.3. Recall Theorem 5.1:

Theorem 5.1 (Asymptotic distribution of $\hat{S}^U_n$)**.** *Define* $C^Q_W := \mathrm{cov}_{y\sim Q}[\psi_W(y), \psi_W(y)]$, $C^P_V := \mathrm{cov}_{x\sim P}[\psi_V(x), \psi_V(x)]$, *and* $C^R_{VW} := \mathrm{cov}_{z\sim R}[\psi_V(z), \psi_W(z)] \in \mathbb{R}^{J_p\times J_q}$. *Let* $S^U := U_P^2 - U_Q^2$, *and* $\mathbf{M} := \begin{pmatrix} \psi^P_V - \psi^R_V & \mathbf{0} \\ \mathbf{0} & \psi^Q_W - \psi^R_W \end{pmatrix} \in \mathbb{R}^{(J_p+J_q)\times 2}$. *Assume that 1) P, Q and R are all distinct, 2) (k_X, V) are chosen such that $U_P^2 > 0$, and (k_Y, W) are chosen such that $U_Q^2 > 0$, 3)* $\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} := \mathbf{M}^\top \begin{pmatrix} C^P_V + C^R_V & C^R_{VW} \\ (C^R_{VW})^\top & C^Q_W + C^R_W \end{pmatrix} \mathbf{M}$ *is positive definite. Then, as* $n \to \infty$,

$$\sqrt{n}\left(\widehat{S}^U_n - S^U\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2)\right).$$

Proof. Consider a random vector $\mathbf{t} := (x, y, z) \in \mathcal{X}^3$, where x, y, and z are independently drawn from P, Q, and R, respectively. Let $\{\mathbf{t}_i\}_{i=1}^n = \{(x_i, y_i, z_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} T$ be i.i.d copies of $\mathbf{t}$. Define two functions

$$\delta^P_V(\mathbf{t}, \mathbf{t}') := (\psi_V(x) - \psi_V(z))^\top(\psi_V(x') - \psi_V(z')),$$
$$\delta^Q_W(\mathbf{t}, \mathbf{t}') := (\psi_W(y) - \psi_W(z))^\top(\psi_W(y') - \psi_W(z')),$$

where $\mathbf{t}' := (x', y', z')$. It can be seen that $\delta^P_V(\mathbf{t}, \mathbf{t}') = \delta^P_V(\mathbf{t}', \mathbf{t})$ and $\delta^Q_W(\mathbf{t}, \mathbf{t}') = \delta^Q_W(\mathbf{t}', \mathbf{t})$ for all $\mathbf{t}, \mathbf{t}' \in \mathcal{X}^3$, and that both functions are valid U-statistic kernels. It is not difficult to see that $\widehat{U_P^2}$ and $\widehat{U_Q^2}$ (estimator given in Section 5.2) can be written in the form of second-order U-statistics [Serfling, 2009, Chapter 5] as

$$\widehat{U_P^2} = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j<i} \delta^P_V(\mathbf{t}, \mathbf{t}') \text{ and } \widehat{U_Q^2} = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j<i} \delta^Q_W(\mathbf{t}, \mathbf{t}').$$

Since $\psi^P_V \neq \psi^R_V$ (because $U_P^2 > 0$), the U-statistic $\widehat{U_P^2}$ is a non-degenerate U-statistic. Since $\psi^Q_W \neq \psi^R_W$, $\widehat{U_Q^2}$ is also non-degenerate [Serfling, 2009, Section 5.5.1]. By Hoeffding [1948, Theorem 7.1], asymptotically their joint distribution is given by a normal distribution:

$$\sqrt{n}\left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix}\right), \tag{5.1}$$

where

$$\begin{aligned}
\zeta_P^2 &= \mathrm{Var}_{\mathbf{t}\sim T}\left[\mathbb{E}_{\mathbf{t}'\sim T}[\delta_V^P(\mathbf{t},\mathbf{t}')]\right] \overset{(a)}{=} (\psi_V^P - \psi_V^R)^\top (C_V^P + C_V^R)(\psi_V^P - \psi_V^R),\\
\zeta_Q^2 &= \mathrm{Var}_{\mathbf{t}\sim T}\left[\mathbb{E}_{\mathbf{t}'\sim T}[\delta_W^Q(\mathbf{t},\mathbf{t}')]\right] \overset{(b)}{=} (\psi_W^Q - \psi_W^R)^\top (C_W^Q + C_W^R)(\psi_W^Q - \psi_W^R),\\
\zeta_{PQ} &= \mathrm{cov}_{\mathbf{t}\sim T}\left(\mathbb{E}_{\mathbf{t}'\sim T}[\delta_V^P(\mathbf{t},\mathbf{t}')], \mathbb{E}_{\mathbf{t}'\sim T}[\delta_W^Q(\mathbf{t},\mathbf{t}')]\right) \overset{(c)}{=} (\psi_V^P - \psi_V^R)^\top C_{VW}^R (\psi_W^Q - \psi_W^R),
\end{aligned}$$

and $C_{VW}^R := \mathrm{cov}_{z\sim R}[\psi_V(z), \psi_W(z)] \in \mathbb{R}^{J_p \times J_q}$. At (a), (b), (c), we rely on the independence among x, y, and z. A direct calculation gives the expressions of ζ_P^2, ζ_Q^2, and ζ_{PQ}. By the continuous mapping theorem, and (5.1),

$$\begin{aligned}
\sqrt{n}\left(\widehat{S}_n^U - S^U\right) &= \sqrt{n}\begin{pmatrix} 1 \\ -1 \end{pmatrix}^\top \left(\begin{pmatrix} \widehat{U_P^2} \\ \widehat{U_Q^2} \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix}\right)\\
&\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, 4\begin{pmatrix} 1 \\ -1 \end{pmatrix}^\top \begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right).
\end{aligned}$$

□

Remark 5.3. The assumption that P, Q, and R are all distinct in Theorem 5.1 is necessary for $\widehat{U_P^2}$ and $\widehat{U_Q^2}$ to follow a non-degenerate normal distribution asymptotically. If $R \in \{P, Q\}$, then $\widehat{U_S^2}$ for $S \in \{P, Q\}$ asymptotically follows a weighted sum of chi-squared random variables, and $U_S^2 = 0$. If $P = Q$, the covariance matrix in (5.1) is rank-defficient.

5.D Details of Experiment 5: Examining GAN Training

LSGAN Architecture We rely on Pytorch code[4] by Hyeonwoo Kang to train the LSGAN [Mao et al., 2017] model that we use in experiment 5. Network architectures of the generator and the discriminator follow the design used in Chen et al. [2016, Section C.1]. We reproduce here in Table 5.D.1 for ease of reference.

Table 5.D.1: Discriminator and generator of LSGAN used in experiment 5.

Discriminator	Generator
Input: 28×28 grayscale image	Input noise vector $\boldsymbol{z} \sim \mathrm{Unif}[0,1]^{62}$
4×4 conv. 64 LRELU. Stride 2.	FC. 1024 RELU. Batch norm.
4×4 conv. 128 LRELU. Stride 2. Batch norm.	FC. $7 \times 7 \times 128$ RELU. Batch norm.
FC. 104 Leaky RELU. Batch norm.	4×4 upconv. 64 RELU. Stride 2. Batch norm.
FC	4×4 upconv. 1 channel.

conv. refers to a convolution layer, FC means a fully-connected layer, RELU means a rectified linear unit, LRELU means Leaky RELU, and upconv is the transposed convolution.

[4] https://github.com/znxlwm/pytorch-generative-model-collections (commit: 0d183bb5ea)

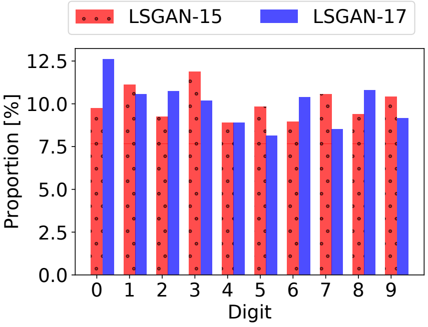


Figure 5.D.1: Proportions of generated digits from the LSGAN models at 15th and 17th epochs. Classification of each generated image is done by a trained convolutional neural network classifier (see Section 5.D).

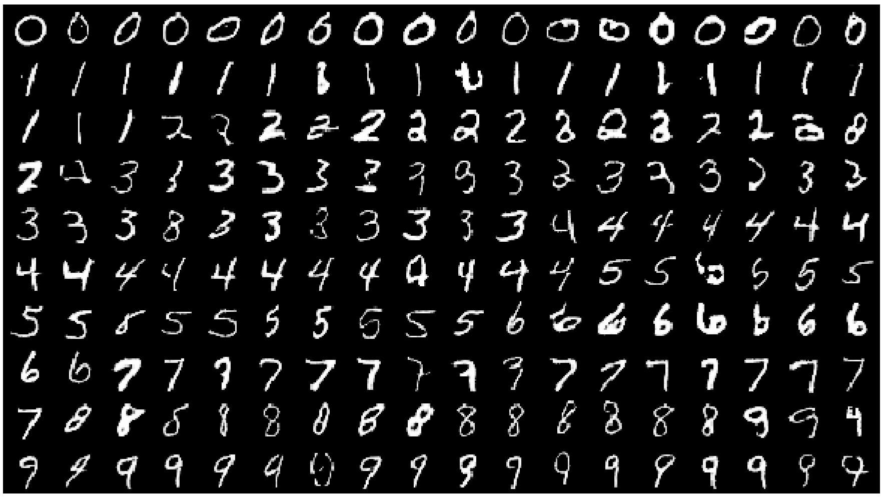
Kernel Function The kernel k is chosen to be an exponentiated quadratic kernel on features extracted from a convolutional neural network (CNN) classifier trained to classify the ten digits of MNIST. Specifically the kernel k is $k(x, y) = \exp\left(-\|f(x) - f(y)\|_2^2/2\nu^2\right)$, where f is the output (in $\mathbb{R}^{10}$) of the last fully-connected layer of a trained CNN classifier.[5] The architecture of the CNN is

$$\begin{aligned}\text{Input: } 28{\times}28 \text{ grayscale image} &\rightarrow 5 \times 5 \text{ conv. } 10 \text{ filters. } 2{\times}2 \text{ max pool}\\ &\rightarrow 5 \times 5 \text{ conv. } 20 \text{ filters. } 2{\times}2 \text{ max pool}\\ &\rightarrow \text{FC. } 50 \text{ RELU.}\\ &\rightarrow \text{FC. } 10 \text{ outputs.}\end{aligned}$$

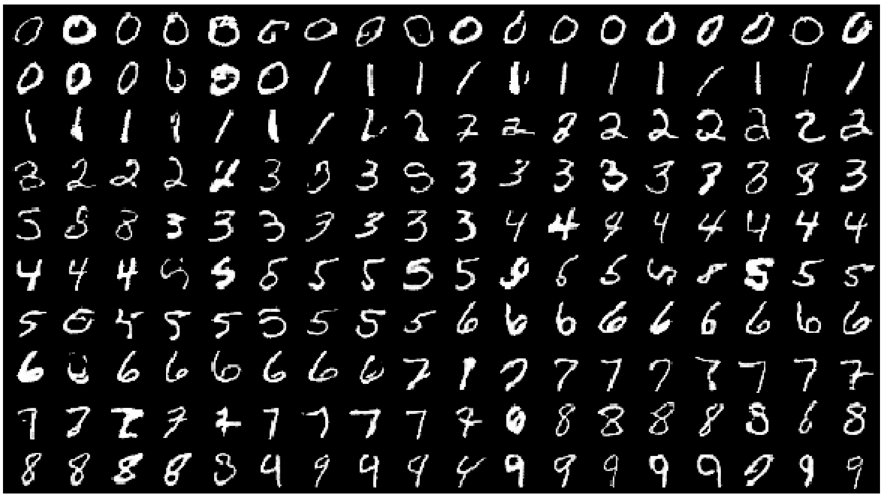
We train the CNN for 30 epochs and achieve higher than 99% accuracy on MNIST's test set. The Gaussian bandwidth ν is set with the median heuristic.

Class Proportion of Generated Digits To examine the proportion of digits in the generated samples, we sample 4000 images from both models P (LSGAN-15, LSGAN model trained for 15 epochs), and Q (LSGAN-17, LSGAN model trained for 17 epochs), and use the CNN classifier to assign a label to each image. The proportions of digits are shown in Figure 5.D.1. We observe that the generated digits from both LSGAN-15 and LSGAN-17 follow the right distribution i.e., uniform distribution, up to variability due to noise. There is no mode collapse problem. This observation means that the difference between P and Q studied in experiment 5 in the main text is not due to the mismatch of class proportions.

[5] Code to train the CNN classifier is taken from `https://github.com/pytorch/examples/blob/master/mnist/main.py` (commit: 75e7c75).



(a) Samples from LSGAN trained for 15 epochs.



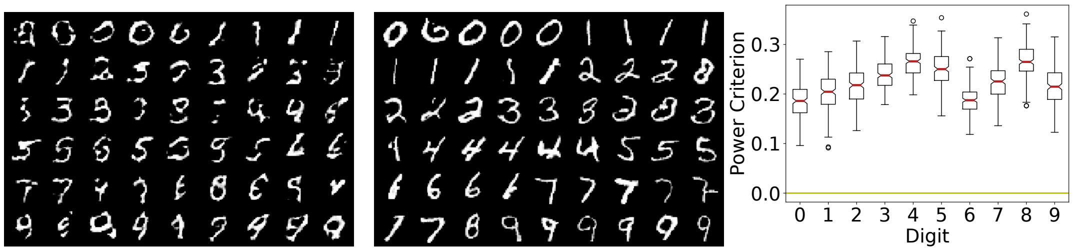
(b) Samples from LSGAN trained for 17 epochs.

Figure 5.D.2: Samples from LSGAN models trained on MNIST. Samples are taken from the models at two different time points: after 15 epochs, and after 17 epochs of training.

5.E Comparing Different GAN Models Trained on MNIST

This section extends experiment 5 in the main text to compare other GAN variants trained on MNIST. All the GAN variants that we consider have the same network architecture as described in Table 5.D.1. We use the notation *AAA-n* to refer to a GAN model of type AAA trained for n epochs. We note that the result presented here for each GAN variant does not represent its best achievable result.

WGAN-GP-10 vs. LSGAN-10 Here we compare P = Wasserstein GAN with Gradient Penalty [Gulrajani et al., 2017] and Q = LSGAN [Mao et al., 2017] trained for ten epochs on MNIST. The results are shown in Figure 5.E.1. From the generated samples from the two models, it appears that LSGAN yields more realistic images of handwritten digits, after training for ten epochs. The positive power criterion values in Figure 5.E.1c further confirm this observation i.e., Q is better at all digits.



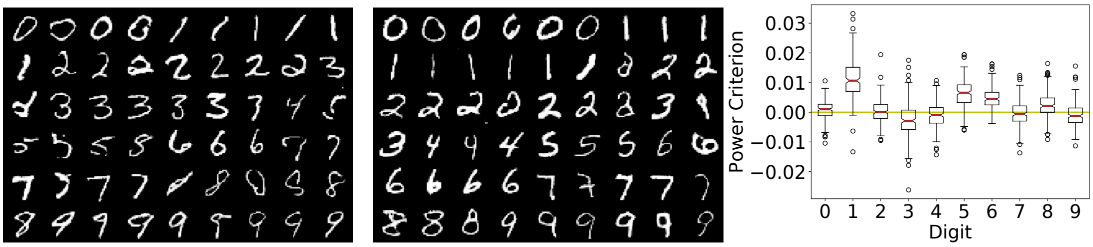
(a) Sample from P = WGAN-GP trained for 10 epochs.

(b) Sample from Q = LSGAN trained for 10 epochs.

(c) Power criterion

Figure 5.E.1: Comparing WGAN-GP (Wasserstein GAN with Gradient Penalty) and LSGAN, trained for ten epochs on MNIST.

GAN-40 vs. LSGAN-40 In this part, we compare P = GAN-40 [Goodfellow et al., 2014] and Q = LSGAN trained for 40 epochs on MNIST. The results are shown in Figure 5.E.2. It can be seen from visual inspection that LSGAN-40 is slightly better overall, except for digits 1 and 5 at which LSGAN-40 appears to be significantly better. This observation is also hinted by the power criterion values at digits 1 and 5 which tend to be positive (see Figure 5.E.2c).



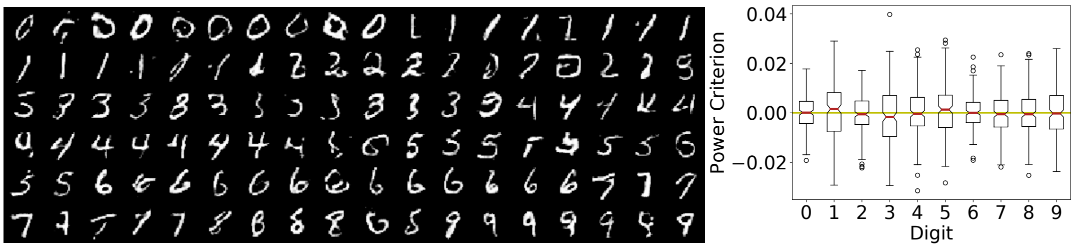
(a) Sample from P = GAN trained for 40 epochs.

(b) Sample from Q = LSGAN trained for 40 epochs.

(c) Power criterion

Figure 5.E.2: Comparing GAN (the original formulation) and LSGAN, trained for 40 epochs on MNIST.

WGAN-30 vs WGAN-30 As a sanity check, we also run the same procedure on a case where $P = Q$. We set $P = Q$ = Wasserstein GAN [WGAN, Arjovsky et al., 2017] trained for 30 epochs on MNIST. The results are shown in Figure 5.E.3. As expected, the power criterion values spread around zero in all cases. We note that we did not modify the procedure to treat this special case. In particular, in each trial, two samples are drawn from P and Q as usual.



(a) Sample from $P = Q =$ WGAN trained for 30 epochs.

(b) Power criterion

Figure 5.E.3: Comparing two models which are the same for sanity checking. The model is set to WGAN trained for 30 epochs.

Chapter 6

Controlling moments with kernel Stein discrepancies

Summary The kernel Stein discrepancy (KSD) enables us to measure the similarity between a given distribution and a distribution defined by a density having an intractable normalizing constant. To use the KSD in practice, one needs to know what can be interpreted from the discrepancy. This chapter investigates the interpretability of the KSD in terms of moments; fundamental statistical quantities. The result in this chapter extends that of Gorham and Mackey [2017], who showed that the KSD controls the bounded-Lipschitz metric. Specifically, we prove that the KSD controls the integral probability metric defined by a class of continuous functions of polynomial growth, generalizing Lipschitz functions.

6.1 Introduction

Consider two probability distributions P and Q over $\mathbb{R}^D$, where P is defined by a density function p and Q is arbitrary. We assume that the density p may contain an unknown normalizing factor. Suppose we are interested in comparing the expectations of a test function f_0 under these two distributions. As a test function is usually only known up to certain properties (e.g., differentiability or growth conditions), a natural measure to consider is the worst case error, or an integral probability metric [Müller, 1997],

$$\sup_{f\in\mathcal{F}} \lvert \mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)] \rvert$$

over a function class $\mathcal{F}$ containing f_0. This setting may be interpreted as either evaluating the goodness of fit of a statistical model P against the data distribution Q, or assessing an approximation Q to a posterior distribution P in Bayesian inference. Our particular focus in this chapter is on continuous functions of polynomial growth. In this case, the IPM above implies disagreement in terms of moments, since they are defined by monomials of coordinates.

As the expectations are rarely analytically tractable, one might compare empirical estimates using samples from both distributions. In some scenarios, however, assuming access to samples from P may be problematic; e.g., *correct* samples from P are unavailable if the goal is to assess the quality of a Markov chain Monte Carlo sampler Q targeting intractable P. One possible way to sidestep the intractable integral is using a Stein discrepancy [Gorham and Mackey, 2015]

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_{Y\sim Q}[\mathcal{T}_P g(Y)]|,$$

where $\mathcal{T}_P$ is a Stein operator inducing functions whose expectations under P are zero, and $\mathcal{G}$ is some function class included in the domain of $\mathcal{T}_P$. Different choices of the operator and the function class yield distinct Stein discrepancies. Two major classes are the graph Stein discrepancies [Gorham and Mackey, 2015, Gorham et al., 2019] and the kernel Stein discrepancy (KSD) [Chwialkowski et al., 2016, Liu et al., 2016, Oates et al., 2017]. Notably, it is possible to compute these discrepancies. As we have seen in Chapter 3, the KSD circumvents the supremum and admits a closed-form expression involving kernel evaluations on samples, whereas the graph Stein discrepancy requires solving a linear program.

While successfully avoiding the intractable expectation, a drawback of Stein discrepancies is that they lack interpretability: the test functions $\mathcal{T}_P g$ in a Stein discrepancy are not immediately interpretable in terms of a given class $\mathcal{F}$ of interest. Following the spirit of Stein's method, this problem has been addressed by lower bounding a Stein discrepancy by a known IPM – convergence in the Stein discrepancy then implies the IPM convergence. Gorham and Mackey [2015] showed that the Langevin graph Stein discrepancy controls the L^1-Wasserstein distance (the IPM defined by 1-Lipschitz functions) for distantly dissipative target distributions; Gorham et al. [2019] later generalized this result to heavy-tailed targets with diffusion graph Stein discrepancies. Gorham and Mackey [2017] proved that the Langevin KSD with the inverse multi-quadratic kernel (IMQ) controls the bounded-Lipschitz metric; Chen et al. [2018] offer other kernel choices.

Despite its computational appeal, the KSD is limited in that it need not control convergence in unbounded functions, particularly functions of polynomial growth. This issue has been in part addressed by the aforementioned works on the graph Stein discrepancies. However, their analyses are limited to linearly growing functions – it is unclear how these results extend to functions of faster growth. Polynomially growing functions are of practical interest since they are related to fundamental statistical quantities, such as mean and variance (the latter corresponds to a quadratically growing function). Our objective is thus to extend the reach of the KSD to functions of arbitrary polynomial growth.

In this chapter, we investigate conditions under which the KSD controls the convergence of expectations of polynomially growing functions. Specifically, we prove that the KSD controls the IPM defined by a class of pseudo-Lipschitz functions, a polynomial generalization of Lipschitz functions. Our KSD bound builds on the result of finite Stein factors by Erdogdu et al. [2018]. Our specific contributions are twofold. First, our analysis considers the diffusion kernel Stein discrepancy (DKSD) [Barp et al., 2019], a generalization of the Langevin KSD [Chwialkowski et al., 2016, Liu et al., 2016, Oates et al., 2017]; this extension allows us to consider heavy-tailed targets. Second, we provide reproducing kernels required for the advertised convergence property to hold.

6.2 Background

We begin with background material required to present our main results.

6.2.1 Definitions and symbols specific to this chapter

Pseudo-Lipschitz functions. A function $h : \mathbb{R}^D \to \mathbb{R}$ is called pseudo-Lipschitz continuous (or simply C-pseudo-Lipschitz) of order q if it satisfies, for some constant $C > 0$,

$$|h(x) - h(y)| \le C(1 + \|x\|_2^q + \|y\|_2^q)\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^D, \tag{6.1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. We denote the smallest constant C satisfying (6.1) by $\tilde{\mu}_{\mathrm{pLip}}(h)_{1,q}$. We denote by $\mathrm{pLip}_{1,q}$ the set of pseudo-Lipschitz functions of order q with $\tilde{\mu}_{\mathrm{pLip}}(h)_{1,q} \le 1$. The pseudo-Lipschitz continuity generalizes the Lipschitz continuity (corresponding to the case $q = 0$) and allows us to describe functions of polynomial growth.

Vectors, matrices, and tensors. For a real vector v, $\|v\|_{\mathrm{op}} = \|v\|_2$. We identify an order L tensor $T \in \mathbb{R}^{D_1 \times \cdots \times D_L}$ as a multilinear map from $\mathbb{R}^{D_1} \times \cdots \times \mathbb{R}^{D_L}$ to $\mathbb{R}$ via the natural inner product

$$T : (u_1, \ldots, u_L) \in \mathbb{R}^{D_1 \times \cdots \times D_L} \mapsto \mathbb{R}, \left\langle T, v^{(1)} \otimes \cdots \otimes v^{(L)} \right\rangle = \sum_{i_1, \ldots, i_L} T_{i_1, \ldots, i_L} v_{i_1}^{(1)} \cdots v_{i_L}^{(L)},$$

and define the operator norm $\|T\|_{\mathrm{op}}$ by

$$\|T\|_{\mathrm{op}} = \sup_{\|u^{(l)}\|_2 = 1} \left| \left\langle T, u^{(1)} \otimes \cdots \otimes u^{(l)} \otimes \cdots \otimes u^{(L)} \right\rangle \right|.$$

Derivatives. The symbol $\nabla = (\partial_1, \ldots, \partial_D)^\top$ denotes the gradient operator with ∂_d denoting the partial derivative with respect to the d-th coordinate. The symbol ∇^m denotes the operator that outputs all the m-th order partial derivatives, defined as $(\nabla^m g(x))_{i_1, \ldots i_m} = \partial_{i_1} \cdots \partial_{i_m} g(x)$. For a vector-valued function $g : \mathbb{R}^D \to \mathbb{R}^{D'}$, we define $\nabla^i g : \mathbb{R}^D \to \underbrace{\mathbb{R}^D \otimes \cdots \otimes \mathbb{R}^D}_{i \text{ times}} \otimes \mathbb{R}^{D'}$ by $(\nabla^i g(x))_{k_1, \ldots, k_{i-1}, d} = (\nabla^i g_d(x))_{k_1, \ldots, k_{i-1}}$; i.e., ∇^i is applied to g element-wise. For a matrix valued function f, we define its column-wise divergence by $\langle \nabla, f(x) \rangle$; i.e., $\langle \nabla, f(x) \rangle_i = \sum_j \partial_j f_{ji}(x)$.

Generalized Fourier transform. Our main result (Proposition 6.13) requires an analogue of the Fourier transform of a function that is not an element of L^1 or L^2, where L^r, $r \in \{1, 2\}$ denotes the Banach space of r-integrable functions with respect to the Lebesgue measure. For our purpose, we use the following generalized Fourier transform.

Definition 6.1 (Generalized Fourier transform [Wendland, 2004, Defintion 8.9])**.** Let Φ be a continuous complex-valued function on $\mathbb{R}^D$ such that for some constant $q \ge 0$, $\Phi(x) = O(\|x\|_2^q)$ in the limit of $\|x\|_2 \to \infty$. A measurable function $\hat{\Phi}$ is called the generalized Fourier transform of Φ if it satisfies the following conditions: (a) the restriction of $\hat{\Phi}$ on every compact set $K \subset \mathbb{R}^D \setminus \{0\}$ is square integrable, and (b) there exists a nonnegative integer m such that

$$\int \Phi(x)\hat{\gamma}(x)\mathrm{d}x = \int \hat{\Phi}(\omega)\gamma(\omega)\mathrm{d}\omega$$

is true for all Schwartz functions γ satisfying $\gamma(\omega) = O(\|\omega\|_2^{2m})$ for $\|\omega\|_2 \to 0$. The integer m is called the order of $\hat{\Phi}$.

The Fourier transform of Φ coincides with the generalized Fourier transform, if it exists. In the following, we limit ourselves generalized Fourier transforms of order zero.

6.2.2 Generators of Itô diffusions and their Stein equations

We first consider the generator of an Itô diffusion as a Stein operator. This object is related to the diffusion Stein operator considered in the next Section; the diffusion Stein operator is a first-order differential operator, whereas generators considered in this section are of the second order.

For a function f pseudo-Lipschitz of order q, consider the Stein equation

$$\mathcal{A}_P u_f = f - \mathbb{E}_{X\sim P}[f(X)], \tag{6.2}$$

defined by the generator $\mathcal{A}_P$ of an Itô diffusion with invariant distribution P, where $u_h : \mathbb{R}^D \to \mathbb{R}$ is a solution to the equation. The diffusion is defined by the following stochastic differential equation:

$$\mathrm{d}Z_t^x = b(Z_t^x)\mathrm{d}t + \sigma(Z_t^x)\mathrm{d}B_t \text{ with } Z_0^x = x. \tag{6.3}$$

Here, $(B_t)_{t\geq 0}$ is a D'-dimensional Wiener process; $b \in C^1 : \mathbb{R}^D \to \mathbb{R}^D$ and $\sigma \in C^1 : \mathbb{R}^D \to \mathbb{R}^{D\times D'}$ represent the drift and the diffusion coefficients. The drift coefficient b is chosen so that the diffusion has P as an invariant measure:

$$b(x) = \frac{1}{2p(x)} \langle \nabla, p(x)\{a(x) + c(x)\}\rangle ,$$

where $a(x) = \sigma(x)\sigma(x)^\top$ is the covariance coefficient, $c(x) = -c(x)^\top \in \mathbb{R}^{D\times D}$ is the skew-symmetric stream coefficient. Then, the generator $\mathcal{A}_P$ is an operator defined by

$$\mathcal{A}_P u_f(x) := \langle b(x), \nabla u_f(x)\rangle + \frac{1}{2}\left\langle \sigma(x)\sigma(x)^\top, \nabla^2 u_f(x)\right\rangle .$$

Characterizing the regularity of a solution u_f to (6.2) requires additional assumptions on the diffusion. Erdogdu et al. [2018] revealed that a solution is a pseudo-Lipschitz function for a fast-converging diffusion. To introduce their result, we first detail required assumptions.

Condition 6.2 (Polynomial growth of coefficients)**.** For some $q_a \in \{0, 1\}$ and any $x \in \mathbb{R}^D$, the drift and the diffusion coefficients of (6.3) satisfy the growth condition

$$\|b(x)\|_2 \leq \frac{\lambda_b}{4}(1+\|x\|_2), \|\sigma(x)\|_{\mathrm{F}} \leq \frac{\lambda_\sigma}{4}(1+\|x\|_2), \text{and } \|\sigma(x)\sigma(x)^\top\|_{\mathrm{op}} \leq \frac{\lambda_a}{4}(1+\|x\|_2^{q_a+1}),$$

with $\lambda_b, \lambda_\sigma, \lambda_a > 0$.

Condition 6.3 (Dissipativity)**.** For $\alpha, \beta > 0$, the diffusion (6.3) satisfies the dissipativity condition

$$\mathcal{A}_P \|x\|_2^2 \leq -\alpha\|x\|_2^2 + \beta.$$

The operator $\mathcal{A}_P$ is the generator of an Itô diffusion with coefficients b and σ, and $\mathcal{A}_P\|x\|_2^2 = 2\langle b(x), x\rangle + \|\sigma(x)\|_F^2$.

Condition 6.4 (Wasserstein rate)**.** For $q \geq 1$, the diffusion Z_t^x has L^q-Wasserstein rate $\rho_q : [0, \infty) \to \mathbb{R}$ if

$$\inf_{\text{couplings}(Z_t^x, Z_t^y)} \mathbb{E}[\|Z_t^x - Z_t^y\|_2^q]^{1/q} \leq \rho_q(t)\|x - y\|_2 \text{ for } x, y \in \mathbb{R}^D \text{ and } t \geq 0,$$

where the infimum is taken over all couplings between Z_t^x and Z_t^y. We further define the relative rates

$$\tilde{\rho}_1(t) = \log(\rho_2(t)/\rho_1(t)) \text{ and } \tilde{\rho}_2(t) = \log[\rho_1(t)/\{\rho_2(t)\rho_1(0)\}]/\log[\rho_1(t)/\rho_1(0)].$$

Erdogdu et al. [2018] shows that the solution u_f to the Stein equation (6.2) satisfies the following property:

Theorem 6.5 (Finite Stein factors from Wasserstein decay, Erdogdu et al. 2018, Theorem 3.2)**.** *Assume that Conditions 6.2, 6.3, and 6.4 for the L^1-Wasserstein distance hold and that f is pseudo-Lipschitz of order q with at most degree-q polynomial growth of its i-th derivatives for $i = 2, 3, 4$. Then, the solution u_f to the equation (6.2) is pseudo-Lipschitz of order q with constant ζ_1, and has i-th order derivative with degree-q polynomial growth for $i = 2, 3, 4$:*

$$\|\nabla^i u_f(x)\|_{\text{op}} \leq \zeta_i(1 + \|x\|_2^q) \textit{ for } i \in \{2, 3, 4\}, \textit{ and } x \in \mathbb{R}^D.$$

The constants ζ_i (called Stein factors) are given as follows:

$$\zeta_i = \tau_i + \xi_i \int_0^\infty \rho_1(t)\omega_{q_a+1}(t + i - 2)\mathrm{d}t \textit{ for } i = 1, 2, 3, 4,$$

where

$$\omega_{q_a+1}(t) = 1 + 4\rho_1(t)^{1-1/(q_a+1)}\rho_1(0)^{\frac{1}{2}}\left[1 + \frac{1}{\tilde{\alpha}_{q_a+1}^q}\{(1 \vee \tilde{\rho}_{q_a+1}(t))2\lambda_a q + 3(q_a + 1)\beta\}^q\right],$$

with $\tilde{\alpha}_1 = \alpha$, $\tilde{\alpha}_2 = \inf_{t\geq 0}[\alpha - q\lambda_a(1 \vee \tilde{\rho}_2(t)]_+$,

$$\tau_1 = 0,\ \tau_i = \tilde{\mu}_{\text{pLip}}(f)_{1,q}\tilde{\pi}(f)_{2:i,q}\tilde{\nu}_{1:q}(\sigma)\kappa_{q_a}(6q) \textit{ for } i = 2, 3, 4,$$
$$\xi_1 = \tilde{\mu}_{\text{pLip}}(f)_{1,q}, \xi_i = \tilde{\mu}_{\text{pLip}}(f)_{1,q}\tilde{\nu}_{1:i}(b)\tilde{\nu}_{0:i-2}(\sigma^{-1})\rho_1(0)\omega_{q_a+1}(1)\kappa_{q_a+1}(6q)^{i-1} \textit{ for } i = 2, 3, 4,$$

where $\tilde{\pi}(f)_{i,q} = \sup_{x\in\mathbb{R}^D}\|\nabla^i f(x)\|_{\text{op}}/(1 + \|x\|^q)$, $\tilde{\pi}(f)_{a:b,q} := \max_{i=a,\ldots,b}\tilde{\pi}(f)_{i,q}$, $\tilde{\nu}_{a:b}(g)$ is *a constant whose precise form is given in the proof of Erdogdu et al. [2018, Theorem 3.2], and*

$$\kappa_{q_a+1}(q) = 2 + \frac{2\beta}{\alpha} + \frac{q\lambda_a}{4\alpha} + \frac{\tilde{\alpha}_{q_a+1}}{\alpha}\left(\frac{q\lambda_a + 6(q_a+1)\beta}{2r\tilde{\alpha}_{q_a+1}}\right).$$

There are two known sufficient conditions for establishing exponential Wasserstein decay (Condition 6.4). The first is uniform dissipativity, which is a simple (but more restrictive) condition leading to exponential L^1- or L^2- exponential decay rates.

Proposition 6.6 (Wasserstein decay from uniform dissipativity, Wang 2020, Theorem 2.5)**.** *A diffusion with drift and diffusion coefficients b and σ has L^q-Wasserstein rate $\rho_q(t) = e^{-rt/2}$, if for all $x, y \in \mathbb{R}^D$,*

$$2\langle b(x) - b(y), x - y\rangle + \|\sigma(x) - \sigma(y)\|_F^2 - (q-2)\|\sigma(x) - \sigma(y)\|_{\mathrm{op}}^2 \leq -r\|x - y\|_2^2.$$

The second and more general condition is distant dissipativity. Explicit L^1-Wasserstein decay rates from distant dissipativity are obtained by the following result of Gorham et al. 2019, which builds upon the analyses of Eberle [2015] and Wang [2020].

Proposition 6.7 (Wasserstein decay from distant dissipativity, Gorham et al. 2019, Corollary 12)**.** *A diffusion with drift and diffusion coefficients b and σ is called distantly dissipative if for the truncated diffusion coefficient*

$$\tilde{\sigma}(x) := (\sigma(x)\sigma(x)^\top - s^2 I)^{1/2} \text{ with } s \in [0, 1/M_0(\sigma^{-1}))$$

with $M_0(\sigma^{-1}) = \sup_{x\in\mathbb{R}^D}\|\sigma^{-1}(x)\|_{\mathrm{op}}$, it satisfies

$$2\frac{\langle b(x) - b(y), x - y\rangle}{s\|x - y\|_2^2} + \frac{\|\tilde{\sigma}(x) - \tilde{\sigma}(y)\|_F^2}{s^2\|x - y\|_2^2} - \frac{\|(\tilde{\sigma}(x) - \tilde{\sigma}(y))^\top (x - y)\|_F^2}{s^2\|x - y\|_2^4}$$
$$\leq \begin{cases} -K & \|x - y\|_2 > R \\ L & \|x - y\|_2 \leq R \end{cases}$$

for some $K > 0$ and $R, L \geq 0$. If the distant dissipativity holds, then the diffusion has Wasserstein rate $\rho_1(t) = 2e^{LR^2/8}e^{-rt/2}$ for

$$s^2 r^{-1} \leq \begin{cases} \frac{e-1}{2}R^2 + e\sqrt{8K^{-1}}R + 4K^{-1} & \text{if } LR^2 \leq 8, \\ 8\sqrt{2\pi}R^{-1}L^{-1/2}(L^{-1} + K^{-1})\exp\left(\frac{LR^2}{8}\right) + 32R^{-2}K^{-2} & \text{otherwise.} \end{cases}$$

These two conditions also conveniently lead to the dissipativity condition (Condition 6.3) defined above. Therefore, we assume that the diffusion satisfies either of these conditions in the following.

6.2.3 Diffusion Stein operators and their Stein equations

In this section, we recall the diffusion Stein operator of Gorham et al. [2019]. For a diffusion process defined in (6.3), the diffusion Stein operator is defined as an operator that takes as input a vector-valued differentiable function $g : \mathbb{R}^D \to \mathbb{R}^D$ and outputs a real-valued function as follows:

$$\begin{aligned}\mathcal{T}_P g(x) &= \frac{1}{p(x)} \langle \nabla, p(x) m(x) g(x) \rangle \\ &= 2\langle b(x), g(x) \rangle + \langle m(x), \nabla g(x) \rangle\end{aligned}$$

where $b(x) = \langle \nabla, p(x)m(x) \rangle / (2p(x))$ and $m(x) = a(x) + c(x)$. Note that we can recover the Langevin Stein operator [Gorham and Mackey, 2015] by taking $a \equiv I$ and $c \equiv 0$.

Recall the Stein equation for the second-order Stein operator $\mathcal{A}_P u_f = f - \mathbb{E}_{X\sim P}[f(X)]$. By relating the definition of $\mathcal{A}_P$ to $\mathcal{T}_P$, we have that the function $g_f = \nabla u_f / 2$ solves the Stein equation $\mathcal{T}_P g = f - \mathbb{E}_{X\sim P}[f(X)]$ [Gorham et al., 2019, Section 2]. As a result of Theorem 6.5, we have the following corollary:

Corollary 6.8. *Let $q \geq 0$. Let $f \in C^3$ be a pseudo-Lipschitz function of order q with derivatives satisfying the polynomial decay condition in Theorem 6.5. The solution $g_f = \nabla u_f/2$ to the Stein equation $\mathcal{T}_P g = f - \mathbb{E}_{X\sim P}[f(X)]$ belongs to the set*

$$\begin{aligned}\mathcal{G} = \{ g : \mathbb{R}^D \to \mathbb{R}^D : \|g(x)\|_2 \leq \sqrt{D}\zeta_1 (1 + \|x\|_2^q), \\ \|\nabla^i g(x)\|_{\text{op}} \leq \zeta_{i+1}(1 + \|x\|_2^q) \textit{ for } i \in \{1,2\}, x \in \mathbb{R}^D \},\end{aligned}$$

where ζ_1, ζ_2, and ζ_3 are the Stein factors from Theorem 6.5.

Proof. The derivative norm bounds follow directly from Theorem 6.5. Note that the bound on $\|g(x)\|_2$ follows from the pseudo-Lipschitzness of u_f as

$$\begin{aligned}|\partial_j u_f(x)| &= \lim_{h\to 0} \left| \frac{u_f(x + he_j) - u_f(x)}{h} \right| \\ &\leq \lim_{h \to 0} \zeta_1 (1 + \|x + he_j\|_2^{q-1} + \|x\|_2^{q-1}) \frac{\|he_j\|_2}{|h|} \\ &= \zeta_1(1 + 2\|x\|_2^{q-1}) \leq 2\zeta_1(1 + \|x\|_2^{q-1}),\end{aligned}$$

where $\{e_1, \ldots, e_D\}$ is the standard basis of $\mathbb{R}^D$. □

The following proposition shows that the diffusion Stein operator induces zero-mean functions (see Appendix 6.A.3 for a proof).

Proposition 6.9 (The diffusion Stein operator generates zero-mean functions)**.** *Let $q_a \in \{0, 1\}$ be the additional growth exponent of $\|a(x)\|_{\text{op}}$ from Condition 6.2. If $q_a = 0$, assume P has a finite q-th moment; if $q_a = 1$, a finite $(q+1)$-th moment. Let $g \in C^1$ be a function with the following growth conditions:*

$$\begin{aligned}\|g(x)\|_2 &\leq C_0(1 + \|x\|_2^{q-1}), \\ \|\nabla g(x)\|_{\text{op}} &\leq C_1(1 + \|x\|_2^{q-1}),\end{aligned}$$

for each $x \in \mathbb{R}^D$, and some positive constants C_0 and C_1. Then, we have $\mathbb{E}_{X\sim P}[\mathcal{T}_P g(X)] = 0$.

6.2.4 The diffusion kernel Stein discrepancy

We recall the definition of the diffusion kernel Stein discrepancy (DKSD) proposed by Barp et al. [2019]. As with the Langevin KSD in Chapter 3, we can construct a computable Stein discrepancy by combining the diffusion Stein operator with an RKHS. Let P and Q be distributions over $\mathbb{R}^D$. The DKSD is defined as follows:

$$\mathcal{S}\big(Q, \mathcal{T}_P, \mathcal{B}_1(\mathcal{G}_\kappa)\big) = \sup_{g \in \mathcal{B}_1(\mathcal{G}_\kappa)} \left|\mathbb{E}_{Y\sim Q}\left[\mathcal{T}_P g(Y)\right]\right|,$$

where $\mathcal{B}_1(\mathcal{G}_\kappa)$ is the unit ball of a vector-valued RKHS $\mathcal{G}_\kappa$ determined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$ [Carmeli et al., 2006]. In the following, by abuse of notation, we denote $\mathcal{S}\big(Q, \mathcal{T}_P, \mathcal{B}_1(\mathcal{G}_\kappa)\big)$ by $\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa)$. Note that the Langevin KSD is obtained as a specific instance of the DKSD by choosing $a(x) \equiv I$, $c \equiv 0$, and $\kappa = kI$ for a scalar-valued kernel k and the D-dimensional identity matrix I. In particular, we obtain the KSD with a reweighted kernel $w(x)k(x,y)w(y)$ by choosing $a(x) = w(x)I$ with w a scalar-valued positive function. As with the Langevin KSD, the use of an RKHS leads to a closed-form expression for the DKSD:

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa)^2 = \mathbb{E}_{X,X'\sim Q\otimes Q}[h_p(X, X')],$$

where

$$h_p(x,y) = \frac{1}{p(x)p(y)} \Big\langle \nabla_y, \langle \nabla_x, (p(x)m(x)\kappa(x,y)m(y)^\top p(y)) \rangle \Big\rangle,$$

provided that $x \mapsto \|\mathcal{T}_p \kappa(x,\cdot)\|_{\mathcal{G}_\kappa}$ is integrable with respect to Q [Barp et al., 2019, Theorem 1]. Although the majority of the following analyses focus on the simple RKHS $\kappa = kI$, we present results using a general vector-valued RKHS where possible.

6.3 Main results

We formalize our objective in this chapter, as outlined in the introduction. Let P be a distribution over $\mathbb{R}^D$ defined by a twice continuously differentiable density function p. For a distribution Q, we are interested in the maximum discrepancy over a function class $\mathcal{F}$: $\sup_{f\in\mathcal{F}}|\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]|$. Our focus is on functions with growing in the polynomial order of $q \geq 1$. Therefore, we specify the function class $\mathcal{F}$ to be a subset of pseudo-Lipschitz functions of order $q-1$

$$\begin{aligned}\mathcal{F}_q \coloneqq \{ & \\ & f : \mathbb{R}^D \to \mathbb{R} : f \in C^3 \text{ with } \tilde{\mu}_{\mathrm{pLip}}(f)_{1,q-1} \leq 1 \\ & \text{and } \sup_x \|\nabla^i f(x)\|_{\mathrm{op}} / \big(1 + \|x\|_2^{q-1}\big) \leq 1 \text{ for } i \in \{2,3\} \\ \}. & \end{aligned}$$

We define an IPM corresponding to the class $\mathcal{F}_q$ by

$$d_{\mathcal{F}_q}(P, Q) = \sup_{f \in \mathcal{F}_q} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right|. \tag{6.4}$$

Appendix 6.B.5 shows that $\mathcal{F}_q$ contains degree-q polynomial functions of $(x_1, \ldots, x_D) \in \mathbb{R}^D$, if scaled appropriately. Note that for this class, we can take the Stein factors $\{\zeta_1, \zeta_2, \zeta_3\}$ that are independent of test functions and only depend on the diffusion through $b(x)$ and $\sigma(x)$. Unfortunately, the IPM $d_{\mathcal{F}_q}$ is not computable as it involves an intractable integral. Thus, we aim to relate the IPM $d_{\mathcal{F}_q}$ to the DKSD, a computable discrepancy measure.

Before presenting results concerning the DKSD, we first show a convergence property of $d_{\mathcal{F}_q}$. In the following, we denote the set of probability measures with finite q-th moments by $\mathcal{P}_q := \{\text{probability measure } \mu : \int \|x\|_2^q \mathrm{d}\mu(x) < \infty\}$.

Proposition 6.10. *Let $P \in \mathcal{P}_q$ be a probability measure on $\mathbb{R}^D$ with a finite q-th moment with $q \geq 1$. For a sequence of probability measures $\{Q_1, Q_2, \ldots, \} \subset \mathcal{P}_q$, the following conditions are equivalent: (a) $d_{\mathcal{F}_q}(Q_n, P) \to 0$ as $n \to \infty$, and (b) as $n \to \infty$, the sequence Q_n converges weakly to P, and $\mathbb{E}_{X \sim Q_n}\left[\|X\|_2^q\right] \to \mathbb{E}_{X \sim P}\left[\|X\|_2^q\right]$.*

Proof. We relate the metric $d_{\mathcal{F}_q}$ to the bounded-Lipschitz metric [see, e.g., Dudley, 2002, Section 11.2] and make use of uniform integrability. See Appendix 6.A.1 for a complete proof. □

The above result shows that the distance $d_{\mathcal{F}_q}$ characterizes both weak convergence (convergence in distribution) and convergence with respect to the q-th moment. For sequences in $\mathcal{P}_q$, another example of a metric characterizing this topology is L^q-Wasserstein distance defined by the Euclidean distance [see, e.g., Villani, 2009, Theorem 6.9].

In the following, we make an additional assumption on the growth of the stream coefficient in the DKSD.

Condition 6.11 (Polynomial growth of the stream coefficient)**.** For any $x \in \mathbb{R}^D$, the stream coefficient of (6.3) satisfies the growth condition

$$\|c(x)\|_{\mathrm{op}} \leq \frac{\lambda_c}{4}(1 + \|x\|_2)^{q_a+1}$$

with $\lambda_c > 0$ and q_a as in Condition 6.2.

Note that under Conditions 6.2, 6.11, we have

$$\|m(x)\|_{\mathrm{F}} \leq \sqrt{D}\|m(x)\|_{\mathrm{op}} \leq \sqrt{D}\lambda_m(1 + \|x\|_2)^{q_a+1} \text{ with } \lambda_m = \frac{\lambda_a \vee \lambda_c}{4}.$$

6.3.1 Uniform integrability and DKSD bounds

To prove our main result, we make use of an analogue of uniform integrability for a family of functions, defined as follows:

Definition 6.12 (Uniform integrability with respect to q-th moments)**.** Let $q > 0$. A sequence of probability measures $\mathcal{Q} = \{Q_1, Q_2 \dots, \} \subset \mathcal{P}_q$ is said to have uniformly integrable q-th moments if

$$\lim_{r\to\infty} \limsup_{n\to\infty} \int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q_n(x) = 0.$$

It is known that for weakly converging sequences, convergence of a moment is equivalent to the uniform integrability of the moment [Ambrosio et al., 2005, Lemma 5.1.7] (see also Lemma 6.32 in Appendix 6.B.2). Intuitively, uniform integrability is a condition enforcing that the probability mass does not diverge too fast along the sequence. If $q = 0$, the above definition is reduced to uniform tightness [see, e.g., Dudley, 2002, Chapter 9]. The uniform integrability is a stricter condition in that it requires the decay rate of the tail probability to stay faster than a q-th degree polynomial.

Our first result shows that for sequences with a uniformly integrable moment, the DKSD implies convergence in $d_{\mathcal{F}_q}$.

Proposition 6.13. *Let $\Phi \in C^2$ be a positive definite function with non-vanishing generalized Fourier transform $\hat{\Phi}$. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w}$ with real numbers $v \geq 1$ and $q_w \geq 0$. Let $\mathcal{G}_{kI}$ be the RKHS defined by kernel kI with $k(x, y) = \Phi_w(x, x') + \ell(x, x')$, where $\Phi_w(x, x') = w(x)w(x')\Phi(x - x')$,and ℓ is an optional positive definite kernel. Then, for a sequence of measures $\mathcal{Q} = \{Q_1, Q_2, \dots, \} \subset \mathcal{P}_{q+q_a}$ with uniformly integrable $(q + q_a)$-th moments, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{kI}) \to 0$ only if $d_{\mathcal{F}_q}(P, Q_n) \to 0$.*

Proof. We only provide a proof sketch here and refer the reader to Appendix 6.A.2 (see also Proposition 6.26 for the precise statement) for the full proof. Our strategy is to bound the IPM $d_{\mathcal{F}_q}$ explicitly by the DKSD. Consider the Stein equation $\mathcal{T}_P g_f = f - \mathbb{E}_{X\sim P}[f(X)]$ for $f \in \mathcal{F}_q$. We approximate $\mathcal{T}_p g_f$ by mollification. Specifically, we decompose the function $\mathcal{T}_P g_f$ into three parts:

$$\mathcal{T}_P g_f = \underbrace{\mathcal{T}_P g_f - \mathcal{T}_P g_{\mathrm{trunc}}}_{\text{Step 1: truncation}} + \underbrace{\mathcal{T}_P g_{\mathrm{trunc}} - \mathcal{T}_P g_{\mathrm{RKHS}}}_{\text{Step 2: approximation}} + \underbrace{\|g_{\mathrm{RKHS}}\|}_{\text{Step 3: measuring RKHS norm}} \frac{\mathcal{T}_P g_{\mathrm{RKHS}}}{\|g_{\mathrm{RKHS}}\|}.$$

In the first step, we truncate $\mathcal{T}_p g_f$ to make it bounded and ignore the tail expectation. In the second step, we approximate the truncated function with a smooth function. In the last step, we show that the smooth function is an RKHS function if the kernel Φ is chosen appropriately. The expectations of truncation and approximation errors are evaluated using the characterization of g_f derived in Corollary 6.8; the expectation of the third term can be bounded by the DKSD. Thus, we obtain a bound on $d_{\mathcal{F}_q}$ in terms of the DKSD. If the DKSD term vanishes, the rest of the error terms can be made arbitrarily small. □

When $q_a = 1$, i.e., we have a quadratic growth of the operator norm of the covariance coefficient, we need approximating distributions $\{Q_1, Q_2, \dots\}$ to have an extra moment. This requirement is placed in order to make the Stein discrepancy upper bound well-defined; otherwise, the upper bound is vacuous. Indeed, if one is willing to make the assumption $Q \neq P$, it is appropriate to assume a higher-order moment. For example, any numerical quadrature method is represented by a discrete measure with a finite support, thereby guaranteeing a finite moment of any order. Note that in proving the uniform integrability of a sequence of measures $\mathcal{Q} = \{Q_1, Q_2 \ldots\}$, it is not sufficient to assume that each $Q \in \mathcal{Q}$ has a finite moment of a higher order (a sufficient condition is having their moments of that order uniformly bounded).

6.3.2 The DKSD detects non-uniform integrability

In the previous section, we assumed that the approximating distributions $\{Q_1, Q_2, \ldots\}$ had uniformly integrable q-th moments if $\|a(x)\|_{\mathrm{op}}$ grows linearly, or $(q+1)$-th moments if quadratically. Proposition 6.13 indicates that we can use any kernel formed by a positive definite function for diagnosing non-convergence. However, such a kernel alone is not enough when the uniform integrability is violated. Indeed, in the case of weak convergence, Gorham and Mackey [2017] demonstrated that an inadequate choice of the kernel yields a discrepancy that converges to zero even when the sequence is not uniformly tight and hence non-convergent; the IMQ kernel suggested by Gorham and Mackey [2017] ensures that vanishing KSD implies the tightness of the sequence. Analogously, we explore conditions that allow us to check the uniform integrability using the DKSD.

The following two lemmas characterize uniform integrability using the DKSD:

Lemma 6.14. *Let $\mathcal{Q} = \{Q_1, Q_2, \ldots\} \subset \mathcal{P}_q$ be a sequence of probability measures for $q > 0$. Let $\mathcal{G}_\kappa$ be the RKHS of $\mathbb{R}^D$-valued functions defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Suppose that for any $\varepsilon > 0$, there exists $r_\varepsilon > 0$ and a function $g \in \mathcal{G}_\kappa$ such that $\mathcal{T}_P g(x) \geq \|x\|_2^q 1\{\|x\|_2 > r_\varepsilon\} - \varepsilon$ for any $x \in \mathbb{R}^D$. Then, $\mathcal{Q}$ has uniformly integrable q-th moments if $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) \to 0$ as $n \to \infty$.*

Proof. For any $\varepsilon > 0$, we have

$$\int_{\{\|x\|_2 > r_\varepsilon\}} \|x\|_2^q \mathrm{d}Q_n(x) \leq \int \mathcal{T}_p g(x) \mathrm{d}Q_n(x) \leq \|g\|_{\mathcal{G}_\kappa} \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) + \varepsilon.$$

Letting $n \to \infty$ concludes the proof. □

Lemma 6.15 (KSD upper-bounds the integrability rate)**.** *Let $\mathcal{G}_\kappa$ be the RKHS of $\mathbb{R}^D$-valued functions defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Suppose there exists a function $g \in \mathcal{G}_\kappa$ such that $\mathcal{T}_P g(x) \geq \nu$ for any $x \in \mathbb{R}^D$ with some constant $\nu \in \mathbb{R}$, and $\liminf \|x\|_2^{-(q+\theta)} \mathcal{T}_P g(x) \geq \eta$ for some $q \geq 0$, $\eta > 0$, and $\theta > 0$ as $\|x\|_2 \to \infty$. Assume $\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) < \infty$ for a distribution $Q \in \mathcal{P}_{q+\theta}$. Then, for sufficiently small $\varepsilon > 0$, we have*

$$\begin{aligned} R_q(Q, \varepsilon) &:= \inf\left\{ r \geq 1 : \int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q(x) \leq \varepsilon \right\} \\ &\leq \left\{ 2\left(1 + \frac{q}{\theta}\right) \left(\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) - \nu}{\eta\varepsilon} \right) \right\}^{\frac{1}{\theta} \vee \frac{q}{\theta}}. \end{aligned}$$

Thus, for a sequence of measures $\{Q_1, Q_2 \dots\} \subset \mathcal{P}_{q+\theta}$, we have

$$\limsup_{n\to\infty} \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) < \infty \Rightarrow \limsup_{n\to\infty} R_q(Q_n, \varepsilon) < \infty.$$

In particular, if the sequence $\{Q_1, Q_2 \dots\}$ does not have uniformly integrable q-th moments, then Stein discrepancy $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa)$ diverges.

Proof. The proof is in Appendix 6.B.4.1. □

The quantity $R_q(Q, \varepsilon)$ (termed an integrability rate) in Lemma 6.15 represents the radius of a ball, outside of which the tail moment integral becomes negligible. Note that $Q \in \mathcal{P}_q$ is equivalent to having $R_q(Q, \varepsilon) < \infty$ for each $\varepsilon > 0$. In particular, if a sequence $\{Q_n\}_{n\geq 1}$ does not have uniformly integrable q-th moments, the integrability rate $R_q(Q_n, \varepsilon)$ diverges.

The above two lemmas require the Stein-modified RKHS to have a function growing at a certain rate. The first lemma requires a stronger condition in that it requires a function that approximates the power function $\|x\|_2^q$ arbitrarily well. The existence proof for the first lemma is left as future work, and we focus on the second lemma. Note that the second lemma in contrast relies on the existence of a function that behaves as $\|x\|_2^q$ outside a ball; we can create an RKHS that satisfy this requirement using a linear kernel and the diffusion Stein operator, provided that the diffusion satisfies the dissipativity condition (Condition 6.3).

Lemma 6.16 (Tilted linear kernels have the lower bound properties)**.** *Suppose the diffusion targeting P satisfies the dissipativity condition (Condition 6.3) with $\alpha, \beta > 0$ and the coefficient condition (Condition 6.2) with $\lambda_a > 0$ and $q_a \in \{0, 1\}$. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w - u}$ with $q_w \geq 0,\ u \geq 0,$ and $v > 0$. Assume $(q_w - u) < 2\alpha/\lambda_a$ if $q_a = 1$. Let*

$$k(x, x') = w(x)w(x')\langle x, x'\rangle.$$

There exists a function $g \in \mathcal{G}_{kI}$ such that $\|g\|_{\mathcal{G}_{kI}} = \sqrt{D}$ and the corresponding diffusion Stein operator $\mathcal{T}_P$ satisfies

$$\mathcal{T}_P g(x) \geq \nu \text{ for any } x \in \mathbb{R}^D, \text{ and } \liminf_{\|x\|_2\to\infty} \|x\|_2^{-2(q_w-u+1)} \mathcal{T}_P g(x) \geq \eta$$

for some $\nu \in \mathbb{R}$ and $\eta > 0$.

Proof. The proof can be found in Appendix 6.B.4.2. □

The next result is an immediate consequence of Lemma 6.16.

Corollary 6.17. *Define symbols as in Lemma 6.16. For the RKHS $\mathcal{G}_{kI}$ of a kernel $k = k_1 + k_2$ with k_1 an arbitrary positive definite kernel and k_2 the kernel from Lemma 6.16, there exists a function $g \in \mathcal{G}_{kI}$ such that the corresponding diffusion Stein operator $\mathcal{T}_P$ satisfies*

$$\mathcal{T}_P g(x) \geq \nu \text{ for any } x \in \mathbb{R}^D, \text{ and } \liminf_{\|x\|_2\to\infty} \|x\|_2^{-2(q_w-u+1)} \mathcal{T}_P g(x) \geq \eta$$

for some $\nu \in \mathbb{R}$ and $\eta > 0$. In particular, if the RKHSs of kernels k_1 and k_2 do not overlap, $\|g\|_{\mathcal{G}_{kI}} = \sqrt{D}$.

6.3.3 Recommended kernel choice

We have established the conditions required for the DKSD to control the pseudo-Lipschitz metric $d_{\mathcal{F}_q}$. In the following, we present our recommended settings.

Linear growth case When $q_a = 0$, we need the uniform integrability with respect to the q-th moment. We recommend the following kernel function

$$k_{\Phi,q,\theta}(x, x') = w_{q,\theta}(x; v)w_{q,\theta}(x'; v)\Big(\Phi(x - x') + \tilde{k}_{\text{lin}}(x, x'; v)\Big), \tag{6.5}$$

where the weight $w_{q,\theta}$ is given by

$$w_{q,\theta}(x; v) = (v^2 + \|x\|_2^2)^{\frac{q+\theta-1}{2}}$$

with $v \geq 1$, and $\tilde{k}_{\text{lin}}$ denotes the normalized linear kernel:

$$\tilde{k}_{\text{lin}}(x, x') = \frac{v^2 + \langle x, x'\rangle}{\sqrt{v^2 + \|x\|_2^2}\sqrt{v^2 + \|x'\|_2^2}}.$$

This choice ensures that the two kernels in the sum (6.5) have the same growth rate. The kernel $\Phi(x - x')$ can be any function with a non-vanishing (generalized) Fourier transform. Examples are the exponentiated quadratic (EQ) kernel, the IMQ kernel, and the Matérn class kernels [Matérn, 1986, Stein, 1999]. Note that the normalized linear kernel enables us to use light-tailed kernels (e.g., the EQ or the Matérn kernels) that are recommended against by Gorham and Mackey [2017]. In particular, one benefit of the Matérn class is that we can use rougher functions: the RKHS of an IMQ kernel consists of infinitely differentiable functions, whereas a Matérn kernel can specify an RKHS of finitely differentiable functions. The increased complexity would render the DKSD more sensitive to the difference between distributions (see Section 6.4.3).

Quadratic growth case When $q_a = 1$, we need the uniform integrability with respect to the $(q + 1)$-th moment. Therefore, we recommend to use the same form of the kernel as in (6.5) except that $w_{q,\theta}$ is replaced with

$$w_{q+1,\theta}(x) = (v^2 + \|x\|_2^2)^{\frac{q+\theta}{2}}.$$

6.3.4 The DKSD detects convergence

We have shown that vanishing DKSD implies convergence in $d_{\mathcal{F}_q}$. Here, we clarify conditions on which the DKSD converges to zero.

Proposition 6.18. *Let $\mathcal{G}_\kappa$ be the RKHS defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Let $q \geq 1$. Assume that any function g in the unit ball $\mathcal{B}_1(\mathcal{G}_\kappa)$ satisfies the following: there exist some constants C_0, C_1, and C_2 such that*

$$\|\nabla^i g(x)\|_{\text{op}} \leq C_i(1 + \|x\|_2^{q-1}) \textit{ for any } x \in \mathbb{R}^D \textit{ and } i \in \{0, 1, 2\}.$$

Assume a linear growth condition on m (Conditions 6.2, 6.11 with $q_a = 0$). Assume that b is $\phi_1(b)$-Lipschitz in the Euclidean norm, and m is $\phi_1(m)$-Lipschitz in the Frobenius norm. Suppose $P \in \mathcal{P}_q$. Then,

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) \leq C_{b,m} d_{\mathrm{pLip}_{1,q}}(Q, P),$$

where

$$C_{b,m} = \frac{\lambda_b C_1(5 + 2^{q-1})}{4} + 4C_0\phi_1(b) + \frac{\lambda_m C_2 D(5 + 2^{q-1})}{2} + 2\sqrt{D}C_1\phi_1(m)$$

In particular, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) \to 0$ if $d_{\mathcal{F}_q}(P, Q_n) \to 0$.

Proof. The claim follows from showing that a Stein-modified RKHS function $\mathcal{T}_P g$ is pseudo-Lipschitz of order q. The full proof is available in Appendix 6.A.4. □

Proposition 6.19. *Let $\mathcal{G}_\kappa$ be the RKHS of $\mathbb{R}^D$-valued functions defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Let $q \geq 1$. Assume that any function g in the unit ball $\mathcal{B}_1(\mathcal{G}_\kappa)$ satisfies the following: there exist some constants C_0, C_1, and C_2 such that*

$$\|\nabla^i g(x)\|_{\mathrm{op}} \leq C_i(1 + \|x\|_2^{q-1}) \textit{ for any } x \in \mathbb{R}^D \textit{ and } i \in \{0, 1, 2\}.$$

Assume a quadratic growth condition on m (Conditions 6.2, 6.11 with $q_a = 1$). Assume that b is $\phi_1(b)$-Lipschitz, and m is pseudo-Lipschitz of order 1 in the operator norm with constant $\tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}$. Suppose $P \in \mathcal{P}_{q+1}$. Then,

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) \leq C_{b,m} d_{\mathrm{pLip}_{1,q+1}}(Q, P),$$

where

$$C_{b,m} = (5 + 2^q)\left(\frac{\lambda_b C_1}{4} + \frac{\lambda_m C_2 D}{2} + C_1 D\tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}\right) + 4\phi_1(b)C_0.$$

In particular, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) \to 0$ if $d_{\mathcal{F}_{q+1}}(P, Q_n) \to 0$.

Proof. The proof proceeds as in the previous proposition and can be found in Appendix 6.A.4. □

Note that the two propositions above require stronger convergence requirements than that of $d_{\mathcal{F}_q}$.

6.4 Experiments

We conduct numerical experiments to examine the theory developed above. In the first two experiments, we investigate the behavior of the KSD using simple light-tailed and heavy-tailed target distributions. Then, we present a negative case study, where the KSD fails to detect discrepancies in moment estimates.

6.4.1 The Langevin KSD

Our first problem studies the behavior of Langevin KSD corresponding to the choice $a \equiv I$, $c \equiv 0$. This setting allows us to contrast with the IMQ kernel previously recommended by Gorham and Mackey [2017], which is known to control the bounded-Lipschitz metric.

6.4.1.1 Non-convergence in mean

We first consider a problem where a sequence does not converge in mean to their target and therefore not in $d_{\mathcal{F}_q}$ $(q = 1)$. We choose a target P and an approximating sequence $\{Q_1, Q_2, \ldots, \}$ as follows:

$$P = \mathcal{N}(-\bar{\mathbf{1}}, I),\ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1} \mathcal{N}\big((n+1)\bar{\mathbf{1}}, I\big),\ n \geq 1,$$

where $\bar{\mathbf{1}} = \mathbf{1}/\sqrt{D}$ and $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution over $\mathbb{R}^D$ with mean μ and covariance Σ. We take $D = 5$. While the sequence $\{Q_1, Q_2, \ldots, \}$ converges weakly to P, it does not converge in mean. Indeed, by construction, the approximating sequence has the following biased limit:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}_{Y\sim Q_n}[Y] &= \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) \mathbb{E}_{X\sim P}[X] + \frac{1}{n+1} \mathbb{E}_{Z\sim\mathcal{N}\left((n+1)\bar{\mathbf{1}}, I\right)}[Z] \\ &= \mathbb{E}_{X\sim P}[X] + \bar{\mathbf{1}}.\end{aligned}$$

We examine the kernel choice (6.5) recommended in Section 6.3.3. For the weight function $w_{1,\theta}$, we take $\theta = 0.5$. We use $v = 1$ for the linear kernel and the weight function. For the translation-invariant kernel Φ, we use the IMQ kernel $k_{\mathrm{IMQ}}(x, x') = (1 + \|x - x'\|_2^2)^{-1/2}$. We also consider the case $\theta = 0$.We compare these two choices against using k_{IMQ} alone.

We estimate the KSD using the U-statistic estimator. Specifically, we draw an i.i.d. sample of size $1,000$ from Q_n and compute a U-statistic; we repeat this process 30 times and take the average of the computed U-statistics.

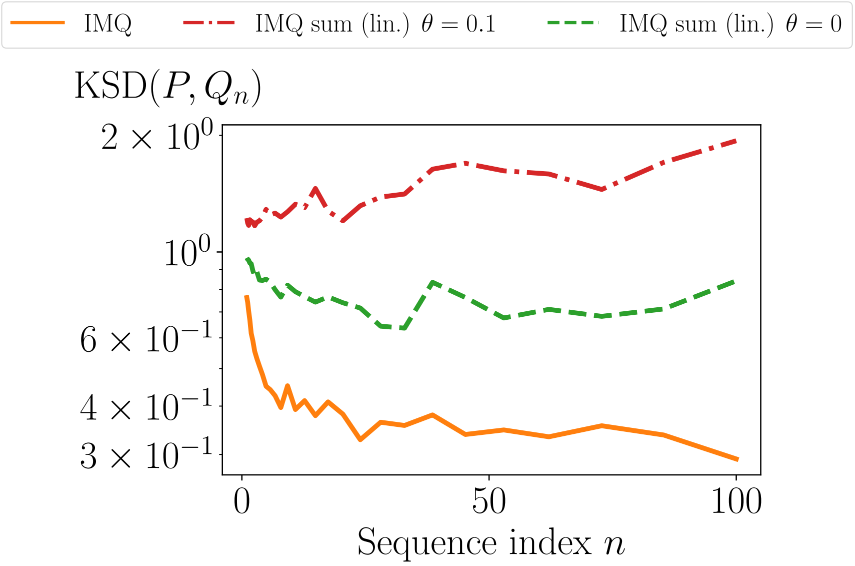


Figure 1: Comparisons of KSD with different kernels. Settings: (a) IMQ; the IMQ kernel k_{IMQ} (solid line), (b) IMQ sum (lin.) $\theta = 0$; the sum of the IMQ kernel and the normalized linear kernel (dashed line), (c) IMQ sum (lin.) $\theta = 0.1$; the sum of the IMQ kernel and the normalized linear kernel with tilting $\theta = 0.1$ (dash-dotted line).

Figure 1 shows the change of the KSD along the sequence for the three kernels. It can be seen that the KSD decreases with the IMQ kernel alone. For the other two choices, the KSD value does not decay; in particular, the kernel with the additional tilting $\theta = 0.1$ diverges as n increases. By design, our kernel choices induce functions growing linearly or super-linearly, and their KSD can therefore capture the non-convergence of the mean. Remarkably, although the case $\theta = 0$ is not guaranteed by our theory, the KSD does not decay to zero, implying possibility for theory improvement.

6.4.1.2 Non-convergence in variance

As in the previous section, we consider a case where a sequence does not converge in variance to their target and therefore not in $d_{\mathcal{F}_q}$ $(q = 2)$. We use the following target and approximating sequence:

$$P = \mathcal{N}(\mathbf{0}, I),\ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\big(\mathbf{0}, 2(n+1)I\big),\ n \geq 1.$$

As with the mean shift problem above, this example is constructed so that the sequence converges in distribution but not in variance, since the Gaussian in the second term always adds diagonal covariance $2I$.

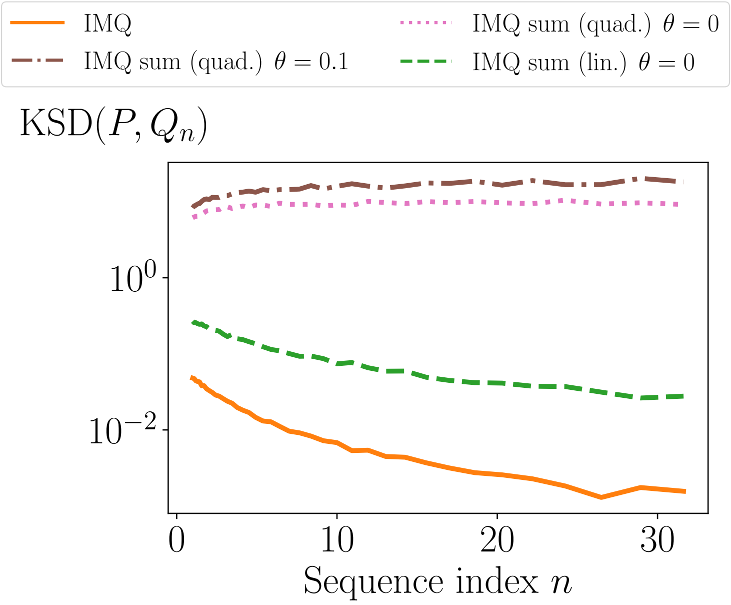


Figure 2: Comparisons of KSD with different kernels. Settings: (a) IMQ; the IMQ kernel k_{IMQ} (solid line), (b) IMQ sum (lin.) $\theta = 0$; the sum of the IMQ kernel and the normalized linear kernel $q = 1$ (dashed line), (c) IMQ sum (quad.) $\theta = 0$; the sum of the IMQ kernel and the normalized linear kernel with quadratic tilting $q = 2$ (dotted line), (d) IMQ sum (quad.) $\theta = 0.1$; The sum of the IMQ kernel and the normalized linear kernel with quadratic tilting $q = 2$ and additional reweighting $\theta = 0.1$ (dash-dotted line).

Figure (2) shows the KSD's transition along the approximating sequence for our four kernel choices. As in the previous experiment, the IMQ-KSD decreases; the mean-characterizing KSD also decays , confirming that having a function of linear growth is not sufficient to detect the non-convergence. For the two kernels yielding quadratically growing Stein-RKHS functions, we see a similar trend as in the previous experiment; in particular, the behavior of the KSD with $\theta = 0$ closely follows that of $\theta = 0.1$.

6.4.2 The DKSD and heavy-tailed distributions

In this section, we turn to heavy-tailed distribution to investigate the performance of the DKSD. Heavy-tailed distributions such as Student's *t*-distribution have bounded score functions, and it is known that the Langevin KSD fail to detect non-convergence for this target class [Gorham and Mackey, 2017, Theorem 10].

As our target P, we use the standard multivariate *t*-distribution with the degrees of freedom $\nu > 1$ defined by the density

$$p(x) = \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{D}{2}}\pi^{\frac{D}{2}}}\left(1 + \|x\|_2^2\right)^{-\frac{\nu+D}{2}}.$$

The *t*-distribution is uniformly dissipative with $\sigma(x) = \sqrt{1 + \nu^{-1}\|x\|_2^2}I$ and $\lambda_a = 4$. If $\nu > 2$, the diffusion is dissipative (Condition 6.3) with $\alpha = 1 - 2\nu^{-1}$ (see Lemma 6.38). With this choice, the DKSD is reduced to the Langevin KSD with the kernel tilted by $1 + \nu^{-1}\|x\|_2^2$.

According to Lemma 6.16, the allowed power of the weight is $q_w < (\alpha+1)/2 < 1$. As Proposition 6.13 requires uniform integrability of the higher moment, this bound on q_w implies that the DKSD can only be used for examining mean convergence. In the following, we take $\nu = 3$ and consider the same perturbation as in the previous section:

$$Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\big((n+1)\bar{\mathbf{1}}, I\big),\ n \geq 1.$$

We consider the quadratic growth case in Section 6.3.3. We compare the following two kernels: (a) $q = 0$ (uniform integrability of the first moment), (b) $q = 1$ (the second moment). The case $q = 1$ conforms to the requirement of Proposition 6.13, whereas the case $q = 0$ violates it. We also include the IMQ kernel as a baseline.

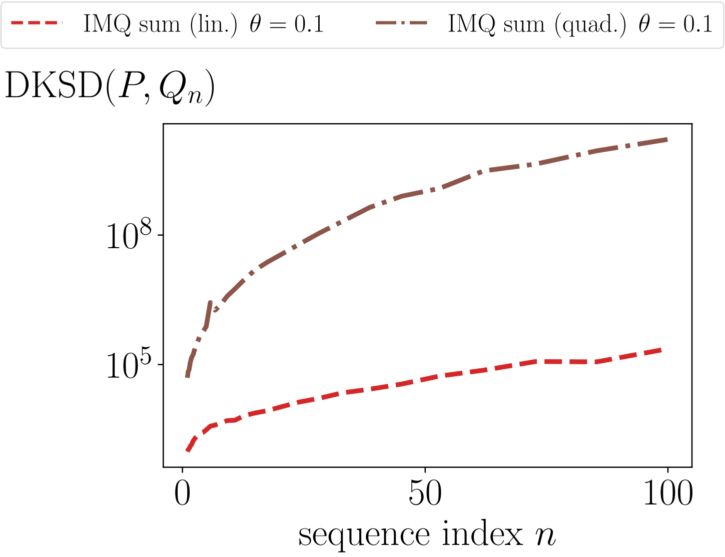


Figure 3: Comparisons of DKSD with different kernels in the t-distribution problem. Settings: (a) IMQ sum (lin.) $\theta = 0.1$; the sum of the IMQ kernel and the normalized linear kernel $q = 0$ (dashed line), (b) IMQ sum (quad.) $\theta = 0.1$; The sum of the IMQ kernel and the normalized linear kernel with quadratic tilting $q = 1$ and additional reweighting $\theta = 0.1$ (dash-dotted line).

Figure 3 demonstrates the result. The curve for the IMQ kernel is omitted, as its KSD decayed extremely fast and sometimes yielded negative values. While our theory requires the uniform integrability of the second moment, the linear-growth DKSD detects non-convergence; the same trend is observed for the quadratic-growth counterpart as expected. Recall that as the linear-growth DKSD characterizes the uniform integrability of the first moment, weak convergence control is sufficient for determining convergence in mean. The success of the linear-growth DKSD may therefore be attributed to its ability to control weak convergence; this feature should be proved without the extra uniform integrability condition, and we leave this task as future work. Finally, the stricter requirement may be considered as an artifact of the strong claim of Proposition 6.13, which deals with the uniform convergence under a class of pseudo-Lipschitz functions.

6.4.3 Failure mode: distribution mixtures with isolated components

Our final experiment concerns the following distributions:

$$P = \frac{1}{2}\mathcal{N}(\mu_1, I) + \frac{1}{2}\mathcal{N}(\mu_2, I),\ \tilde{Q}_n = r_n\mathcal{N}(\mu_1, I) + (1 - r_n)\mathcal{N}(\mu_2, I),$$

where $r_1 = 0.05$ and $r_n \nearrow 1/2$ as $n \to \infty$. The target P is supported by our theory, since the Gaussian mixtures are known to be distantly dissipative [Gorham et al., 2019, Example 3]. However, when the distance $\|\mu_1 - \mu_2\|_2$ between the two modes is large, the KSD is unable to capture the discrepancy of the mixture ratio r_n. Indeed, the Wasserstein rate $\rho_1(t) = 2e^{LR^2/8}e^{-rt/2}$ in Proposition 6.7 has an exponent depending on $R = \|\mu_1 - \mu_2\|_2$, and the diffusion therefore suffers from the slow convergence when R is large, rendering the KSD insensitive to this difference.

We detail our experimental procedure. In contrast to the previous experiments, rather than estimating the KSD between P and $\tilde{Q}_n$, we use a sample $\{x_i\}_{i=1}^{N}$from $\tilde{Q}_n$ to form a sequence of empirical distributions $Q_n = N^{-1}\sum_{i\leq N}\delta_{x_i}$. In this case, the KSD between P and Q_n is possible to compute exactly. This consideration is more relevant in practice since we are interested in how well a particular sample approximates P. In the following, we set $\mu_1 = -30\cdot\mathbf{1}$, $\mu_2 = -10\cdot\mathbf{1}$ and $D = 5$. The mixture ratio r_n is chosen from a regular grid of size 30 on the logarithmic scale with base 10. For each r_n, we draw 100 sample of size $N = 500$ and compute the KSD; we repeat this procedure 100 times and report the average KSD.

Our main question is whether the KSD can detect convergence in the first moment. It is clear that the mean of Q_n changes significantly as r_n increases. To this end, we consider the sum kernel in (6.5) with $q = 1$, ensuring that the Stein-modified RKHS has a function of linear growth. As $\{Q_1, Q_2 \ldots, \}$ has uniformly integrable first moments, we simply do not apply the additional tilting (i.e., $\theta = 0$). We examine two choices of the translation invariant kernel Φ : the IMQ kernel $k_{\mathrm{IMQ}}(x, x') = \left(1 + \|x - x'\|_2^2/\sigma^2\right)^{-1/2}$ and the Matérn kernel

$$k_{\mathrm{Mat}}(x, x') = \left(1 + \frac{\sqrt{3}\|x - x'\|_2}{\sigma}\right)\exp\left(-\frac{\sqrt{3}}{\sigma}\|x - x'\|_2\right).$$

First, we fix the bandwidth σ to 1 for both kernels. Figure 4 plots the KSD value against the mixture ratio with error bars representing standard deviations. For both kernels, their KSDs do not change along the sequence.

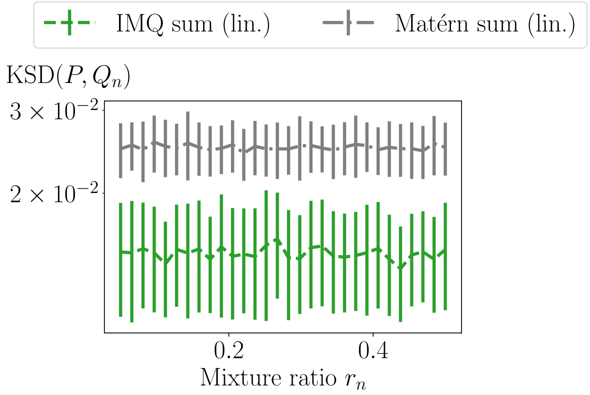
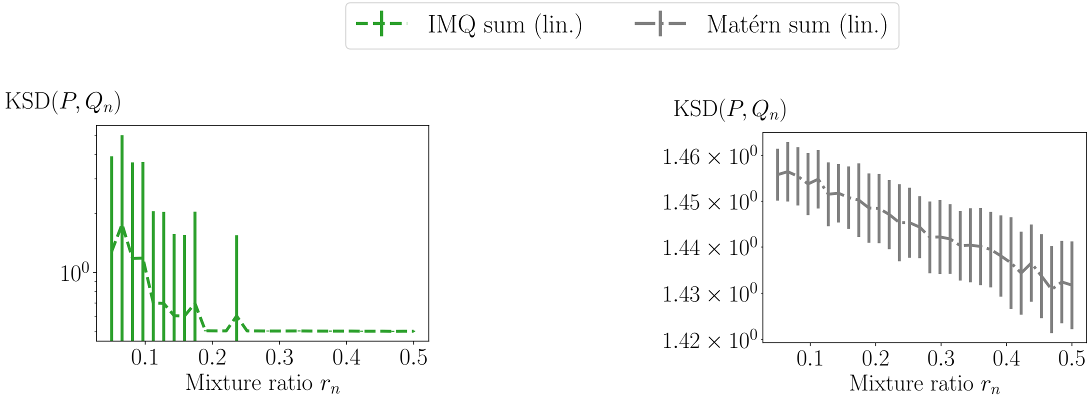


Figure 4: Comparison between the IMQ and the Matérn kernels in the mixture problem. Without bandwidth optimization. Error bars represent standard deviations.

This observation implies that the bandwidth choice may be suboptimal, making the KSD too weak to detect the change in the mixing proportion. Therefore, we next consider optimizing the bandwidth σ for each Q_n to improve the sensitivity. Following the approaches in non-parametric hypothesis testing [Gretton et al., 2012b, Sutherland et al., 2016, Jitkrittum et al., 2016, 2017b], we choose a bandwidth by optimizing the power of a test using the objective $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{kI})^2/\sqrt{v_{H_1}}$, where $v_{H_1} = \mathrm{Var}_{X\sim Q_n}\big[\mathbb{E}_{X'\sim Q_n}[h_p(X, X')]\big]$. This objective is a proxy of the power of the KSD goodness-of-fit tests [Chwialkowski et al., 2016, Liu et al., 2016] and can be computed exactly. The optimization procedure may be interpreted as seeking a bandwidth that allows us to conclude $P \neq Q_n$ using as few sample points as possible, without using the whole of Q_n.

Figure 5 shows the result. The IMQ kernel stays at the same value past the point $r_n = 0.3$, whereas the Matérn kernel shows a decreasing curve and captures the discrepancy. The Matérn kernel can therefore be thought of as more stable and inducing a stringent discrepancy measure.



(a) IMQ kernel

(b) Matérn kernel

Figure 5: Comparison between the IMQ and the Matérn kernels in the mixture problem. With optimized bandwidth for each kernel. Error bars represent standard deviations.

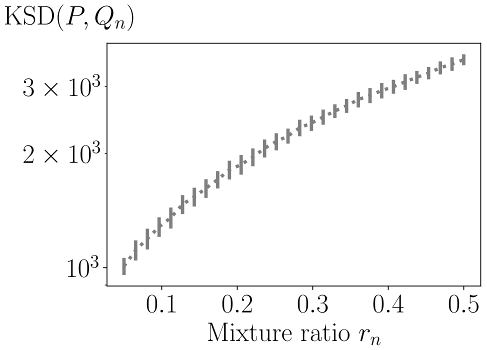


Figure 6: The KSD defined by the Matérn kernel in the mixture problem. The kernel's bandwidth is optimized for each r_n. Error bars represent standard deviations.

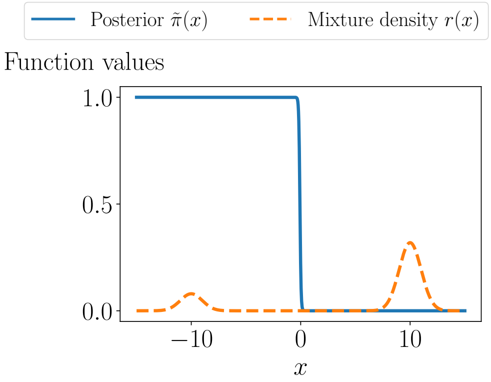


Figure 7: A two component Gaussian mixture and its posterior probability $\tilde{\pi}(x)$ of component assignment.

Although Figure 5b indicates that the optimization approach might be helpful, the following result casts doubt on this observation. We raise the growth rate of the Matérn sum kernel to $q = 2$ (and $\theta = 0$) in order to investigate the second moment convergence. Contrary to our expectation, Figure 6 demonstrates that the KSD grows as the mixture ratio approaches to the true value. We observed the same trend even without optimization. Note that since the Stein-modified RKHS contains quadratically growing functions, the KSD takes larger values with more samples observed around the mode $\mu_1 = -30 \cdot \mathbf{1}$ as $r_n \nearrow 1/2$. If the contribution of the mixture ratio is negligible, the change in the KSD value is therefore largely determined by the change of sampling locations.

We can attribute the KSD's weak dependence on the mixture ratio to the score function. To see this, consider a density

$$r_\pi(x) = \pi p_1(x) + (1 - \pi) p_2(x), 0 < \pi < 1.$$

The score function of r_π is given by

$$\mathbf{s}_{r,\pi}(x) = \tilde{\pi}(x) \mathbf{s}_{p_1}(x) + \big(1 - \tilde{\pi}(x)\big) \mathbf{s}_{p_2}(x), \text{ where } \tilde{\pi}(x) = \frac{\pi p_1(x)}{\pi p_1(x) + (1 - \pi) p_2(x)}.$$

The function $\tilde{\pi}(x)$ represents the posterior probability of an observation x arising from the mixture component p_1. We claim that for any two $\pi \neq \pi'$, the difference between $\mathbf{s}_{r,\pi}$ and $\mathbf{s}_{r,\pi'}$ is virtually absent. Our reasoning is as follows: $\mathbf{s}_{r,\pi}(x)$ becomes $\mathbf{s}_{p_1}$ (or $\mathbf{s}_{p_2}$) in the high-probability region of p_1 (or p_2); thus, the difference between two score functions is negligible under $r_{\pi'}$ for any other configuration π' so long as two components are concentrated in separate regions as in the Gaussian mixture example. This pathology is due to the following behavior of the posterior density $\tilde{\pi}(x)$: it becomes effectively a binary-valued function that outputs 1 in the high-density region of p_1 and 0 in the counterpart of $p_2(x)$. Indeed, if $p_1(x) \gg p_2(x)$, the posterior $\tilde{\pi}(x) \approx 1$ and is effectively independent of π. Figure 7 illustrates this situation for a Gaussian mixture in $D = 1$. Hence, in this case, varying the mixing proportion π does not modify the score function $\mathrm{s}_{r,\pi}$ significantly.

As the KSD between two densities depends on the difference of their score functions [Liu et al., 2016, Definition 3.2], the KSD is therefore insensitive to mismatches of the mixture ratios. Our experiments shows that kernel choice and optimization alone cannot solve this issue, calling for a more fundamental solution to this failure mode.

6.5 Conclusion

In this chapter, we have dealt with the question of the KSD's interpretability. We have shown that the KSD upper bounds the pseudo-Lipschitz metric $d_{\mathcal{F}_q}$, the worst-case expectation error over a class of polynomially growing function. A particular consequence of this result is that we may interpret the KSD in terms of convergence of moments. Our experiments confirm that the KSD with our proposed kernel choices indeed detect non-convergence in $d_{\mathcal{F}_q}$.

A theoretical shortcoming of our result is that we need a coercive function to characterize the uniform integrability (Lemma 6.15). Lemma 6.14 shows that we only need a function dominating a function of the form $x \mapsto \|x\|^q 1_{\{\|x\|>r\}}$. A related question to address is whether we can remove the uniform integrability of a higher order moment in Proposition 6.13. As heavy-tailed distributions have finitely many moments, assuming a higher-order moment can be restrictive.

Finally, we have seen a failure mode of the KSD in Section 6.4.3. Being unable to detect a mixture-ratio mismatch affects our inference significantly. We should therefore combine the KSD with other diagnostic tools if available. Given diverse applications of the KSD, it is desirable to develop a new computable discrepancy addressing this shortcoming.

6.A Proofs of main results

6.A.1 Characterization of pseudo-Lipschitz metrics

Proposition 6.20. *Let $P \in \mathcal{P}_q$ be a probability measure on $\mathbb{R}^D$ with a finite q-th moment with $q \geq 1$. For a sequence of probability measures $\{Q_1, Q_2, \ldots, \} \subset \mathcal{P}_q$, the following conditions are equivalent: (a) $d_{\mathcal{F}_q}(Q_n, P) \to 0$ as $n \to \infty$, and (b) as $n \to \infty$, the sequence Q_n converges weakly to P, and $\mathbb{E}_{X\sim Q_n}\left[\|X\|_2^q\right] \to \mathbb{E}_{X\sim P}\left[\|X\|_2^q\right]$.*

Proof. (a) $\Leftarrow$ (b) Our goal is to show that the following quantity can be made arbitrarily small by taking sufficiently large n :

$$\begin{aligned}\sup_{f\in\mathcal{F}_q}\left|\int f \mathrm{d}Q_n - \int f \mathrm{d}P\right| &= \sup_{f\in\mathcal{F}_q}\left|\int \big(f - f(0)\big)\,\mathrm{d}(Q_n - P) + \underbrace{\int f(0)\,\mathrm{d}(Q_n - P)}_{=0}\right| \\ &= \sup_{f\in\mathcal{F}_q}\left|\int \bar{f}\,\mathrm{d}(Q_n - P)\right|,\end{aligned}$$

where $\bar{f}$ denotes $f - f(0)$. To this end, we smoothly truncate a function $\bar{f}$ using the bump function $1_{R,1}$ of Lemma 6.37 with $r = R$ and $\delta = 1$ for some $R \geq 1$ (the function $1_{R,1}(x)$ vanishes if $\|x\|_2 > R + 1$). Specifically, we break up the integral on the RHS as

$$\sup_{f\in\mathcal{F}_q}\left|\int \bar{f}\,\mathrm{d}(Q_n - P)\right| = \sup_{f\in\mathcal{F}_q}\left|\int \bar{f}1_{R,1}\,\mathrm{d}(Q_n - P) + \int \bar{f}(1 - 1_{R,1})\,\mathrm{d}(Q_n - P)\right|$$

and evaluate each term below.

As a preparatory step, we clarify some properties of $\bar{f}$. Note that for any $f \in \mathcal{F}_q$, we have $|\bar{f}(x)| = |f(x) - f(0)| \leq (1 + \|x\|_2^{q-1})\|x\|_2$ for $x \in \mathbb{R}^D$. This implies that we have

$$\|\bar{f}1_{R,1}\|_\infty = \sup_{x\in\mathbb{R}}|\bar{f}(x)1_{R,1}(x)| < (R + 1) + (R + 1)^q =: C_{1,R}$$

and as $\bar{f}1_{R,1}(x)$ is everywhere differentiable,

$$\begin{aligned}\|\bar{f}1_{R,1}\|_{\mathrm{L}} = \sup_{x,y\in\mathbb{R}^D, x\neq y}&\frac{|\bar{f}1_{R,1}(x) - \bar{f}1_{R,1}(y)|}{\|x - y\|_2} \\ &\leq \sup_{x\in\mathbb{R}^D} 1_{R,1}(x)\|\nabla\bar{f}(x)\|_2 + \bar{f}(x)\|\nabla 1_{R,1}(x)\|_2 \\ &\leq \sqrt{D}\big\{1 + 2(R + 1)^{q-1}\big\} + 8e^{-1}\{1 + R + (R + 1)^q\} =: C_{2,R}\end{aligned}$$

Let $C_R = 2(C_{1,R} \vee C_{2,R})$ Then, $\bar{f}1_{R,1}/C_R$ belongs to the set of bounded Lipschitz functions $\mathrm{BL}_1 = \{f : \mathbb{R}^D \to \mathbb{R} : \|f\|_\infty + \|f\|_{\mathrm{L}} \leq 1\}$.

We are ready to bound the quantify of interest. For $\varepsilon > 0$, using the weak convergence assumption, take n large enough so that

$$d_{\mathrm{BL}_1}(P, Q_n) := \sup_{f \in \mathrm{BL}_1} \left| \int f \mathrm{d}Q_n - \int f \mathrm{d}P) \right| < \frac{\varepsilon}{2C_R},$$

which is possible as the bounded Lipschitz metric d_{BL_1} metrizes weak convergence [Dudley, 2002, Section 11.3]. The definition of R has been left unspecified; here, we take R such that

$$\int_{\{\|x\|_2 > R\}} \|x\|_2^q \mathrm{d}P(x) \vee \sup_{n \geq 1} \int_{\{\|x\|_2 > R\}} \|x\|_2^q \mathrm{d}Q_n(x) < \frac{\varepsilon}{8}.$$

The existence of R is guaranteed by Lemma 6.32: for a weakly converging sequence of probability measures $\{Q_1, Q_2, \dots\}$, the convergence in the q-th moment is equivalent to the q-th moment uniform integrability. Then,

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_q} \left| \int \bar{f}\, \mathrm{d}(Q_n - P) \right| \\
&\leq \sup_{f \in \mathcal{F}_q} C_R \left| \int C_R^{-1} \cdot \bar{f} 1_{R,1}\, \mathrm{d}(Q_n - P) \right| + \sup_{f \in \mathcal{F}_q} \left| \int \bar{f}(1 - 1_{R,1})\, \mathrm{d}(Q_n - P) \right| \\
&\leq C_R \sup_{f \in \mathrm{BL}_1} \left| \int f\, \mathrm{d}(Q_n - P) \right| + \sup_{f \in \mathcal{F}_q} \left| \int \bar{f}(1 - 1_{R,1}) \mathrm{d}Q_n \right| + \sup_{f \in \mathcal{F}_q} \left| \int \bar{f}(1 - 1_{R,1}) \mathrm{d}P \right| \\
&\leq C_R \sup_{f \in \mathrm{BL}_1} \left| \int f\, \mathrm{d}(Q_n - P) \right| + \sup_{f \in \mathcal{F}_q} \int_{\{\|x\|_2 > R\}} |\bar{f}| \mathrm{d}Q_n + \sup_{f \in \mathcal{F}_q} \int_{\{\|x\|_2 > R\}} |\bar{f}| \mathrm{d}P \\
&\leq C_R \cdot \frac{\varepsilon}{2C_R} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.
\end{aligned}$$

The third line follows from the bounded Lipschitzness of $\bar{f} 1_{R,1} / C_R$ and the triangle inequality; the fourth line is another application of the triangle inequality and $1 - 1_{R,1}(x) \leq 1\{\|x\|_2 > R\}$; the fifth line is true because $|\bar{f}(x)| = |f(x) - f(0)| \leq (1 + \|x\|_2^{q-1}) \|x\|_2$ for any $x \in \mathbb{R}^D$, and because of the definition of R.

(a) $\Rightarrow$ (b) We first prove that convergence in $d_{\mathcal{F}_q}$ implies weak convergence. For any $\varepsilon > 0$ and $f \in \mathrm{BL}_1$, define the Gaussian convolution $\tilde{f}_\varepsilon(x) = \mathbb{E}_G[f(x - \varepsilon G)]$ with G is a standard Gaussian random vector. The function is infinitely differentiable and satisfies, by the Lipschitzness of f,

$$\sup_{x \in \mathbb{R}^D} |f(x) - \tilde{f}_\varepsilon(x)| \leq \varepsilon \mathbb{E}_G \left[\|G\|_2 \right].$$

By Lemma 6.30, we have constant bounds on the operator norms of the derivatives up to the third order; the bounding constants depend on ε. With C_ε the maximum of the bounding constants, we $C_\varepsilon^{-1} \tilde{f}_\varepsilon \in \mathcal{F}_q$. Then,

$$\begin{aligned}
&d_{\mathrm{BL}_1}(Q_n, P) = \sup_{f \in \mathrm{BL}_1} \left| \int f \mathrm{d}Q_n - \int f \mathrm{d}P) \right| \\
&\leq \sup_{f \in \mathrm{BL}_1} \int |f - \tilde{f}_\varepsilon| \mathrm{d}Q_n + \int |f - \tilde{f}_\varepsilon| \mathrm{d}P + C_\varepsilon \left| \int C_\varepsilon^{-1} \tilde{f}_\varepsilon\, \mathrm{d}(Q_n - P) \right| \\
&\leq 2\varepsilon \mathbb{E}_G \left[\|G\|_2 \right] + C_\varepsilon d_{\mathcal{F}_q}(P, Q_n).
\end{aligned}$$

Taking the successive limits of n, ε shows that $d_{\mathrm{BL}_1}(Q_n, P) \to 0$, implying the weak convergence of the sequence.

Next, we show that convergence in $d_{\mathcal{F}_q}$ implies $\mathbb{E}_{X\sim Q_n}\|X\|_2^q \to \mathbb{E}_{X\sim P}\|X\|_2^q$. As the weak convergence has been established above, we only need to show that the sequence $\{Q_1, Q_2, \dots, \}$ has uniformly integrable q-th moments. As $P \in \mathcal{P}_q$, we can take R satisfying

$$\int_{\|x\|_2 > R} \|x\|_2^q \mathrm{d}P(x) \le \varepsilon.$$

Consider the function $f_{q,R}(x) = \|x\|_2^q(1 - 1_{R,1}(x))$ with the smooth bump function $1_{R,1}$ of Lemma 6.37. This function has derivatives, up to the third order, growing in the order of $\|x\|_2^{q-1}$, and therefore with a proper scaling $C_R > 0$ (depending on R), we have $\tilde{f}_{q,R} \coloneqq C_R f_{q,R} \in \mathcal{F}_q$. By the convergence in $d_{\mathcal{F}_q}$, for any $\varepsilon > 0$ we can take N such that for any $n > N$,

$$\int \tilde{f}_{q,R} \mathrm{d}Q_n \le \int \tilde{f}_{q,R} \mathrm{d}P + C_R \varepsilon.$$

These choices yield,

$$\begin{aligned}\int_{\|x\|_2 > R+1} \|x\|_2^q \mathrm{d}Q_n(x) &\le \int f_{q,R} \mathrm{d}Q_n \\ &\le C_R^{-1}\left(\int \tilde{f}_{q,R} \mathrm{d}P + C_R\varepsilon\right) \\ &\le \int_{\|x\|_2 > R+1} \|x\|_2^q \mathrm{d}P + \varepsilon \le 2\varepsilon.\end{aligned}$$

Thus, we have arrived at the desired conclusion $\lim_{r\to\infty} \limsup_{n\to\infty} \int_{\|x\|_2 > r} \|x\|_2^q \mathrm{d}Q_n(x) = 0$. □

Corollary 6.21. *For $q \ge 1$, let $d_{\mathrm{pLip}_{1,q-1}}(P, Q)$ be the IPM defined by $\mathrm{pLip}_{1,q-1}$, the set of functions that are pseudo-Lipschitz of order $q-1$ with its pseudo-Lipschitz constant bounded by 1. For a sequence of probability measures $\{Q_1, Q_2, \dots, \} \subset \mathcal{P}_q$, the following conditions are equivalent: (a) $d_{\mathrm{pLip}_{1,q-1}}(Q_n, P) \to 0$ as $n \to \infty$, and (b) as $n \to \infty$, the sequence Q_n converges weakly to P, and $\mathbb{E}_{X\sim Q_n}\|X\|_2^q \to \mathbb{E}_{X\sim P}\|X\|_2^q$.*

Proof. The direction(a) $\Rightarrow$ (b) results from Proposition 6.10, as $d_{\mathcal{F}_q}(P, Q) \le d_{\mathrm{pLip}_{1,q-1}}(P, Q)$. The other direction (b) $\Rightarrow$ (a) can be shown as in the proof of Proposition 6.10. □

6.A.2 Uniform integrability and a DKSD lower bound

Our goal is to show a KSD bound on the IPM $d_{\mathcal{F}_q}(P, Q)$.

Let g be a solution to the Stein equation $\mathcal{T}_P g = f - \mathbb{E}_{X\sim P}[f(X)]$ for $f \in \mathcal{F}_q$. We approximate $\mathcal{T}_p g$ by mollification. Specifically, we decompose the function $\mathcal{T}_P g$ into three parts:

$$\mathcal{T}_P g = \underbrace{\mathcal{T}_P g - \mathcal{T}_P g_{\mathrm{trunc}}}_{\text{Step 1: truncation}} + \underbrace{\mathcal{T}_P g_{\mathrm{trunc}} - \mathcal{T}_P g_{\mathrm{RKHS}}}_{\text{Step 2: approximation}} + \underbrace{\|g_{\mathrm{RKHS}}\|}_{\text{Step 3: measuring RKHS norm}} \frac{\mathcal{T}_P g_{\mathrm{RKHS}}}{\|g_{\mathrm{RKHS}}\|}.$$

We elaborate on each step as follows.

Step 1: Smoothly truncating g. Let us start with the following notion.

Definition 6.22 (Integrability rate of order q)**.** For any probability measure Q and $\epsilon > 0$, the integrability rate $R_q(Q, \varepsilon)$ of order q is defined as

$$R_q(Q, \varepsilon) := \inf \left\{ r \geq 1 : \int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q(x) \leq \varepsilon \right\},$$

where we use the convention $\inf \emptyset = \infty$.

Note that $Q \in \mathcal{P}_q$ is equivalent to having $R_q(Q, \varepsilon) < \infty$ for each $\varepsilon > 0$. In the following, for each $\varepsilon > 0$ and a probability measure Q, we consider the integrability rate $R = R_{q+q_a}(Q, \varepsilon)$. Let where $1_{R,1}$ be a smooth bump function from Lemma 6.37, which vanishes outside the centered Euclidean ball of radius $R + 1$. For a function $g \in \mathcal{G}$, we consider a truncated version $g_{R,1} := 1_{R,1} \cdot g$. The truncation results yields the following error estimate:

Lemma 6.23. *Let $Q \in \mathcal{P}_{q+q_a}$ a probability measure and $R = R_{q+q_a}(Q, \varepsilon)$. For each $\varepsilon > 0$, we have*

$$\left| \int \mathcal{T}_P g \mathrm{d}Q - \int \mathcal{T}_P g_{R,1} \mathrm{d}Q \right| \leq c_{P,D} \cdot \varepsilon$$

with $c_{P,D} = 2\sqrt{D}\zeta_1 \left(\lambda_b + 8e^{-1}\right) + D\lambda_m \zeta_2$.

Proof. Observe that for each $x \in \mathbb{R}^D$,

$$\begin{aligned}
&|\mathcal{T}_P g(x) - \mathcal{T}_P g_{R,1}(x)| \\
&\leq |\langle 2b(x), g(x)(1 - 1_{R,1})(x) \rangle| + |\langle m(x), \nabla g(1 - 1_{R,1})(x) \rangle| \\
&\leq 2\|b(x)\|_2 \|g(x)\|_2 |(1 - 1_{R,1})(x)| + D|1 - 1_{R,1}(x)| \|m(x)\|_{\mathrm{op}} \|\nabla g(x)\|_{\mathrm{op}} \\
&\quad + \|\nabla 1_{R,1}(x)\|_2 \|g(x)\|_2 \\
&\leq 1\{\|x\|_2 > R\} \left[\left\{ \frac{\lambda_b}{2}(1 + \|x\|_2) + 8e^{-1} \right\} \cdot D\zeta_1 + D\zeta_2 \lambda_m \big(1 + \|x\|_2^{q_a+1}\big) \right] (1 + \|x\|_2^{q-1}).
\end{aligned}$$

By the definition of R, we have

$$\int_{\{\|x\|_2 > R\}} \|x\|_2^{q'} \mathrm{d}Q \leq \int_{\{\|x\|_2 > R\}} \|x\|_2^{q+q_a} \mathrm{d}Q \leq \varepsilon.$$

for any $0 \leq q' < q + q_a$. Thus,

$$\begin{aligned}
&\left| \int \mathcal{T}_P g \mathrm{d}Q - \int \mathcal{T}_P g_{R,1} \mathrm{d}Q \right| \\
&\leq \int_{\|x\|_2 > R} \left[\left\{ \frac{\lambda_b}{2}(1 + \|x\|_2) + 8e^{-1} \right\} \cdot \sqrt{D}\zeta_1 (1 + \|x\|_2^{q-1}) \right. \\
&\qquad \left. + D\zeta_2 \lambda_m \big(1 + \|x\|_2^{q_a+1}\big)\big(1 + \|x\|_2^{q-1}\big) \right] \mathrm{d}Q(x)
\end{aligned}$$

$$\leq \left\{2\sqrt{D}\zeta_1\left(\lambda_b + 8e^{-1}\right) + D\lambda_m\zeta_2\right\}\varepsilon.$$

□

Step 2: Constructing a smooth approximation to $g_{R,1}$. We consider approximating the function $g_{R,1}$ with an RKHS function. For later use, we consider factorizing $g_{R,1}$ using a differentiable positive function $w(x) : \mathbb{R}^D \to [1, \infty)$; i.e.,

$$g_{R,1}(x) = w(x)g^w_{R,1},$$

where $g^w_{R,1} := g_{R,1}/w$. We assume that $\sup_{x\in\mathbb{R}^D} \|\nabla \log w(x)\|_2 =: M_w < \infty$. We also define a helper function

$$B_w(z) := \sup_{x\in\mathbb{R}^D, u\in[0,1]} \frac{w(x)}{w(x-uz)}.$$

The form of w will be specified below, which will be of the form $w(x) = \left(v^2 + \|x\|_2^2\right)^{q_w}$; for this w, we have $B_w(z) = O(\|z\|_2^{q_w})$. Note that $0 < \rho \leq 1$, $B_w(\rho z) \leq B_w(z)$.

For fixed $\rho > 0$, we define a smooth approximation g^w_ρ to by convolution,

$$g^w_\rho(x) := \mathbb{E}_Z\left[g^w_{R,1}(x - \rho Z)\right],$$

where Z is a $\mathbb{R}^D$-valued random variable with $\mathbb{E}\big[\|Z\|_2 B_w(Z)\big] < \infty$ (its law will be specified in the sequel). The following result quantifies the approximation error and mirrors the proof of Lemma 12 of Gorham and Mackey [2017]; here, we do not assume the Lipschitzness of the drift b (assuming the Lipschitzness will improve the D-dependency of the bound).

Lemma 6.24. *Let $g^w_{R,1} := g_{R,1}/w$,and $g^w_\rho(x) := \mathbb{E}_Z\left[g^w_{R,1}(x - \rho Z)\right]$. For each fixed $\rho \in (0, 1]$, the approximation g^w_ρ satisfies the following:*

$$\begin{aligned}&\left|\int \mathcal{T}_P(wg^w_\rho)\mathrm{d}Q - \int \mathcal{T}_P g_{R,1}\mathrm{d}Q\right| \\ &\leq \rho \cdot U_{P,D,w} \cdot (1 + R + 2\varepsilon) \cdot \{1 + (R+1)^{q-1}\}\end{aligned}$$

with

$$U_{P,D,w} = 4\left\{(2\lambda_b + M_w D\lambda_m) \cdot \tilde{u}^{(1)}_{P,D,w} + D\lambda_m \tilde{u}^{(2)}_{P,D,w}\right\}$$

where $\tilde{u}^{(1)}_{P,D,w}$ and $\tilde{u}^{(2)}_{P,D,w}$ are constants given respectively in Lemmas 6.39, 6.40 with $\delta = 1$.

Proof. By Lemmas 6.39, 6.40, for each $x \in \mathbb{R}^D$, we have

$$\begin{aligned}&|\mathcal{T}_P w g^w_\rho(x) - \mathcal{T}_P w g^w_{R,1}(x)| \\ &\leq |2w(x)\langle b(x), g^w_\rho(x) - g^w_{R,1}(x)\rangle| + |w(x)\langle m(x), \nabla g^w_\rho(x) - \nabla g^w_{R,1}(x)\rangle| \\ &\qquad + w(x)|\langle m(x), \nabla \log w(x) \otimes (g^w_\rho(x) - g^w_{K,\delta}(x))\rangle| \\ &\leq w(x)\{2\|b(x)\|_2 + \|m(x)\|_{\mathrm{op}}\|\nabla \log w(x)\|_2\}\|g^w_\rho(x) - g^w_{K,\delta}(x)\|_2 \\ &\quad + Dw(x)\|m(x)\|_{\mathrm{op}}\|\nabla g^w_\rho(x) - \nabla g^w_{K,\delta}(x)\|_{\mathrm{op}}\end{aligned}$$

$$\le \rho \cdot \{1 + (R+1)^{q-1}\} \left[\left\{ \frac{\lambda_b}{2}(1 + \|x\|_2) + M_w \lambda_m (1 + \|x\|_2^{q_a+1}) \right\} \tilde{u}^{(1)}_{P,D,w} \right.$$
$$\left. + D\lambda_m \cdot (1 + \|x\|_2^{q_a+1}) \tilde{u}^{(2)}_{P,D,w} \right]$$

Note that

$$\int (1 + \|x\|_2^{q_a+1}) \mathrm{d}Q(x) \le 1 + R^{q_a+1} + \int_{\{\|x\|_2 > R\}} (1 + \|x\|_2^{q_a+1}) \mathrm{d}Q(x)$$
$$\le 1 + R^{q_a+1} + 2\varepsilon.$$

As a consequence,

$$\left| \int \mathcal{T}_P w g^w_\rho \mathrm{d}Q - \int \mathcal{T}_P g^w_{R,1} \mathrm{d}Q \right|$$
$$\le \rho(1 + R^{q_a+1} + 2\varepsilon) \left[\left\{ \frac{\lambda_b}{2} + M_w \lambda_m \right\} \tilde{u}^{(1)}_{P,D,w} + D\lambda_m \tilde{u}^{(2)}_{P,D,w} \right] \{1 + (R+1)^{q-1}\}.$$

□

Step 3: Measuring the RKHS norm of wg^w_ρ. The RKHS norm of wg^w_ρ is derived once we specify the convolution distribution and the RKHS kernel function. The former will be specified below; for the latter, we assume that the scalar kernel is given by a weighted kernel

$$k(x, x') = w(x) w(x') \tilde{k}(x, x'),$$

where $\tilde{k}(x, x')$ is another positive definite kernel. This choice yields

$$\|wg^w_\rho\|_{\mathcal{G}_{kI}} = \sqrt{\sum_{d=1}^{D} \left\| w(g^w_\rho)_d \right\|^2_{\mathcal{G}_k}} = \sqrt{\sum_{d=1}^{D} \left\| (g^w_\rho)_d \right\|^2_{\mathcal{G}_{\tilde{k}}}},$$

where $(g^w_\rho)_d$ is the dth component of g^w_ρ. Thus, we only need to check the $\mathcal{G}_{\tilde{k}}$ norm of each coordinate function of g^w_ρ.

We consider the tilting function $w(x) = (v^2 + \|x\|_2^2)^{q_w}$ with $v \ge 1$ and $q_w \in [0, \infty)$. We specify the law of the convolution variable Z in 6.A.2 (i.e., the convolution kernel) with the sinc density from Lemma 6.34

$$s_{\tilde{q}_w}(x) = \prod_{d=1}^{D} \frac{\mathrm{sinc}^{4\tilde{q}_w}(x_d)}{S_{\tilde{q}_w}},$$

where $S_{\tilde{q}_w} = \int_{-\infty}^{\infty} \mathrm{sinc}^{4\tilde{q}_w}(x) \mathrm{d}x$, $\tilde{q}_w = \lceil q_w \rceil + 1$ with the symbol$\lceil x \rceil$ denoting the smallest integer greater than or equal to x. Lemma 6.34 guarantees that $s_{\tilde{q}_w}$ is well-defined and has a finite $\tilde{q}_w$-th moment. Note that by the convolution theorem and $\|s_{\tilde{q}_w}\|_{L^1} = 1$, the Fourier transform $\hat{s}_{\tilde{q}_w}$ of $s_{\tilde{q}_w}$ satisfies

$$|\hat{s}_{\tilde{q}_w}(\omega)|^2 \le (2\pi)^{-D} 1\{\|\omega\|_\infty \le 4\tilde{q}_w\} \text{ for } \omega \in \mathbb{R}^D,$$

where $\|\omega\|_\infty = \max_{d=1,\ldots,D} |\omega_i|$.

Lemma 6.25. *Let $g_{R,1} = g \cdot 1_{R,1}$ with $1_{R,1}$ a smooth bump function from Lemma 6.37. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w}$ with $v \geq 1$ and $q_w > 0$. Let $\Phi_w(x, x') = w(x)w(x')\Phi(x - x')$ where $\Phi \in C^2$ is a positive definite function with non-vanishing generalized Fourier transform $\hat{\Phi}$. Define the convolution g_ρ^w of $g_{R,1}^w = g_{R,1}/w$ by*

$$g_\rho^w(x) \coloneqq \mathbb{E}_Z \left[g_{R,1}^w(x - \rho Z) \right],$$

where Z has the law with density $s_{\tilde{q}_w}$ from Lemma 6.34 with $\tilde{q}_w = \lceil q_w \rceil + 1$. Let $k = \Phi_w$. Then, the RKHS norm $\|wg_\rho^w\|_{\mathcal{G}_{kI}}$ is evaluated as

$$\|wg_\rho^w\|_{\mathcal{G}_{kI}} \leq (2\pi)^{-D/4} D\zeta_1 \sqrt{\sup_{\|\omega\|_\infty \leq 4\tilde{q}_w/\rho} \hat{\Phi}(\omega)^{-1}} \cdot \sqrt{\mathrm{Vol}(\mathcal{B}_{R+1})} \cdot (1 + (R+1)^{q - 2q_w - 1}),$$

where $\mathrm{Vol}(\mathcal{B}_{R+1})$ is the volume of the Euclidean ball of radius $R + 1$. In particular, the same conclusion holds if the kernel k is given by $k = \Phi_w + \ell$, where ℓ is another positive definite kernel.

Proof. We first address the case $k(x, x') = \Phi_w(x, x')$. As $\left(g_\rho^w\right)_d$ is given by the convolution with $s(x) = \rho^{-D} s_{\tilde{q}_w}(x/\rho)$, by the convolution theorem [Wendland, 2004, Theorem 5.16], its Fourier transform of is expressed by the product

$$(2\pi)^{D/2} \widehat{(g_{R,1}^w)}_d(\omega) \hat{s}(\omega) = (2\pi)^{D/2} \widehat{(g_{R,1}^w)}_d(\omega) \hat{s}_{\tilde{q}_w}(\rho\omega),$$

(note that we are using the definition of Fourier transform in [Wendland, 2004, Definition 5.15]). Thus, by [Wendland, 2004, Theorem 10.21], the RKHS norm of $(g_\rho^w)_d$ is given by

$$\begin{aligned}
\left\|(g_\rho^w)_d\right\|_{\mathcal{G}_{\Phi_w}}^2 &= (2\pi)^{-D/2} \int \frac{|\widehat{(g_\rho^w)_d}(\omega)|^2}{\hat{\Phi}(\omega)} \mathrm{d}\omega = (2\pi)^{D/2} \int \frac{|\widehat{(g_{R,1}^w)}_d(\omega) \hat{s}_{\tilde{q}_w}(\rho\omega)|^2}{\hat{\Phi}(\omega)} \mathrm{d}\omega \\
&\leq (2\pi)^{-D/2} \sup_{\|\omega\|_\infty \leq 4\tilde{q}_w/\rho} \hat{\Phi}(\omega)^{-1} \cdot \|\widehat{(g_{R,1}^w)}_d\|_{L^2}^2 \\
&= (2\pi)^{-D/2} \sup_{\|\omega\|_\infty \leq 4\tilde{q}_w/\rho} \hat{\Phi}(\omega)^{-1} \cdot \|(g_{R,1}^w)_d\|_{L^2}^2,
\end{aligned}$$

where the last equality is a result of the Plancherel theorem, since $(g_{R,1}^w)_d \in L^1 \cap L^2$. One can check $\|(g_{R,1}^w)_d\|_{L^2} \leq \sqrt{\mathrm{Vol}(\mathcal{B}_{R+1})} \cdot \sup_{x \in \mathbb{R}^D} |(g_{R,1}^w)_d|$ and

$$\begin{aligned}
\sup_{x \in \mathbb{R}^D} |(g_{R,1}^w)_d| &\leq \sqrt{D}\zeta_1 \sup_{x \in \mathbb{R}^D} 1_{R,1}(x) \left\{ 1 + \frac{\|x\|_2^{q-1}}{(v^2 + \|x\|_2^2)^{q_w}} \right\} \\
&\leq \sqrt{D}\zeta_1 (1 + (R+1)^{q - 2q_w - 1}),
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\|wg_\rho^w\|_{\mathcal{G}_{kI}} \\
&= \sqrt{\sum_{d=1}^{D} \left\|w(g_\rho^w)_d\right\|_{\mathcal{G}_k}^2} = \sqrt{\sum_{d=1}^{D} \left\|(g_\rho^w)_d\right\|_{\mathcal{G}_{\Phi_w}}^2}
\end{aligned}$$

$$\leq (2\pi)^{-D/4} D\zeta_1 \sqrt{\sup_{\|\omega\|_\infty \leq 4\tilde{q}_w/\rho} \hat{\Phi}(\omega)^{-1}} \cdot \sqrt{\mathrm{Vol}(\mathcal{B}_{R+1})} \cdot (1 + (R+1)^{q-2q_w-1}).$$

We next deal with the case $k = \Phi_w + \ell$. According to Aronszajn [1950], because k is the sum of two kernels, the RKHS norm of $\mathcal{G}_k$ of a (scalar-valued) function $g \in \mathcal{G}_k$ is given by

$$\|g\|_{\mathcal{G}_k} = \inf\{\|g_1\|_{\mathcal{G}_{\Phi_w}} + \|g_2\|_{\mathcal{G}_\ell} : g = g_1 + g_2, g_1 \in \mathcal{G}_{\Phi_w}, g_2 \in \mathcal{G}_\ell\}.$$

As we have shown that each component of wg_ρ^w is an element of $\mathcal{G}_{\Phi_w}$, we have $\|(wg_\rho^w)_d\|_{\mathcal{G}_k} \leq \|(wg_\rho^w)_d\|_{\mathcal{G}_{\Phi_w}}$. The rest of the proof follows from the definition of the vector-valued RKHS $\mathcal{G}_{kI}$. □

The following proposition summarizes the above three steps.

Proposition 6.26. *Let $\varepsilon > 0$ be an arbitrary fixed positive real number. Let $\rho \in (0, 1]$ be a fixed number. Let $\Phi \in C^2$ be a positive definite function with non-vanishing generalized Fourier transform $\hat{\Phi}$. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w}$ with real numbers $v \geq 1$ and $q_w \geq 0$. Let $\mathcal{G}_{kI}$ be the RKHS defined by kernel kI with $k(x, y) = \Phi_w(x, x') + \tilde{\ell}(x, x')$ where $\Phi_w(x, x') = w(x)w(x')\Phi(x - x')$. Let Q be a probability measure in $\mathcal{P}_{q+q_a}$ and $R := R_{q+q_a}(Q, \varepsilon/c_{P,D})$. Then, we have*

$$\begin{aligned}
&d_{\mathcal{F}_q}(P, Q) \\
&\leq \varepsilon + \rho U_{P,D,w} \cdot (1 + R^{q_a+1} + 2\varepsilon) \cdot \{1 + (R+1)^{q-1}\} \\
&\quad + (2\pi)^{-\frac{D}{4}} D\zeta_1 \sqrt{\mathrm{Vol}\big(\mathcal{B}_1(\mathbb{R}^D)\big)(R+1)^D} \cdot (1 + (R+1)^{q-2q_w-1}) F_\Phi(4\tilde{q}_w\rho^{-1}) \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}),
\end{aligned}$$

where $\mathrm{Vol}\big(\mathcal{B}_1(\mathbb{R}^D)\big)$ denotes the volume of the unit ball in the Euclidean space $\mathbb{R}^D$, $F_\Phi(t) = \sup_{\|\omega\|_\infty \leq t} \hat{\Phi}(\omega)^{-1}$, and $\tilde{q}_w = \lceil q_w \rceil + 1$. Therefore, for a sequence of measures $\{Q_1, Q_2, \ldots, \}$ in $\mathcal{P}_{q+q_a}$ with uniformly integrable $(q + q_a)$-th moments, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{kI}) \to 0$ only if $d_{\mathcal{F}_q}(P, Q_n) \to 0$.

Proof. Using the results obtained so far, we have for a function $f \in \mathcal{F}_q$,

$$\begin{aligned}
&|\mathbb{E}_{Y\sim Q} f(Y) - \mathbb{E}_{X\sim P}[f(X)]| = |\mathbb{E}_{Y\sim Q} \mathcal{T}_P g(Y)| \\
&\leq |\mathbb{E}_{Y\sim Q}[\mathcal{T}_P g(Y) - \mathcal{T}_P g_{R,1}(Y)]| \\
&\qquad + |\mathbb{E}_{Y\sim Q}[\mathcal{T}_P g_{R,1}(Y) - \mathcal{T}_P w g_\rho^w(Y)]| + |\mathbb{E}_{Y\sim Q} \mathcal{T}_P w g_\rho^w(Y)| \\
&\leq |\mathbb{E}_{Y\sim Q}[\mathcal{T}_P g(Y) - \mathcal{T}_P g_{R,1}(Y)]| \\
&\quad + |\mathbb{E}_{Y\sim Q}[\mathcal{T}_P g_{R,1}(Y) - \mathcal{T}_P w g_\rho^w(Y)]| + \left\|w g_\rho^w\right\|_{\mathcal{G}_{kI}} \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) \\
&\leq \varepsilon + \rho U_{P,D,w}(1 + R^2 + 2\varepsilon) \cdot \{1 + (R+1)^{q-1}\} \\
&\qquad + (2\pi)^{-D/4} D\zeta_1 F_\Phi(4\tilde{q}_w\rho^{-1}) \sqrt{\mathrm{Vol}\big(\mathcal{B}_1(\mathbb{R}^D)\big)(R+1)^D} \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_k).
\end{aligned}$$

Taking the supremum over $\mathcal{F}_q$ completes the proof of the first claim.

For the second claim, if the sequence has uniformly integrable q-th moments, we can take finite $r \geq 1$ such that $R_{q+q_a}(Q_n, \varepsilon) \leq r$ for all $n \geq 1$. Thus, if $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{kI}) \to 0$, taking successive limits of n, ρ, ϵ shows $d_{\mathcal{F}_q}(P, Q_n) \to 0$. □

6.A.3 The diffusion Stein operator and zero-mean functions

Proposition 6.27 (The diffusion Stein operator generates zero-mean functions)**.** *Let $q_a \in \{0, 1\}$ be the additional growth exponent of $\|a(x)\|_{\mathrm{op}}$ from Condition 6.2. If $q_a = 0$, assume P has a finite q-th moment; if $q_a = 1$, a finite $(q+1)$-th moment. Let $g \in C^1$ be a function with the following growth conditions:*

$$\|g(x)\|_2 \leq C_0(1 + \|x\|_2^{q-1}),$$
$$\|\nabla g(x)\|_{\mathrm{op}} \leq C_1(1 + \|x\|_2^{q-1}),$$

for each $x \in \mathbb{R}^D$, and some positive constants C_0 and C_1. Then, we have $\mathbb{E}_{X\sim P}[\mathcal{T}_P g(X)] = 0$.

Proof. The proof is essentially that of Gorham et al. [2019, Proposition 3]. Note that by the moment assumption on P, we have $\mathbb{E}_{X\sim P}\left[(1 + \|X\|_2^{q_a+1})(1 + \|X\|_2^{q-1})\right] < \infty$ and thus

$$\mathbb{E}_{X\sim P}|\mathcal{T}_P g(X)| \leq 2\mathbb{E}_{X\sim P}\|b(X)\|_2\|g(X)\|_2 + D\mathbb{E}_{X\sim P}\|m(x)\|_{\mathrm{op}}\|\nabla g(X)\|_{\mathrm{op}} < \infty.$$

Thus, we may apply the dominated convergence theorem and then the divergence theorem to obtain

$$\begin{aligned}\mathbb{E}_{X\sim P}[\mathcal{T}_P g(X)] &= \lim_{r\to\infty} \int_{B_r} \langle \nabla, p(x)\{a(x) + c(x)\}g(x)\rangle \mathrm{d}x \\ &= \lim_{r\to\infty} \int_{\partial B_r} \langle n_r(z), \{a(x) + c(x)\}g(z)\rangle p(z)\mathrm{d}z,\end{aligned}$$

where $\mathrm{d}z$ denotes the $(D-1)$-dimensional Hausdorff measure. Let

$$f(r) = \int_{\partial B_r} \|a(z) + c(z)\|_{\mathrm{op}}\|g(z)\|_2 p(z)\mathrm{d}z$$

Then we have

$$\int_{\partial B_r} \langle n_r(z), g(z)\rangle p(z)\mathrm{d}z \leq f(r).$$

By the coarea formula (and integration with polar coordinates), we have

$$\begin{aligned}\int_0^\infty f(r)\mathrm{d}r &= \int_0^\infty \left\{\int_{\partial B_r} \|a(x) + c(x)\|_{\mathrm{op}}\|g(z)\|_2 p(z)\mathrm{d}z\right\}\mathrm{d}r \\ &\leq 2\lambda_m C_0 \int_0^\infty \left\{\int_{\partial B_r} (1 + \|z\|_2^{q_a+1})(1 + \|z\|_2^{q-1})p(z)\mathrm{d}z\right\}\mathrm{d}r \\ &= 2\lambda_m C_0 \int (1 + \|x\|_2^{q_a+1})(1 + \|x\|_2^{q-1})p(x)\mathrm{d}x < \infty.\end{aligned}$$

Thus, we have $\liminf_{r\to\infty} f(r) = 0$ and therefore $\mathbb{E}_{X\sim P}[\mathcal{T}_P g(X)] = 0$. □

6.A.4 DKSD upper bounds

Proposition 6.28. *Let $\mathcal{G}_\kappa$ be the RKHS defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Let $q \geq 1$. Assume that any function g in the unit ball $\mathcal{B}_1(\mathcal{G}_\kappa)$ satisfies the following: there exist some constants C_0, C_1, and C_2 such that*

$$\|\nabla^i g(x)\|_{\mathrm{op}} \leq C_i(1 + \|x\|_2^{q-1}) \textit{ for any } x \in \mathbb{R}^D \textit{ and } i \in \{0, 1, 2\}.$$

Assume a linear growth condition on m (Conditions 6.2, 6.11 with $q_a = 0$). Assume that b is $\phi_1(b)$-Lipschitz in the Euclidean norm, and m is $\phi_1(m)$-Lipschitz in the Frobenius norm. Suppose $P \in \mathcal{P}_q$. Then,

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) \leq C_{b,m} d_{\mathrm{pLip}_{1,q}}(Q, P),$$

where

$$C_{b,m} = \frac{\lambda_b C_1(5 + 2^{q-1})}{4} + 4C_0\phi_1(b) + \frac{\lambda_m C_2 D(5 + 2^{q-1})}{2} + 2\sqrt{D}C_1\phi_1(m)$$

In particular, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) \to 0$ if $d_{\mathcal{F}_q}(P, Q_n) \to 0$.

Proof. For any $g \in \mathcal{B}_1(\mathcal{G}_\kappa)$, we show that $\mathcal{T}_P g$ is a pseudo-Lipschitz function of order q. By the derivative assumptions, we have

$$\|g(x) - g(y)\|_2 \leq \frac{C_1}{2}\left(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}\right)\|x - y\|_2,$$

and

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq \frac{C_2}{2}\left(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}\right)\|x - y\|_2.$$

Also,

$$\frac{(1 + \|x\|_2)\left(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}\right)}{\left(1 + \|x\|_2^q + \|y\|_2^q\right)}\left(1 + \|x\|_2^q + \|y\|_2^q\right) \leq 5 + 2^{q-1}$$

Using these estimates, we obtain

$$\begin{aligned}
|\mathcal{T}_P g(x) - \mathcal{T}_P g(y)| &\leq 2\|b(x)\|_2\|g(x) - g(y)\|_2 + 2\|b(x) - b(y)\|_2\|g(y)\|_2 \\
&\quad + D\|m(x)\|_{\mathrm{op}}\|\nabla g(x) - \nabla g(y)\|_{\mathrm{op}} + \sqrt{D}\|m(x) - m(y)\|_{\mathrm{F}}\|\nabla g(y)\|_{\mathrm{op}} \\
&\leq C_{b,m}(1 + \|x\|_2^q + \|y\|_2^q)\|x - y\|_2,
\end{aligned}$$

where

$$C_{b,m} = \frac{\lambda_b C_1(5 + 2^{q-1})}{4} + 4C_0\phi_1(b) + \frac{\lambda_m C_2 D(5 + 2^{q-1})}{2} + 2\sqrt{D}C_1\phi_1(m).$$

As a result, for any $g \in \mathcal{B}_1(\mathcal{G}_\kappa)$,

$$\begin{aligned}
|\mathbb{E}_{X\sim Q}\mathcal{T}_P g(X)| &= C_{b,m}C_{b,m}^{-1}\,|\mathbb{E}_{X\sim Q}\mathcal{T}_P g(X) - \mathbb{E}_{Y\sim P}\mathcal{T}_P g(Y)| \\
&\leq C_{b,m} d_{\mathcal{F}_q}(Q_n, P),
\end{aligned}$$

where the equality holds since $\mathbb{E}_{Y\sim P}\mathcal{T}_P g(Y) = 0$ by Proposition 6.9. Taking the supremum over $\mathcal{B}_1(\mathcal{G}_\kappa)$ provides the required relation $\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) \le C_{b,m} d_{\mathcal{F}_q}(Q_n, P)$. □

Proposition 6.29. *Let $\mathcal{G}_\kappa$ be the RKHS of $\mathbb{R}^D$-valued functions defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Let $q \ge 1$. Assume that any function g in the unit ball $\mathcal{B}_1(\mathcal{G}_\kappa)$ satisfies the following: there exist some constants C_0, C_1, and C_2 such that*

$$\|\nabla^i g(x)\|_{\mathrm{op}} \le C_i(1 + \|x\|_2^{q-1}) \textit{ for any } x \in \mathbb{R}^D \textit{ and } i \in \{0, 1, 2\}.$$

Assume a quadratic growth condition on m (Conditions 6.2, 6.11 with $q_a = 1$). Assume that b is $\phi_1(b)$-Lipschitz, and m is pseudo-Lipschitz of order 1 in the operator norm with constant $\tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}$. Suppose $P \in \mathcal{P}_{q+1}$. Then,

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) \le C_{b,m} d_{\mathrm{pLip}_{1,q+1}}(Q, P),$$

where

$$C_{b,m} = (5 + 2^q)\left(\frac{\lambda_b C_1}{4} + \frac{\lambda_m C_2 D}{2} + C_1 D \tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}\right) + 4\phi_1(b) C_0.$$

In particular, we have $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) \to 0$ if $d_{\mathcal{F}_{q+1}}(P, Q_n) \to 0$.

Proof. The proof proceeds as in Proposition 6.18. For any $g \in \mathcal{B}_1(\mathcal{G}_\kappa)$, we have

$$\begin{aligned}
&|\mathcal{T}_P g(x) - \mathcal{T}_P g(y)| \\
&\le 2\|b(x)\|_2 \|g(x) - g(y)\|_2 + 2\|b(x) - b(y)\|_2 \|g(y)\|_2 \\
&\quad + D\|m(x)\|_{\mathrm{op}} \|\nabla g(x) - \nabla g(y)\|_{\mathrm{op}} + \sqrt{D}\|m(x) - m(y)\|_{\mathrm{F}} \|\nabla g(y)\|_{\mathrm{op}} \\
&\le \frac{\lambda_b C_1}{4}(1 + \|x\|_2)\big(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}\big)\|x - y\|_2 + 2\phi_1(b) C_0 (1 + \|x\|_2^{q-1})\|x - y\|_2 \\
&\quad + \frac{\lambda_m C_2 D}{8}(1 + \|x\|_2^2)(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1})\|x - y\|_2 \\
&\quad + C_1 D \tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}(1 + \|x\|_2^{q-1})(1 + \|x\|_2 + \|y\|_2)\|x - y\|_2 \\
&\le C_{b,m}(1 + \|x\|_2^{q+1} + \|y\|^{q+1})\|x - y\|_2,
\end{aligned}$$

where

$$C_{b,m} = (5 + 2^q)\left(\frac{\lambda_b C_1}{4} + \frac{\lambda_m C_2 D}{2} + C_1 D \tilde{\mu}_{\mathrm{pLip}}(m)_{1,1}\right) + 4\phi_1(b) C_0.$$

□

6.B Auxiliary results

6.B.1 Results from previous work

Lemma 6.30 (An extended version of Gorham et al. [2019, Lemma 17])**.** *Let G be a D-dimensional standard normal random vector, and fix $s > 0$. If $f : \mathbb{R}^D \to \mathbb{R}$ bounded and* *measurable, and* $f_s(x) = \mathbb{E}[f(x+sG)]$, *then*

$$M_0(f_s) \leq M_0(f),\ M_1(f_s) \leq \sqrt{\frac{2}{\pi}}\frac{M_0(f)}{s},$$
$$M_2(f_s) \leq \sqrt{2}\frac{M_0(f)}{s^2},\ \textit{and}\ M_3(f_s) \leq \frac{3M_0(f)}{s^3},$$

where $M_0(f_s) = \sup_{x\in\mathbb{R}^D}|f_s(x)|$, *and* $M_i(f_s) = \sup_{x\in\mathbb{R}^D}\|\nabla^i f_s(x)\|_{\mathrm{op}}$ *for* $i \in \{1,2,3\}$.

Proof. We prove the bound on $M_3(f_s)$, as the other bounds are given in Gorham et al. [2019, Lemma 17]. Let $\phi_s \in C^\infty$ be the density of sG and $*$ be the convolution operator. By Leibniz's rule,

$$\left\langle \nabla^3 f_s, u_1 \otimes u_2 \otimes u_3 \right\rangle = \left\langle (f * \nabla^3\phi_s)(x), u_1 \otimes u_2 \otimes u_3 \right\rangle.$$

The RHS can be evaluated as

$$\begin{aligned}
&\left|\left\langle (f * \nabla^3\phi_s)(x), u_1 \otimes u_2 \otimes u_3 \right\rangle\right| \\
&= \left|\int f(x-y)\left\langle \nabla^3\phi_s(y), u_1 \otimes u_2 \otimes u_3 \right\rangle \mathrm{d}y\right| \\
&\leq \frac{M_0(f)}{s^6}\int\left|\left\{\prod_{i=1}^3 \langle y, u_i\rangle - s^2\sum_{ijk}\langle u_i, y\rangle\langle u_j, u_k\rangle\right\}\right|\phi_s(y)\mathrm{d}y \\
&\leq \frac{M_0(f)}{s^6}\sqrt{\int\left\{\prod_{i=1}^3 \langle y, u_i\rangle - s^2\sum_{ijk}\langle u_i, y\rangle\langle u_j, u_k\rangle\right\}^2\phi_s(y)\mathrm{d}y} \\
&= \frac{M_0(f)}{s^6}\sqrt{s^6\mathbb{E}\left[\left(\sum_{ijk}\langle u_i, G\rangle\langle u_j, u_k\rangle\right)^2\right]} \\
&\leq \frac{3M_0(f)}{s^3}\cdot\|u_1\|_2\|u_2\|_2\|u_3\|_2.
\end{aligned}$$

where $\sum_{ijk}$ denotes the sum over the choices of (i,j,k) from $\{(1,2,3),(2,1,3),(3,1,2)\}$, the equality holds by Isserlis' theorem, and the last inequality follows from the Cauchy-Schwarz inequality after expanding the square. □

6.B.2 Miscellaneous results

Definition 6.31 (f-uniform integrability)**.** Let $\mathcal{Q}$ be a set of probability measures on $\mathbb{R}^D$, and $f : \mathbb{R}^D \to [0,\infty)$ be a nonnegative function that is integrable for each $Q \in \mathcal{Q}$. The function f is called uniformly integrable with respect to $\mathcal{Q}$, or the set $\mathcal{Q}$ is called f-uniformly integrable if

$$\lim_{r\to\infty}\sup_{Q\in\mathcal{Q}}\int_{\{f(x)>r\}} f\mathrm{d}Q = 0.$$

Note that for a sequence$\mathcal{Q} = \{Q_1, Q_2 \ldots\}$, the f-uniform integrability is equivalent to having

$$\lim_{r\to\infty}\limsup_{n\to\infty}\int_{\{f(x)>r\}} f\mathrm{d}Q_n = 0.$$

The following lemma is an analogue of Kallenberg [2021, Lemma 5.11]:

Lemma 6.32. *Let $Q_1, Q_2, \dots$ be a sequence of probability measures on a separable metric space $\mathcal{X}$ weakly converging to Q. Then, for a nonnegative continuous function f, we have*

1. $\int f \mathrm{d}Q \leq \liminf_{n\to\infty} \int f \mathrm{d}Q_n$,
2. $\lim_{n\to\infty} \int f \mathrm{d}Q_n \to \int f \mathrm{d}Q < \infty \Leftrightarrow \mathcal{Q} = \{Q_1, Q_2, \dots\}$ is f-uniformly integrable.

Proof. The proof follows Kallenberg [2021, Lemma 5.11]. Below, we use the notation $\mu(f) = \int f \mathrm{d}\mu$ for a measure μ.

For any $r > 0$, $x \mapsto f(x) \wedge r$ is a bounded continuous function. Thus, by the weak convergence assumption,

$$\begin{aligned}\liminf_{n\to\infty} Q_n(f) &\geq \liminf_{n\to\infty} Q_n(f \wedge r)\\ &= Q(f \wedge r).\end{aligned}$$

The first claim follows as we let $r \to \infty$.

For the second claim, let us assume that $\mathcal{Q}$ is f-uniformly integrable. For any $r > 0$, we have

$$\begin{aligned}&|Q_n(f) - Q(f)|\\ &\leq |Q_n(f) - Q_n(f \wedge r)| + |Q_n(f \wedge r) - Q_n(f \wedge r)| + |Q(f \wedge r) - Q(f)|.\end{aligned}$$

The first term on the RHS satisfies

$$\begin{aligned}|Q_n(f) - Q_n(f \wedge r)| &= Q_n\big((f - r)1_{\{f > r\}}\big)\\ &\leq 2Q_n\big(f 1_{\{f > r\}}\big)\\ &\leq 2 \sup_n Q_n\big(f 1_{\{f > r\}}\big).\end{aligned}$$

Note that $Q(f) \leq \liminf Q_n(f) < \infty$. Thus, letting $n \to \infty$ and then $r \to \infty$ proves the claim. For the other direction, assume $Q_n(f) \to Q(f) < \infty$. With a fixed $r > 0$, as $n \to \infty$,

$$\begin{aligned}Q_n(f 1_{\{f > r\}}) &\leq Q_n\big(f - f \wedge (r - f)_+\big)\\ &\to Q\big(f - f \wedge (r - f)_+\big),\end{aligned}$$

where we denote $(a)_+ = \max(a, 0)$. Since $x \mapsto f(x) \wedge (r - f(x))_+$ converges to f point-wise as $r \to \infty$, by the dominated convergence theorem we have $Q\big(f - f \wedge (r - f)_+\big) \to 0$. □

Lemma 6.33. *Let $w(x) = \left(a + b\|x\|_2^2\right)^q$ with $a \geq 1$, $b > 0$, and $q > 0$. Then,*

$$\begin{aligned}B_w(z) &:= \sup_{x \in \mathbb{R}^D, u \in [0,1]} \frac{w(x)}{w(x - uz)}\\ &\leq \left(1 + 2\left(1 + \frac{b\|z\|_2^2}{a}\right)\right)^q\end{aligned}$$

and

$$M_w := \sup_{x \in \mathbb{R}^D} \|\nabla \log w(x)\|_2 = \sup_{x \in \mathbb{R}^D} \frac{2bq\|x\|_2}{a + b\|x\|_2^2} \leq q\sqrt{\frac{b}{a}}.$$

Proof. We have

$$\begin{aligned} \frac{w(x+uz)}{w(x)} &= \left(\frac{a + b\|x+uz\|_2^2}{a + b\|x\|_2^2}\right)^q \leq \left(1 + \frac{\|x+uz\|_2^2}{(a/b) + \|x\|_2^2}\right)^q \\ &\leq \left(1 + 2\frac{\|x\|_2^2 + \|z\|_2^2}{(a/b) + \|x\|_2^q}\right)^q \\ &\leq \left(1 + 2\left(1 + \frac{b\|z\|_2^2}{a}\right)\right)^q. \end{aligned}$$

The first claim follows by observing

$$\sup_{x \in \mathbb{R}^D, u \in [0,1]} \frac{w(x)}{w(x-uz)} \leq \sup_{u \in [0,1]} \sup_{x \in \mathbb{R}^D} \frac{w(x+uz)}{w(x)}.$$

The second claim can be checked easily. □

Lemma 6.34 (A sinc function density and its moments)**.** *For an integer $q \geq 1$, let $s_q : \mathbb{R}^D \to (0, \infty)$ be a probability density function defined by*

$$s_q(x) = \left(\frac{1}{I_{4q,\infty}}\right)^D \prod_{d=1}^D (\operatorname{sinc}(x_d))^{4q},$$

where

$$\operatorname{sinc} : \mathbb{R} \to \mathbb{R},\ \operatorname{sinc}(x) = \begin{cases} \frac{\sin(x)}{x} & x \neq 0 \\ 1 & x = 0 \end{cases},$$

and $I_{4q,t} = 2\int_0^t \operatorname{sinc}^{4q}(x)\mathrm{d}x$ for $t \in (0, \infty]$. Let Z be a random variable with the law specified by s_{4q}. Then, the q-th moment $\mathbb{E}[\|Z\|_2^q]$ is finite. Specifically, it has the following upper bound:

$$\mathbb{E}[\|Z\|_2^{q'}] \leq \begin{cases} \frac{3D\log 2}{\pi} & q = 1, \\ D^{q'}\left(\frac{I_{4q,1}}{I_{4q,\infty}} + \frac{1}{2q-q'}\frac{1}{I_{4q,\infty}}\right)^D, & q > 1, \end{cases}$$

where q' is an integer such that $1 \leq q' \leq q$.

Proof. The density s_q is nonnegative due to the even power. We have

$$I_{4q,\infty} = \int_0^\infty \operatorname{sinc}^{4q}(x)\mathrm{d}x \leq \int_0^\infty \operatorname{sinc}^2(x)\mathrm{d}x = \frac{\pi}{2},$$

where the inequality holds as $|\operatorname{sinc}^2(x)| \leq 1$ everywhere, and the integral value is obtained by integration by parts and using $\int_0^\infty \operatorname{sinc}(x) = \pi/2$ [see, e.g., Abramowitz and Stegun, 1965, Chapter 5]. Thus, $I_{4q,\infty} < \infty$, and the existence of the density s_q is guaranteed.

Assume $q > 1$, the moment estimates are derived as follows:

$$\begin{aligned} &I_{4q,\infty}^{D}\mathbb{E}[\|Z\|_2^{q'}] \\ &= \int \|z\|_2^{q'} s_q(z)\mathrm{d}z \leq \int \|z\|_1^{q'} s_q(z)\mathrm{d}z \\ &= \sum_{i_1+\cdots+i_D=q'} \left(\frac{q'!}{i_1!i_2!\cdots i_D!}\right) \prod_{d=1}^{D} 2\int_0^\infty \frac{\sin^{2q}(z_d)}{|z_d|^{2q-i_d}} \left(\frac{\sin(z_d)}{z_d}\right)^{2q} \mathrm{d}z_d. \end{aligned}$$

The inequality holds because $\|\cdot\|_1$ upper bounds $\|\cdot\|_2$. By noting that $\mathrm{sinc}(x) \leq 1$ for all $x \in \mathbb{R}$, we can further evaluate the integral:

$$\begin{aligned} &\int_0^\infty \frac{\sin^{2q}(z_d)}{|z_d|^{2q-i_d}} \left(\frac{\sin(z_d)}{z_d}\right)^{2q} \mathrm{d}z_d \\ &\leq \int_0^1 \left(\frac{\sin(z_d)}{z_d}\right)^{2q} \mathrm{d}z_d + \int_1^\infty \frac{1}{|z_d|^{2q-i_d}}\mathrm{d}z_d \\ &= \frac{I_{4q,1}}{2} + \frac{1}{(2q-i_d)-1} \leq \frac{I_{4q,1}}{2} + \frac{1}{(2q-q')-1}. \end{aligned}$$

Thus, we have

$$\mathbb{E}[\|Z\|_2^{q'}] \leq D^{q'} \left(\frac{I_{4q,1}}{I_{4q,\infty}} + \frac{2}{(2q-q')-1}\frac{1}{I_{4q,\infty}}\right)^D.$$

When $q = 1$, a similar computation shows

$$\begin{aligned} \mathbb{E}[\|Z\|_2] &\leq \frac{2D}{I_{4,\infty}^D} \int_0^\infty \frac{\sin^4(z_d)}{z_d^3}\mathrm{d}z_d \cdot \prod_{d'\neq d} \int \frac{\sin^4(z_d)}{z_d^4}\mathrm{d}z_{d'} \\ &= \frac{2D\log 2}{I_{4,\infty}} = \frac{3D\log 2}{\pi}. \end{aligned}$$

□

Remark 6.35. The tedious expectation estimate is derived to ensure that the inside of the power function is small (typically less than 1).

Lemma 6.36. *Let* $f(t) = e^{-1/t}1_{(0,\infty)}(t)$. *Let*

$$h(t) = \frac{f\big(r+\delta-t\big)}{f\big(r+\delta-t\big) + f\big(t-r\big)}.$$

Then, $h(t) = 0$ *for* $t \geq r+\delta$, $h(t) = 1$ *for* $t \leq r$, *and* $0 < h(t) < 1$ *otherwise. Moreover, the first two derivatives of* h *satisfies the following: for any* $t \geq 0$,

$$\left|\frac{\mathrm{d}h}{\mathrm{d}t}\right|(t) \leq \begin{cases} \frac{2}{\delta^2} & \delta \leq 1/2 \\ 8e^{1/\delta-2} & \delta > 1/2 \end{cases},$$

$$\left|\frac{\mathrm{d}^2 h}{\mathrm{d}t^2}\right|(t) \le \begin{cases} 2\left(\frac{1-2\delta}{\delta^4}\right) + \frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & 0 < \delta \le 1/2 - \sqrt{1/12}, \\ \frac{96\sqrt{3}}{(\sqrt{3}-1)^4} e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}} e^{1/\delta} + \frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & 1/2 - \sqrt{1/12} < \delta \le 1/2, \\ \left(\frac{96\sqrt{3}}{(\sqrt{3}-1)^4} e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}} + 32e^{-4}(e^{1/\delta}+4)\right) e^{1/\delta}. & \delta > 1/2. \end{cases}$$

Proof. The cutoff property of h is well known, and therefore we focus only on the bounds on the derivatives.

As a preparatory step, let us write down the first three derivatives of f, which are given, for $t > 0$, as follows:

$$\begin{aligned} f'(t) &= \frac{1}{t^2} e^{-1/t}, \\ f''(t) &= \frac{1}{t^4} e^{-1/t}(-2t+1), \text{ and} \\ f^{(3)}(t) &= \frac{1}{t^6} e^{-1/t}(6t^2 - 6t + 1) = \frac{6}{t^6} e^{-1/t} \left\{ t - \left(\frac{1}{2} + \sqrt{\frac{1}{12}} \right) \right\} \left\{ t - \left(\frac{1}{2} - \sqrt{\frac{1}{12}} \right) \right\}. \end{aligned}$$

The first derivative f' increases from 0 to $1/2$ and then decreases. Thus, $f'(t) \le f'(\delta \wedge 1/2)$ for $\delta > 0$ and $0 < t \le \delta$. The second derivative f'' increases from zero to its maximum at $t = \left(1/2 - \sqrt{1/12}\right)$, and we have $f''(t) \le f''\{\delta \wedge (1/2 - \sqrt{1/12})\}$ for $\delta > 0$ and $0 < t \le \delta$. Also, note that we have for $0 \le t - r \le \delta$,

$$\frac{1}{f\big(\delta - (t-r)\big) + f(t-r)} \le \frac{1}{f(\delta)}.$$

Now we evaluate the first derivative of h. We consider the region $r < t < r + \delta$, as otherwise h is constant. Using $h \le 1$, we obtain

$$\begin{aligned} \left|\frac{\mathrm{d}h}{\mathrm{d}t}\right| &= \left| -\frac{f'\big(r+\delta-t\big) f\big(t-r\big)}{\big(f\big(r+\delta-t\big) + f\big(t-r\big)\big)^2} + \frac{f\big(r+\delta-t\big) f'\big(t-r\big)}{\big(f\big(r+\delta-t\big) + f\big(t-r\big)\big)^2} \right| \\ &\le \left| \frac{f'\big(r+\delta-t\big)}{f\big(r+\delta-t\big) + f\big(t-r\big)} \right| + \left| \frac{f'\big(t-r\big)}{f\big(r+\delta-t\big) + f\big(t-r\big)} \right| \\ &\le 2\frac{f'(\delta \wedge 1/2)}{f(\delta)} \\ &= \begin{cases} \frac{2}{\delta^2} & \delta \le 1/2 \\ 8e^{1/\delta - 2} & \delta > 1/2 \end{cases}. \end{aligned}$$

The second derivative can be similarly evaluated as follows:

$$\begin{aligned} \left|\frac{\mathrm{d}^2 h}{\mathrm{d}t^2}\right| = \Bigg| &-\frac{f''(r+\delta-t) f(t-r) - 2f'(r+\delta-t) f'(r-t) + f''(t-r) f(r+\delta-t)}{\big(f\big(r+\delta-t\big) + f\big(t-r\big)\big)^2} \\ &-2\frac{\big(-f'(r+\delta-t) f(r-t) + f'(t-r) f(r+\delta-t)\big)^2}{\big(f\big(r+\delta-t\big) + f\big(t-r\big)\big)^3} \Bigg| \end{aligned}$$

$$
\begin{aligned}
&\le \left|\frac{f''(r+\delta-t)}{f\big(r+\delta-t\big)+f\big(t-r\big)}\right| + 2\left|\frac{f'(r+\delta-t)f'(r-t)}{\big(f\big(r+\delta-t\big)+f\big(t-r\big)\big)^2}\right| \\
&\quad + \left|\frac{f''(t-r)}{f\big(r+\delta-t\big)+f\big(t-r\big)}\right| \\
&\quad + 2\left|\frac{-f'(r+\delta-t)f(r-t)+f'(t-r)f(r+\delta-t)}{\big(f\big(r+\delta-t\big)+f\big(t-r\big)\big)}\right|^2 \frac{1}{f\big(r+\delta-t\big)+f\big(t-r\big)} \\
&\le \left|\frac{f''(r+\delta-t)}{f\big(r+\delta-t\big)+f\big(t-r\big)}\right| + 2\left|\frac{f'(r+\delta-t)f'(t-r)}{\big(f\big(r+\delta-t\big)+f\big(t-r\big)\big)^2}\right| \\
&\quad + \left|\frac{f''(t-r)}{f\big(r+\delta-t\big)+f\big(t-r\big)}\right| + 2\frac{\big(|f'(r+\delta-t)|+|f'(t-r)|\big)^2}{f\big(r+\delta-t\big)+f\big(t-r\big)}.
\end{aligned}
$$

Evaluating each term yields

$$
\begin{aligned}
\left|\frac{\mathrm{d}^2 h}{\mathrm{d}t^2}\right|(t) &\le 2\left|\frac{f''\{\delta\wedge(1/2-\sqrt{1/12})\}}{f(\delta)}\right| + 2\left|\frac{f'(\delta\wedge 1/2)}{f(\delta)}\right|^2 + 8\frac{|f'(\delta\wedge 1/2)|^2}{f(\delta)}. \\
&\le \begin{cases}
2\left(\frac{1-2\delta}{\delta^4}\right)+\frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & 0<\delta\le\frac{1}{2}-\frac{1}{\sqrt{12}}, \\
\frac{96\sqrt{3}}{(\sqrt{3}-1)^4}e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}}e^{1/\delta}+\frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & \frac{1}{2}-\frac{1}{\sqrt{12}}<\delta\le 1/2, \\
\left(\frac{96\sqrt{3}}{(\sqrt{3}-1)^4}e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}}+32e^{-4}(e^{1/\delta}+4)\right)e^{1/\delta}. & \delta>\frac{1}{2}.
\end{cases}
\end{aligned}
$$

□

Lemma 6.37. *Let $f(t)=e^{-1/t}1_{(0,\infty)}(t)$. Let $1_{r,\delta}(x)=h\big(\|x\|_2\big)$ be a smooth bump function defined by*

$$
h(t)=\frac{f\big(r+\delta-t\big)}{f\big(r+\delta-t\big)+f\big(t-r\big)}.
$$

Then, we have $1_{r,\delta}(x)=1$ for $\|x\|_2\le r, 0<1_{r,\delta}(x)<1$ for $r<\|x\|_2<r+\delta$, and $1_{r,\delta}(x)=0$ for $\|x\|_2\ge r+\delta$; the gradient $\nabla I_{r,\delta}(x)$ vanishes outside $\{x\in\mathbb{R}^D : r<\|x\|_2<r+\delta\}$. Furthermore, we can uniformly bound the derivative norms $\|\nabla_x 1_{r,\delta}(x)\|_2$ and $\|\nabla_x 1_{r,\delta}(x)\|_2$ by constants depending only on δ. Specifically,

$$
\|\nabla_x 1_{r,\delta}(x)\|_2 \le \frac{2}{\delta^2}1_{(0,1/2]}(\delta)+8e^{1/\delta-2}1_{(1/2,\infty)}(\delta)
$$

and

$$
\begin{aligned}
\|\nabla_x^2 1_{r,\delta}(x)\|_{\mathrm{op}} &< \left|\frac{\mathrm{d}^2 h}{\mathrm{d}r^2}\big(\|x\|_2\big)\right| + \frac{2}{r}\left|\frac{\mathrm{d}h}{\mathrm{d}r}\big(\|x\|_2\big)\right| \\
&\le \begin{cases}
\frac{2}{r\delta^2}+2\left(\frac{1-2\delta}{\delta^4}\right)+\frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & 0<\delta\le\frac{1}{2}-\frac{1}{\sqrt{12}}, \\
\frac{2}{r\delta^2}+\frac{96\sqrt{3}}{(\sqrt{3}-1)^4}e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}}e^{1/\delta}+\frac{2}{\delta^4}\left(1+4e^{-1/\delta}\right) & \frac{1}{2}-\frac{1}{\sqrt{12}}<\delta\le\frac{1}{2}, \\
\frac{8}{r}e^{1/\delta-2}+\left(\frac{96\sqrt{3}}{(\sqrt{3}-1)^4}e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}}+32e^{-4}(e^{1/\delta}+4)\right)e^{1/\delta}. & \delta>\frac{1}{2}.
\end{cases}
\end{aligned}
$$

Proof. We restrict the evaluation of the derivatives to the region $r<\|x\|_2\le r+\delta$, as otherwise

h is constant and has zero derivatives. According to Lemma 6.36, the first derivative satisfies

$$\begin{aligned}\|\nabla_x h(\|x\|_2)\|_2 &= \left\| \frac{x}{\|x\|_2} \left.\frac{\mathrm{d}h}{\mathrm{d}t}\right|_{t=\|x\|_2} \right\|_2 \\ &= \left| \frac{\mathrm{d}h}{\mathrm{d}t}(\|x\|) \right| \\ &\leq \frac{2}{\delta^2} 1_{(0,1/2]}(\delta) + 8e^{1/\delta-2} 1_{(1/2,\infty)}(\delta).\end{aligned}$$

The second derivative can also be evaluated as

$$\nabla_x^2 h(\|x\|_2) = \frac{\mathrm{d}^2 h}{\mathrm{d}t^2} \frac{x \otimes x}{\|x\|^2} + \frac{\mathrm{d}h}{\mathrm{d}t} \frac{1}{\|x\|} \left(I - \frac{x \otimes x}{\|x\|^2} \right).$$

Consequently, we have

$$\begin{aligned}\|\nabla_x^2 h(\|x\|_2)\|_{\mathrm{op}} &= \sup_{\|u\|_2 = \|v\|_2 = 1,} \langle u, \nabla_x^2 h(\|x\|_2) v \rangle \\ &< \left| \frac{\mathrm{d}^2 h}{\mathrm{d}r^2}(\|x\|_2) \right| + \frac{2}{r} \left| \frac{\mathrm{d}h}{\mathrm{d}r}(\|x\|_2) \right|.\end{aligned}$$

The rest of the proof follows from the estimates of Lemma 6.36. □

Lemma 6.38. *For $\nu > 2$, let the standard multivariate t-distribution be*

$$p(x) = \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \nu^{\frac{D}{2}} \pi^{\frac{D}{2}}} \left(1 + \|x\|_2^2\right)^{-\frac{\nu+D}{2}}.$$

Then, the t-distribution satisfies the dissipativity condition in Proposition 6.6 with $q = 1$ and $\sigma(x) = \sqrt{1 + \nu^{-1}\|x\|_2^2} I$. For this choice, we have $\lambda_a = 4$, and the diffusion is dissipative (Condition 6.3) with $\alpha = 1 - 2\nu^{-1}$.

Proof. For this density, we have

$$\begin{aligned}\nabla \log p(x) = -\frac{\nu + D}{\nu} \frac{x}{1 + \nu^{-1}\|x\|_2^2}, \ a(x) = 1 + \nu^{-1}\|x\|_2^2, \text{ and} \\ 2b(x) = a(x) \nabla \log p(x) + \langle \nabla, a(x) \rangle = -\frac{\nu + D - 2}{\nu} x.\end{aligned}$$

Thus,

$$\begin{aligned}&2\langle b(x) - b(y), x - y \rangle + \|\sigma(x) - \sigma(y)\|_{\mathrm{F}}^2 - \|\sigma(x) - \sigma(y)\|_{\mathrm{op}}^2 \\ &= -\frac{\nu + D - 2}{\nu} \|x - y\|_2^2 + (D-1) \left(\sqrt{1 + \nu^{-1}\|x\|_2^2} - \sqrt{1 + \nu^{-1}\|y\|_2^2} \right)^2 \\ &\leq -\frac{\nu + D - 2}{\nu} \|x - y\|_2^2 + \frac{D-1}{\nu} \|x - y\|_2^2 \\ &= -\left(1 - \frac{1}{\nu}\right) \|x - y\|_2^2.\end{aligned}$$

In particular,

$$\begin{aligned}\mathcal{A}_P\|x\|_2^2 =& 2\langle b(x), x\rangle + \|\sigma(x)\|_{\mathrm{F}}^2 \\ =& -\frac{\nu+D-2}{\nu}\|x\|^2 + D(1+\nu^{-1}\|x\|_2^2) \\ =& -(1-2\nu^{-1})\|x\|^2 + D.\end{aligned}$$

□

6.B.3 Results concerning approximation

Lemma 6.39. *Let* $B_w(z) = \sup_{x\in\mathbb{R}^D, u\in[0,1]} w(x)/w(x-uz)$ *and* $M_w = \sup_{x\in\mathbb{R}^D}\|\nabla \log w(x)\|_2$. *Let* Z *be a random variable with* $\mathbb{E}_Z\big[\|Z\|_2 B_w(Z)\big] < \infty$. *With fixed* $\rho \in (0,1]$, *define for* $g^w_{R,\delta} := g_{R,\delta}/w$, *the convolution*

$$g^w_\rho(x) = \mathbb{E}_Z\left[\frac{g_{R,\delta}(x-\rho Z)}{w(x-\rho Z)}\right].$$

We have

$$\|g^w_{R,\delta} - g^w_\rho\|_2 \leq \frac{\rho}{w(x)}\cdot \tilde{u}^{(1)}_{P,\delta,D,w}\{1+(R+\delta)^{q-1}\}$$

with a constant

$$\tilde{u}^{(1)}_{P,\delta,D,w} = \Big[\sqrt{D}\zeta_1 M_w \mathbb{E}\Big[\|Z\|_2\, B_w(\rho Z)\Big] + \{\zeta_2 + C_\delta\zeta_1\}\Big],$$

where C_δ *is a constant bounding* $\|\nabla 1_{R,\delta}(x)\|_2$ *satisfying* $C_\delta = 8e^{-\delta}$ *for* $\delta > 1/2$.

Proof. By the definition of g_ρ, we have

$$g^w_\rho(x) - g^w_{R,\delta}(x) = \mathbb{E}_Z\left[g_{R,\delta}(x-\rho Z)\left\{\frac{1}{w(x-\rho Z)} - \frac{1}{w(x)}\right\} + \frac{g_{R,\delta}(x-\rho Z) - g_{R,\delta}(x)}{w(x)}\right].$$

To bound each quantity inside the expectation, we derive their norm estimates. For $g_{R,\delta}$, we have

$$\begin{aligned}M_0(g_{R,\delta}) :=& \sup_{x\in\mathbb{R}^D}\|g_{R,\delta}(x)\|_{\mathrm{op}} \\ =& \max_{\|x\|_2\leq R+\delta}\|g(x)\|_2 = \sqrt{D}\zeta_1\{1+(R+\delta)^{q-1}\}\end{aligned} \tag{6.6}$$

and

$$\begin{aligned}\tilde{\pi}(g_{R,\delta})_{1,0} :=& \sup_{x'\in\mathbb{R}^D}\|\nabla g_{R,\delta}(x')\|_{\mathrm{op}} \\ \leq& \sup_{x'\in\mathbb{R}^D} 1_{R,\delta}(x)\,\|\nabla g(x)\|_{\mathrm{op}} + \sup_{x'\in\mathbb{R}^D}\|\nabla 1_{R,\delta}(x)\|_2\,\|g(x)\|_2 \\ \leq& (\zeta_2 + C_\delta\zeta_1)\,\{1+(R+\delta)^{q-1}\},\end{aligned} \tag{6.7}$$

where $C_\delta$is a universal constant depending on δ. Further, by the fundamental theorem of calculus,

$$
\begin{aligned}
\|g_{K,\delta}(x-\rho Z)-g_{K,\delta}(x)\|_2 &\le \rho\left\|\int_0^1 \nabla g_{R,\delta}(x-t\rho Z)Z_2\mathrm{d}t\right\|_2\\
&\le \rho\int_0^1 \|\nabla g_{R,\delta}(x-t\rho Z)\|_{\mathrm{op}}\|Z\|_2\mathrm{d}t_2\\
&\le \rho\|Z\|_2\tilde{\pi}(g_{R,\delta})_{1,0}.
\end{aligned}
$$

We also have

$$
\begin{aligned}
\left|\frac{1}{w(x-\rho Z)}-\frac{1}{w(x)}\right| &= \rho\int_0^1\left|\left\langle\frac{\nabla w}{w^2}(x-t\rho Z),Z\right\rangle\right|\mathrm{d}t\\
&\le \rho\left\|Z\right\|_2\frac{1}{w(x)}\int_0^1\frac{w(x)}{w(x-t\rho Z)}\left\|\frac{\nabla w}{w}(x-t\rho Z)\right\|_2\mathrm{d}t \qquad (6.8)\\
&\le \frac{\rho M_w\left\|Z\right\|_2 B_w(\rho Z)}{w(x)}.
\end{aligned}
$$

Combining these evaluations yields

$$
\begin{aligned}
\|g^w_\rho(x)-g^w_{R,\delta}(x)\|_2 &\le \mathbb{E}_Z\left[\|g_{R,\delta}(x-\rho Z)\|_2\left|\frac{1}{w(x-\rho Z)}-\frac{1}{w(x)}\right|\right]\\
&\quad+\mathbb{E}_Z\left[\left\|\frac{g_{R,\delta}(x-\rho Z)-g_{R,\delta}(x)}{w(x)}\right\|_2\right]\\
&\le \frac{\rho}{w(x)}\left[M_w\mathbb{E}\left[\left\|Z\right\|_2 B_w(\rho Z)\right]M_0(g_{R,\delta})+\tilde{\pi}(g_{R,\delta})_{1,0}\right]\\
&=\left[\sqrt{D}\zeta_1 M_w\mathbb{E}\left[\left\|Z\right\|_2 B_w(\rho Z)\right]+\{\zeta_2+C_\delta\zeta_1\}\right]\cdot\frac{\rho\{1+(R+\delta)^{q-1}\}}{w(x)}
\end{aligned}
$$

□

Lemma 6.40. *Define symbols as in Lemma 6.39. For each fixed $\rho\in(0,1]$, we have*

$$
\|\nabla g^w_\rho(x)-\nabla g^w_{R,\delta}(x)\|_{\mathrm{op}}\le\frac{\rho}{w(x)}\tilde{u}^{(2)}_{P,\delta,D,w}(x)\cdot\{1+(R+\delta)^{q-1}\},
$$

where

$$
\begin{aligned}
\tilde{u}^{(2)}_{P,\delta,D,w}=\Big\{&(1+M_w)\left(\zeta_3+2C_\delta\zeta_2+C_{R,\delta}\zeta_1\right)\mathbb{E}_Z\left[\|Z\|_2\right]\\
&+M_w\left(1+2M_w\right)\left(\left(\zeta_2+C_\delta\zeta_1\right)\right)\mathbb{E}_Z\left[\left\|Z\right\|_2 B_w(Z)\right]\Big\},
\end{aligned}
$$

where C_δ and $C_{R,\delta}$ are respective uniform bounds on $\|\nabla 1_{R,\delta}(x)\|_2$ and $\|\nabla^2 1_{R,\delta}(x)\|_{\mathrm{op}}$, satisfying $C_\delta=8e^{-\delta}$ and

$$
C_{R,\delta}=\frac{8}{R}e^{1/\delta-2}+\left(\frac{96\sqrt{3}}{(\sqrt{3}-1)^4}e^{-\frac{2\sqrt{3}}{\sqrt{3}-1}}+32e^{-4}(e^{1/\delta}+4)\right)e^{1/\delta}.
$$

for $\delta>1/2$.

Proof. Before the proof, we introduce a notation. We denote the l-mode (vector) product of a tensor $T \in \mathbb{R}^{D_1 \times \cdots \times D_L}$ with a vector $v \in \mathbb{R}^{D_l}$ by $T \bar{\times}_l v$. The resulting tensor is of order $L-1$; its size is $(D_1, \ldots, D_{l-1}, D_{l+1}, \ldots, D_L)$; it is expressed element-wise as

$$(T \bar{\times}_l v)_{i_1 \cdots i_{l-1} i_{l+1} \cdots i_L} = \sum_{i_l=1}^{D_l} T_{i_1 \cdots i_l \cdots L} v_{i_l}.$$

First, note that

$$\begin{aligned} \|\nabla g_\rho^w(x) - \nabla g_{R,\delta}^w(x)\|_{\mathrm{op}} &= \left\| \mathbb{E}_Z \Big[\nabla g_{R,\delta}^w(x - \rho Z) - \nabla g_{R,\delta}^w(x) \Big] \right\|_{\mathrm{op}} \\ &\leq \underbrace{\left\| \mathbb{E}_Z \left[\frac{\nabla g_{R,\delta}(x-\rho Z)}{w(x-\rho Z)} - \frac{\nabla g_{R,\delta}(x)}{w(x)} \right] \right\|_{\mathrm{op}}}_{\text{(a)}} \\ &\quad + \underbrace{\left\| \mathbb{E}_Z \left[\frac{\nabla \log w(t-\rho Z)}{w(x-\rho Z)} \otimes \nabla g_{R,\delta}(x-\rho Z) - \frac{\nabla \log w(x)}{w(x)} \otimes \nabla g_{R,\delta}(x) \right] \right\|_{\mathrm{op}}}_{\text{(b)}}, \end{aligned} \tag{6.9}$$

In the first line, we have exchanged the gradient and the expectation operation. We evaluate each term below.

The term (a) is evaluated as

$$\begin{aligned} \left\| \mathbb{E}_Z \left[\frac{\nabla g_{R,\delta}(x-\rho Z)}{w(x-\rho Z)} - \frac{\nabla g_{R,\delta}(x)}{w(x)} \right] \right\|_{\mathrm{op}} &\leq \underbrace{\left\| \mathbb{E}_Z \left[\nabla g_{R,\delta}(x-\rho Z) \left(\frac{1}{w(x-\rho Z)} - \frac{1}{w(x)} \right) \right] \right\|_{\mathrm{op}}}_{\text{(i)}} \\ &\quad + \underbrace{\frac{1}{w(x)} \left\| \mathbb{E}_Z \left[\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x) \right] \right\|_{\mathrm{op}}}_{\text{(ii)}}, \end{aligned} \tag{6.10}$$

The term (i) is bounded as

$$\begin{aligned} \text{(i)} &\leq \mathbb{E}_Z \left[\left| \frac{1}{w(x-\rho Z)} - \frac{1}{w(x)} \right| \|\nabla g_{R,\delta}(x-\rho Z)\|_{\mathrm{op}} \right] \\ &\leq \frac{\rho}{w(x)} M_w \tilde{\pi}(g_{R,\delta})_{1,0} \mathbb{E}_Z \left[\|Z\|_2 B_w(\rho Z) \right]. \end{aligned}$$

The term (ii) is evaluated as follows. By the fundamental theorem of calculus,

$$\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x) = -\rho \int_0^1 \nabla^2 g_{R,\delta}(x - t\rho Z) \bar{\times}_1 Z \mathrm{d}t.$$

Note that the operator norm of $\nabla^2 g_{R,\delta}$ is bounded by

$$\begin{aligned}
&\tilde{\pi}(g_{R,\delta})_{2,0} := \sup_{x\in\mathbb{R}^D} \|\nabla^2 g_{R,\delta}(x)\|_{\mathrm{op}} \\
&\leq \sup_{x\in\mathbb{R}^D} \left(1_{R,\delta}(x)\|\nabla^2 g(x)\|_{\mathrm{op}} + 2\|\nabla 1_{R,\delta}(x)\|_2\|g_{R,\delta}(x)\|_2 + \|\nabla^2 1_{R,\delta}(x)\|_2\|g_{R,\delta}(x)\|_2\right) \\
&\leq \sup_{x\in\mathbb{R}^D} \left(1_{R,\delta}(x)\zeta_3 + 2\|\nabla 1_{R,\delta}(x)\|_2\zeta_2 + \|\nabla^2 1_{R,\delta}(x)\|_2\zeta_1\right)(1+\|x\|_2^{q-1}) \\
&\leq (\zeta_3 + 2C_\delta\zeta_2 + C_{R,\delta}\zeta_1)\left\{1+(R+\delta)^{q-1}\right\}.
\end{aligned} \tag{6.11}$$

Thus, the operator norm in the term (ii) is bounded as

$$\begin{aligned}
&\left\|\mathbb{E}_Z\left[\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x)\right]\right\|_{\mathrm{op}} \\
&\leq \mathbb{E}_Z\left[\|\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x)\|_{\mathrm{op}}\right] \\
&= \rho\mathbb{E}_Z\left[\|Z\|_2 \sup_{\|u^{(3)}\|_2=1}\left\|\left\{\int_0^1 \nabla^2 g_{R,\delta}(x-t\rho Z)\bar{\times}_1\frac{Z}{\|Z\|_2}\mathrm{d}t\right\}\bar{\times}_3 u^{(3)}\right\|_2\right] \\
&\leq \rho\mathbb{E}_Z\left[\|Z\|_2 \sup_{\|u^{(l)}\|_2=1, l\in\{1,2,3\}}\left|\left\{\int_0^1 \nabla^2 g_{R,\delta}(x-t\rho Z)\bar{\times}_1 u^{(1)}\mathrm{d}t\right\}\bar{\times}_3 u^{(3)}\bar{\times}_2 u^{(2)}\right|\right] \\
&\leq \rho\mathbb{E}_Z\left[\|Z\|_2\int_0^1 \sup_{\|u^{(l)}\|=1}|\langle\nabla^2 g_{R,\delta}(x-t\rho Z), u^{(1)}\otimes u^{(2)}\otimes u^{(3)}\rangle|\mathrm{d}t\right] \\
&\leq \rho\mathbb{E}_Z\left[\|Z\|_2\int_0^1 \|\nabla^2 g_{R,\delta}(x-t\rho Z)\|_{\mathrm{op}}\mathrm{d}t\right] \\
&\leq \rho\tilde{\pi}(g_{R,\delta})_{2,0}\mathbb{E}_Z\|Z\|_2.
\end{aligned}$$

Thus, the term (ii) is bounded by

$$\frac{\rho}{w(x)}\tilde{\pi}(g_{K,\delta})_{2,0}\mathbb{E}_Z\|Z\|_2.$$

Similarly, we can evaluate the term (b) in (6.9)

$$\begin{aligned}
\text{(b)} &= \frac{\|\mathbb{E}_Z[\nabla w(t-\rho Z)\otimes\nabla g_{R,\delta}(x-\rho Z) - \nabla w(x)\otimes\nabla g_{R,\delta}(x)]\|_{\mathrm{op}}}{w(x)} \\
&\leq \frac{1}{w(x)}\left\|\mathbb{E}_Z\left[\frac{\nabla\log w(t-\rho Z)}{w(x-\rho Z)}\otimes\left(\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x)\right)\right]\right\|_{\mathrm{op}} \\
&\quad + \frac{1}{w(x)}\left\|\mathbb{E}_Z\left[\left(\frac{\nabla\log w(t-\rho Z)}{w(x-\rho Z)} - \frac{\nabla\log w(x)}{w(x)}\right)\otimes\nabla g_{R,\delta}(x)\right]\right\|_{\mathrm{op}} \\
&\leq \frac{1}{w(x)}\mathbb{E}_Z\left\|\frac{\nabla\log w(t-\rho Z)}{w(x-\rho Z)}\right\|_2\|\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x)\|_{\mathrm{op}} \\
&\quad + \frac{1}{w(x)}\|\nabla g_{R,\delta}(x)\|_2\mathbb{E}_Z\left\|\frac{\nabla\log w(t-\rho Z)}{w(x-\rho Z)} - \frac{\nabla\log w(x)}{w(x)}\right\|_2 \\
&\leq \frac{M_w}{w(x)}\mathbb{E}_Z\|\nabla g_{R,\delta}(x-\rho Z) - \nabla g_{R,\delta}(x)\|_{\mathrm{op}}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2\rho M_w^2}{w(x)^2}\tilde{\pi}(g_{R,\delta})_{1,0}\mathbb{E}_Z\Big[\|Z\|_2\, B_w(\rho Z)\Big]\\
&\leq \frac{\rho M_w}{w(x)}\left(\mathbb{E}_Z\|Z\|_2\tilde{\pi}(g_{R,\delta})_{2,0} + \frac{2M_w}{w(x)}\tilde{\pi}(g_{w,\delta})_{1,0}\mathbb{E}_Z\Big[\|Z\|_2\, B_w(\rho Z)\Big]\right).
\end{aligned}
$$

With $B(\rho Z) \leq B(Z)$ in mind, combining the results, we obtain

$$
\|\nabla g_\rho^w(x) - \nabla g_{R,\delta}^w(x)\|_{\mathrm{op}} \leq \frac{\rho}{w(x)}\tilde{u}^{(2)}_{P,\delta,D,w}(x)\{1 + (R+\delta)^{q-1}\},
$$

where

$$
\begin{aligned}
\tilde{u}^{(2)}_{P,\delta,D,w} &= (1+M_w)\mathbb{E}_Z\|Z\|_2\tilde{\pi}(g_{R,\delta})_{2,0}\\
&\quad + M_w\left(1+2M_w\right)\mathbb{E}_Z\Big[\|Z\|_2\, B_w(Z)\Big]\tilde{\pi}(g_{R,\delta})_{1,0}\\
&= \Big\{(1+M_w)\left(\zeta_3 + 2C_\delta\zeta_2 + C_{R,\delta}\zeta_1\right)\mathbb{E}_Z\big[\|Z\|_2\big]\\
&\quad + M_w\left(1+2M_w\right)\left((\zeta_2 + C_\delta\zeta_1)\right)\mathbb{E}_Z\Big[\|Z\|_2\, B_w(Z)\Big]\Big\}.
\end{aligned}
$$

□

6.B.4 Results for DKSD

6.B.4.1 DKSD and uniform integrability

Lemma 6.41 (KSD upper-bounds the integrability rate)**.** *Let $\mathcal{G}_\kappa$ be the RKHS of $\mathbb{R}^D$-valued functions defined by a matrix-valued kernel $\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D\times D}$. Suppose there exists a function $g \in \mathcal{G}_\kappa$ such that $\mathcal{T}_P g(x) \geq \nu$ for any $x \in \mathbb{R}^D$ with some constant $\nu \in \mathbb{R}$, and $\liminf \|x\|_2^{-(q+\theta)}\mathcal{T}_P g(x) \geq \eta$ for some $q \geq 0$, $\eta > 0$, and $\theta > 0$ as $\|x\|_2 \to \infty$. Assume $\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_\kappa) < \infty$ for a distribution $Q \in \mathcal{P}_{q+\theta}$. Then, for sufficiently small $\varepsilon > 0$, we have*

$$
\begin{aligned}
R_q(Q,\varepsilon) &:= \inf\left\{r \geq 1 : \int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q(x) \leq \varepsilon\right\}\\
&\leq \left\{2\left(1+\frac{q}{\theta}\right)\left(\frac{\mathcal{S}(Q,\mathcal{T}_P,\mathcal{G}_\kappa) - \nu}{\eta\varepsilon}\right)\right\}^{\frac{1}{\theta}\vee\frac{q}{\theta}}.
\end{aligned}
$$

Thus, for a sequence of measures $\{Q_1, Q_2 \ldots\} \subset \mathcal{P}_{q+\theta}$, we have

$$
\limsup_{n\to\infty} \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa) < \infty \Rightarrow \limsup_{n\to\infty} R_q(Q_n, \varepsilon) < \infty.
$$

In particular, if the sequence $\{Q_1, Q_2 \ldots\}$ does not have uniformly integrable q-th moments, then Stein discrepancy $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_\kappa)$ diverges.

Proof. Let $g \in \mathcal{G}_{kI}$ be a function with the stated properties. Let $f(x) = \|x\|_2^q 1\{\|x\|_2 > r\}$. We consider the integral

$$
\int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q(x) = \int f(x)\mathrm{d}Q(x) = \int_0^\infty Q(\{f(x) > t\})\mathrm{d}t.
$$

By dividing the range of the integral, we obtain

$$\begin{aligned}\int_0^\infty Q(\{f(x) > t\})\mathrm{d}t &= r^q Q(\{\|x\|_2 > r\}) + \int_{r^q}^\infty Q(\{f(x) > t\})\mathrm{d}t \\ &= r^q Q(\{\|x\|_2 > r\}) + \int_r^\infty Q\{\|x\|_2 > t^{1/q}\}\mathrm{d}t,\end{aligned}$$

where we regard the second term as zero when $q = 0$.

We evaluate the tail probabilities in terms of the Stein discrepancy. Following the proof of Gorham and Mackey [2017, Lemma 17], we define $\gamma(r) = \inf\{\mathcal{T}_P g(x) - \nu : \|x\|_2 \geq r\}$. By assumption, there exists $r_\eta > 0$ such that $\mathcal{T}_P g \geq \eta\|x\|_2^{q+\theta}$ for $\|x\|_2 > r_\eta$. Define $r_\gamma := r_\eta \vee 2(|\nu|/\eta)^{1/(q+\theta)}$. It is straightforward to check that for $r \geq r_\gamma$, we have $\gamma(r) \geq \eta r^{q+\theta}/2$. By Markov's inequality,

$$Q(\{\|x\|_2 > r\}) \leq \frac{\mathbb{E}_{X\sim Q}\gamma(\|X\|_2)}{\gamma(r)} \leq \frac{\mathbb{E}_{X\sim Q}[\mathcal{T}_P g(X) - \nu]}{\gamma(r)} \leq \frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\gamma(r)}.$$

These observations yield the following estimate of the above integral:

$$\begin{aligned}\int_{\{\|x\|_2 > r\}} \|x\|_2^q \mathrm{d}Q(x) &= r^q Q(\{\|x\|_2 > r\}) + 1_{\{q>0\}}\int_r^\infty Q(\{\|x\|_2 > t^{1/q}\})\mathrm{d}t \\ &\leq r^q \frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\gamma(r)} + 1_{\{q>0\}}\int_r^\infty \frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\gamma(t)}\mathrm{d}t \\ &\leq 2\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta r^\theta} + 21_{\{q>0\}}\int_r^\infty \frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta t^{1+\theta/q}}\mathrm{d}t \\ &\leq 2\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta}\left(\frac{1}{r^\theta} + \frac{q}{\theta}\frac{1}{r^{\theta/q}}1_{\{q>0\}}\right) \\ &\leq 2\left(1 + \frac{q}{\theta}\right)\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta}\frac{1}{r^{\theta\wedge\theta/q}}\end{aligned}$$

where we assume $r \geq r_\gamma \vee 1$. Therefore, for $\epsilon > 0$, by taking sufficiently large $r_\epsilon \geq 1$ such that

$$2\left(1 + \frac{q}{\theta}\right)\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta}\frac{1}{r_\varepsilon^{\theta\wedge\theta/q}} \leq \varepsilon \text{ and } r_\varepsilon \geq r_\gamma,$$

we have

$$\int_{\|x\|_2 > r_\epsilon} \|x\|_2^q \mathrm{d}Q \leq \varepsilon.$$

Therefore, the order-q integrability rate $R_q(Q, \varepsilon)$ satisfies

$$R_q(Q, \varepsilon) \leq \left\{2\left(1 + \frac{q}{\theta}\right)\left(\frac{\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{kI}) - \nu}{\eta\varepsilon}\right)\right\}^{\frac{1}{\theta}\vee\frac{q}{\theta}} \vee r_\gamma.$$

For sufficiently small ε, the Stein discrepancy term dominates r_γ. Thus, the claim has been proved. □

6.B.4.2 The DKSD with the tilted linear kernel detects non-uniform integrability

Lemma 6.42 (Tilted linear kernels have the lower bound properties)**.** *Suppose the diffusion targeting P satisfies the dissipativity condition (Condition 6.3) with $\alpha, \beta > 0$ and the coefficient condition (Condition 6.2) with $\lambda_a > 0$ and $q_a \in \{0, 1\}$. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w - u}$ with $q_w \geq 0,\ u \geq 0,$ and $v > 0$. Assume $(q_w - u) < 2\alpha/\lambda_a$ if $q_a = 1$. Let*

$$k(x, x') = w(x)w(x')\langle x, x' \rangle.$$

There exists a function $g \in \mathcal{G}_{kI}$ such that $\|g\|_{\mathcal{G}_{kI}} = \sqrt{D}$ and the corresponding diffusion Stein operator $\mathcal{T}_P$ satisfies

$$\mathcal{T}_P g(x) \geq \nu \text{ for any } x \in \mathbb{R}^D, \text{ and } \liminf_{\|x\|_2 \to \infty} \|x\|_2^{-2(q_w - u + 1)} \mathcal{T}_P g(x) \geq \eta$$

for some $\nu \in \mathbb{R}$ and $\eta > 0$.

Proof. We prove that the function $g(x) = -w(x)x$ satisfies the properties in the statement. Note that the kernel k is the linear kernel tilted by $(v^2 + \|x\|_2^2)^{q_w - u}$.The function g then belongs to the RKHS $\mathcal{G}_{kI}$, since each component is the product of two functions from the RKHSs of the respective kernels. It is straightforward to check $\|g\|_{\mathcal{G}_{kI}} = \sqrt{D}$. Applying the diffusion Stein operator yields

$$\begin{aligned} \mathcal{T}_P g(x) &= w(x) \left(-2\langle b(x), x \rangle - \langle m(x), I \rangle - \frac{2(q_w - u)}{(v^2 + \|x\|_2^2)} \langle m(x), x \otimes x \rangle \right) \\ &= w(x) \left(-\mathcal{A}_P \|x\|_2^2 - \frac{2(q_w - u)\|x\|_2^2}{v^2 + \|x\|_2^2} \frac{\langle \sigma(x)\sigma(x)^\top, x \otimes x \rangle}{\|x\|_2^2} \right). \end{aligned}$$

We first address the case $q_w - u \leq 0$. In this case, we have

$$\mathcal{T}_P g(x) \geq w(x)(\alpha \|x\|_2^2 - \beta),$$

where the inequality follows from Condition 6.3 and the nonnegative second term inside the parentheses above. From this estimate, the following relation holds for $\|x\|_2 > R = \sqrt{\beta/\alpha + 1}$,

$$\mathcal{T}_P g(x) \geq \eta \|x\|^{2(q_w - u + 1)},$$

where $\eta = \left(1 + v^2/R^2\right)^{q_w - u} \alpha\{1 - \beta/(\alpha + \beta)\} > 0$. As $\mathcal{T}_p g(x)$ is continuous, it has the minimum in the centered closed ball of radius R. Thus,

$$\begin{aligned} \mathcal{T}_P g(x) \geq \nu &:= 0 \vee \min_{0 \leq \|x\|_2 \leq R} \mathcal{T}_P g(x) \\ &= -\beta w(0) \end{aligned}$$

We next show the case $q_w - u > 0$. Using Condition 6.3 and the growth condition (Condition

6.2), we obtain

$$\mathcal{T}_P g(x) \geq w(x)\left(\alpha\|x\|_2^2 - \beta - \frac{\lambda_a(q_w-u)}{2}\frac{\|x\|_2^2}{(v^2+\|x\|_2^2)}(1+\|x\|_2^{q_a+1})\right)$$

This estimate provides us the lower bound

$$\mathcal{T}_P g(x) \geq \eta\|x\|_2^{2(q_w-u+1)}$$

for $\|x\|_2 > R_1 = R_0 + 1$,with $\eta = \left(1+v^2/R^2\right)^{q_w-u} f(R_1)$ where

$$f(r) = \left\{\alpha - \frac{1}{r^2}\left(\beta + \frac{\lambda_a(q_w-u)}{2}\frac{r^2}{v^2+r^2}(1+r^{q_a+1})\right)\right\},$$

and R_0 is chosen such that $f(R_0) = 0$. The existence of such R_0 is guaranteed if f is increasing and achieves a positive value; the case $q_a = 0$ automatically satisfies this requirement, whereas the case $q_a = 1$ further requires

$$\alpha > \frac{\lambda_a(q_w-u)}{2}.$$

To show a uniform lower bound, note that with R_1 above, $\mathcal{T}_P g(x) \geq 0$ for $\|x\|_2 > R_1$, and

$$\mathcal{T}_P g(x) \geq -w(R_1)\left\{\beta + \frac{\lambda_a(q_w-u)}{2}\frac{R_1^2}{v^2+R_1^2}(1+R_1^{q_a+1})\right\}$$

for $\|x\|_2 \leq R_1$. □

6.B.4.3 The KSD with the tilted IMQ kernel detects non-uniform integrability

Lemma 6.43. *Suppose the diffusion targeting P satisfies the dissipativity condition (Condition 6.3) with $\alpha, \beta > 0$ and the coefficient condition (Condition 6.2) with $\lambda_a > 0$ and $q_a \in \{0,1\}$. Let $w(x) = (v^2 + \|x\|_2^2)^{q_w-u}$ with $q_w \geq 0,\ u \geq 0,$ and $v > 0$. Assume $(q_w - u) < 2\alpha/\lambda_a$ if $q_a = 1$. Let*

$$k(x,x') = w(x)w(y)(v_0^2 + \|x-x'\|^2)^{-t}$$

for $t \in (0,1)$. Then, there exists a function $g \in \mathcal{G}_{kI}$ such that with any fixed $s \in (0,(t+1)/2)$,

$$\mathcal{T}_P g(x) \geq \nu \text{ for any } x \in \mathbb{R}^D \text{ and } \liminf_{\|x\|\to\infty} \|x\|_2^{-2(q_w-u+s)}\mathcal{T}_P g(x) \geq \eta$$

for some $\nu \in \mathbb{R}$ and $\eta > 0$.

Proof. From the proof of Lemma 16 of Gorham and Mackey [2017], for any fixed $2s \in (0, t+1)$ and $w > v_0/2$, we have that the function

$$g(x) = -2\alpha\frac{x}{(w^2+\|x\|_2^2)^{1-s}}$$

is an element of $\mathcal{G}_{kI}$ with the RKHS norm $\mathcal{D}(w, v_0, s, t)^{1/2} < \infty$ [see Lemma 16 of Gorham and Mackey, 2017, for the norm estimate] The rest of the proof proceeds as in Lemma 6.16. □

6.B.5 Polynomial functions are pseudo-Lipschitz functions

We show that the class $\mathcal{F}_q$ suffices for characterizing convergence in moments.

Lemma 6.44. *Let $q \geq 1$ be an integer. The q-th power $\|x\|_2^q$ of the Euclidean norm is a pseudo-Lipschitz function of order $q-1$. Its pseudo-Lipschitz constant is bounded by $1 \vee q/2$.*

Proof. The case $q = 1$ follows from the triangle inequality. For $q \geq 2$, the claim follows by observing

$$\begin{aligned}
|\|x\|_2^q - \|y\|_2^q| &\leq q \int_0^1 \left| \|tx + (1-t)y\|_2^{q-1} \langle tx + (1-t)y, x - y \rangle \right| \mathrm{d}t \\
&\leq q\|x-y\|_2 \int_0^1 \|tx + (1-t)y\|_2^{q-1} \mathrm{d}t \\
&\leq q\|x-y\|_2 \int_0^1 t\|x\|_2^{q-1} + (1-t)\|y\|_2^{q-1} \mathrm{d}t \\
&= \frac{q}{2}(\|x\|_2^{q-1} + \|y\|_2^{q-1})\|x-y\|_2 \\
&\leq \frac{q}{2}(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1})\|x-y\|_2,
\end{aligned}$$

where we have applied Jensen's inequality to derive the third line. □

Lemma 6.45. *Let $q \geq 1$ be an integer. Let $\boldsymbol{q} = (q_1, \ldots, q_D) \in \{0, \ldots, q\}^D$ be a multi-index such that $\sum_{d=1}^D q_d = q \geq 1$. Then, $x^{\boldsymbol{q}} := \prod_{d=1}^D x_d^{q_d}$ is pseudo-Lipschitz of order $q-1$. Its pseudo-Lipschitz constant $\tilde{\mu}_{\mathrm{pLip}}(x^{\boldsymbol{q}})_{1,q-1}$ is bounded by $1 \vee (2(D-1)+1) \cdot q/2$, and its degree $q-1$ polynomial derivative coefficient $\tilde{\pi}(x^{\boldsymbol{q}})_{q-1,i}$ is bounded by $\max_{d=1,\ldots,D} q_d! \binom{i+D-1}{D-1}$.*

Proof. We first prove the following relationship:

$$|x^{\boldsymbol{q}} - y^{\boldsymbol{q}}| \leq C_D \frac{q}{2}(\|x\|^{q-1} + \|y\|^{q-1})\|x-y\|_2,$$

where $C_D = 2(D-1)+1$. From the proof of Lemma 6.44, for $D = 1$, we have

$$|x^q - y^q| \leq \frac{q}{2}(|x|^{q-1} + |y|^{q-1})|x-y|.$$

For $D > 1$, suppose that the relation is true for $D-1$. Take a multi-index $\boldsymbol{q}$ of size q. For $q = 1$, the claim is true with Lipschitz constant 1. Without loss of generality, we may assume $\|x\|_2 \geq \|y\|_2$. Then, for $q > 1$,

$$\begin{aligned}
&|x^{\boldsymbol{q}} - y^{\boldsymbol{q}}| \\
&= \left| \prod_{d=1}^{D-1} x_d^{q_d} \cdot x_D^{q_D} - \prod_{d=1}^{D-1} x_d^{q_d} \cdot y_D^{q_D} + \prod_{d=1}^{D-1} x_d^{q_d} \cdot y_D^{q_D} - \prod_{d=1}^{D-1} y_d^{q_d} \cdot y_D^{q_D} \right| \\
&\leq \left| \prod_{d=1}^{D-1} x_d^{q_d} \right| \cdot |x_D^{q_D} - y_D^{q_D}| + |y_D^{q_D}| \cdot \left| \prod_{d=1}^{D-1} x_d^{q_d} - \prod_{d=1}^{D-1} y_d^{q_d} \right| \\
&\leq \|x\|^{\sum_{d=1}^{D-1} q_d} \cdot \frac{q_D}{2}(|x_D|^{q_D-1} + |y_D|^{q_D-1})|x_D - y_D|
\end{aligned}$$

$$\begin{aligned}
&+|y_D|\cdot\frac{\sum_{d=1}^{D-1}q_d}{2}\Big(\|x\|_2^{\sum_{d=1}^{D-1}q_d-1}+\|y\|_2^{\sum_{d=1}^{D-1}q_d-1}\Big)\|(x_1,\dots,x_{D-1})-(y_1,\dots,y_{D-1})\|_2\\
&\le\|x-y\|_2\left\{\frac{q_D}{2}\|x\|_2^{\sum_{d=1}^{D-1}q_d}\cdot\big(\|x\|_2^{q_D-1}+\|y\|_2^{q_D-1}\big)\right.\\
&\qquad\qquad\left.+\|y\|_2^{q_D}C_{D-1}\frac{\sum_{d=1}^{D-1}q_d}{2}\Big(\|x\|_2^{\sum_{d=1}^{D-1}q_d-1}+\|y\|_2^{\sum_{d=1}^{D-1}q_d-1}\Big)\right\}\\
&\le\|x-y\|_2\left\{\frac{q_D}{2}2\|x\|_2^{q-1}+C_{D-1}\frac{\sum_{d=1}^{D-1}q_d}{2}\big(\|x\|_2^{q-1}+\|y\|_2^{q-1}\big)\right\}\\
&\le\frac{q}{2}\|x-y\|_2\{(C_{D-1}+2)\|x\|_2^{q-1}+C_{D-1}\|y\|_2^{q-1}\}\\
&\le\frac{q}{2}(C_{D-1}+2)\|x-y\|_2\big(\|x\|_2^{q-1}+\|y\|_2^{q-1}\big).
\end{aligned}$$

Solving $C_D=C_{D-1}+2$ yields $C_D=2(D-1)+1$. Therefore,

$$|x^{\boldsymbol{q}}-y^{\boldsymbol{q}}|\le\frac{q}{2}(2(D-1)+1)\cdot\big(1+\|x\|_2^{q-1}+\|y\|_2^{q-1}\big)\|x-y\|_2.$$

Next we check the degree q polynomial coefficient of the ith derivatives. We assume $q\ge i$ below, as the derivatives are zero otherwise. Note that we have

$$(\nabla^i x^{\boldsymbol{n}})_{l_1,\dots,l_i}=\prod_{d=1}^{D}\frac{n_d!}{(n_d-m_d)!}\cdot x_d^{q_d-m_d}\cdot 1_{\{q_d\ge m_d\}},$$

where $m_d:=\#\{l_d:l_d=d\}$, and

$$\begin{aligned}
\|\nabla^i x^{\boldsymbol{q}}\|_{\mathrm{op}}&=\sup_{\|u^{(d)}\|_2=1}\left|\sum_{\sum m_d=i}\prod_{d=1}^{D}\left(\frac{q_d!}{(q_d-m_d)!}\cdot x_d^{q_d-m_d}1_{\{q_d\ge m_d\}}\right)u_{l_d}^{(d)}\right|\\
&\le\max_d q_d!\sup_{\|u^{(d)}\|_2=1}\sum_{\sum m_d=i}\prod_{d=1}^{D}\big|\|x\|_2^{q_d-m_d}u_{i_d}^{(d)}\big|\\
&\le\max_d q_d!\sup_{\|u^{(d)}\|_2=1}\sum_{\sum m_d=i}\prod_{d=1}^{D}\|x\|_2^{q_d-m_d}\|u^{(d)}\|_2\\
&\le\max_d q_d!\sum_{\sum m_d=i}\|x\|_2^{\sum_d q_d-m_d}\\
&\le\max_d q_d!\binom{i+D-1}{D-1}\|x\|_2^{q-i}.
\end{aligned}$$

Therefore,

$$\tilde{\pi}(x^{\boldsymbol{q}})_{q-1,i=}\sup_{x\in\mathbb{R}^D}\frac{\|\nabla^i x^{\boldsymbol{q}}\|_{\mathrm{op}}}{1+\|x\|_2^{q-1}}\le\max_d q_d!\binom{i+D-1}{D-1}\sup_{x\in\mathbb{R}^D}\frac{\big(\|x\|_2^{q-i}\big)}{1+\|x\|_2^{q-1}}.$$

Consider the function $f(r)=r^m/(1+r^{m+i-1})$ on $[0,\infty)$ with $m>0,i\ge1$. If $i=1$, the function is monotonically increasing and its supremum is $\lim_{r\to\infty}f(r)=1$. If $i>1$, the function is nonnegative, and by taking the derivative, it can be shown that the function takes its maximum at $r^* = \{m/(i-1)\}^{m/(m+i-1)}$ with its value

$$f(r^*) = \frac{\left(\frac{m}{i-1}\right)^{\frac{m}{m+i-1}}}{1+\frac{m}{i-1}} \le \frac{\frac{m}{i-1}}{1+\frac{m}{i-1}} < 1.$$

Thus, we have

$$\tilde{\pi}(x^{\boldsymbol{q}})_{i,q} \le \max_{d=1,\dots,D} q_d! \binom{i+D-1}{D-1}.$$

□

The above result indicates that if we divide a given monomial $x^{\boldsymbol{q}} := \prod_{d=1}^{D} x_d^{q_d}$ by the maximum of $1 \vee (2(D-1)+1) \cdot q/2$ and $\max_{d=1,\dots,D} q_d! \binom{3+D-1}{D-1}$, we have $x^{\boldsymbol{q}} \in \mathcal{F}_q$, where $\mathcal{F}_q$ is the psuedo-Lipschitz class used in (6.4).

Chapter 7

Conclusion and further research

As modern applications involve increasingly complex statistical models, it is critical to provide appropriate tools to investigate their limitations. This thesis addresses two main challenges in assessing generative models: intractability, and interpretability. In Chapters 3, 4, we have treated the question of intractability and developed a test for comparing latent variable models. In an attempt to address the question of interpretability, in Chapter 5, we have introduced a novel test for comparing generative models; in Chapter 6, we have established an interpretable goodness-of-fit measure based on moments. In addition to proposals in each chapter, these works suggest the following two potential research directions. The first direction concerns intractability. The extension of the Kernel Stein Discrepancy to latent variable models enables us to consider a large class of models defined by intractable marginals with unknown normalizing constants. While this extension covers a large class, there are interesting model classes that are ruled out from this framework. An example class is simulation-based models [Cranmer et al., 2020], where the likelihood function is not explicitly given or is challenging to compute. While discrepancy measures such as the MMD or the UME statistic in Chapter 5 apply to this class, these may not be practical, especially when the simulation cost is high. However, as we have seen in Chapter 3, information from the model helps us sample-efficiently detect discrepancies that may be challenging to uncover absent a large number of observed samples. Building upon this idea, it would be interesting to extend evaluation techniques such as the KSD to broader model classes, including implicit models.

A related question arises in approximate Bayesian inference, where evaluating posterior approximations remains challenging. A benefit of the KSD approach is that we can characterize the deviation of an approximation from a given target distribution by exploiting the density function. However, the KSD does not apply if the likelihood or the prior is not computable. Such situations are typical for posterior distributions associated with sequential models (e.g., state space models) [Chopin and Papaspiliopoulos, 2020] or simulation-based models [Lintusaari et al., 2017, Cranmer et al., 2020]. Developing a characterization like the KSD for such distribution classes is an important open question.

In respect of interpretability, the two tests developed in Chapter 5 are based on kernel-based distributional discrepancies, where the user can specify features of interest; the test also allows the user to search for features distinguishing two models. A limitation of the discrepancy measures used in our tests is that they can only measure the incompatibility of generative processes in terms of the resulting marginal distributions. It would be of practical interest to develop an informative measure of discrepancy for a specific, possibly latent component in a generative process. Exploiting model structures as in Chapter 3 might allow us to achieve this goal, and this task is an important direction to explore.

Bibliography

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. DOVER PUBN INC, June 1965. ISBN 0486612724. URL https://www.ebook.de/de/product/1675514/handbook_of_mathematical_functions_with_formulas_graphs_and_mathematical_tables.html. 154

E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg. Introduction to mixed membership models and methods. In E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*, chapter 1, pages 3–10. Chapman and Hall/CRC, nov 2014. doi: 10.1201/b17520. 46

V. Alba Fernández, M. Jiménez-Gamero, and J. Muñoz Garcia. A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics and Data Analysis*, 52:3730–3748, 2008. 97

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005. ISBN 978-3-7643-2428-5; 3-7643-2428-7. 128

B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, 2016. 111

A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, L. Mackey, C. J. Oates, G. Reinert, and Y. Swan. Stein's method meets computational statistics: A review of some recent developments. To appear in Statistical Science, May 2021. 23

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of The 34th International Conference on Machine Learning*, 2017. 97, 116

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68 (3):337–337, Mar. 1950. 14, 18, 28, 148

J. N. Arvesen. Jackknifing U-statistics. *Annals of Mathematical Statistics*, 40:2076–2100, 1969. ISSN 0003-4851. doi: 10.1214/aoms/1177697287. 56, 77

A. D. Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, (Special Vol. 25A):175–184, 1988. doi: 10.1017/s0021900200040341. 26

L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988. 98

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 15, 29, 33, 52, 100, 120, 126

A. Basilevsky. *Statistical factor analysis and related methods*. John Wiley & Sons, Inc., jun 1994. doi: 10.1002/9780470316894. 25

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7679-7. doi: 10.1007/978-1-4419-9096-9. With a preface by Persi Diaconis. 17, 20

J. Besag. Comments on 'Representations of knowledge in complex systems' by U. Grenander and M.I. Miller. *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)*, 56:591–592, 1994. 40

S. Betsch and B. Ebner. Fixed point characterizations of continuous univariate probability distributions and their applications. *Annals of the Institute of Statistical Mathematics*, 73(1):31–59, nov 2019a. doi: 10.1007/s10463-019-00735-1. 26

S. Betsch and B. Ebner. Testing normality via a distributional fixed point property in the Stein characterization. *TEST*, 29(1):105–138, feb 2019b. doi: 10.1007/s11749-019-00630-0. 26

S. Betsch and B. Ebner. A new characterization of the Gamma distribution and associated goodness-of-fit tests. *Metrika*, 82(7):779–806, jan 2019c. doi: 10.1007/s00184-019-00708-7. 26

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *6th International Conference on Learning Representations , ICLR 2018, Conference Track Proceedings*. 2018. 108, 109

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006. ISBN 0387310738. 13

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003. 15, 25, 46

K. Borgwardt, E. Ghisu, F. Llinares-López, L. O'Bray, and B. Rieck. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends® in Machine Learning*, 13(5-6):531–712, 2020. doi: 10.1561/2200000076. 38, 39

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3), jun 2020. doi: DOI:10.1214/19-AAP1528. 40, 65

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. 27, 39, 41, 42, 66, 67, 81, 82, 84, 98, 99, 101, 102, 106, 111

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976. 98

G. Bresler and D. Nagaraj. Stein's method for stationary distributions of Markov chains and application to Ising models. *Annals of Applied Probability*, 29, 2019. 26, 58

H. Callaert and P. Janssen. The Berry-Esseen theorem for U-statistics. *The Annals of Statistics*, 6(2): 417–421, 1978. 54, 92

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006. ISSN 0219-5305. doi: 10.1142/S0219530506000838. 126

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

L. H. Y. Chen. Poisson approximation for dependent trials. *The Annals of Probability*, 3(3):534–545, 1975. ISSN 0091-1798. doi: 10.1214/aop/1176996359. 26

W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, pages 844–853, 2018. URL `https://proceedings.mlr.press/v80/chen18f.html`. 15, 120

W. Y. Chen, A. Barp, F. Briol, J. Gorham, M. A. Girolami, L. W. Mackey, and C. J. Oates. Stein point Markov chain Monte Carlo. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1011–1021, 2019. 15, 38

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural InformationProcessing Systems*, volume 29, pages 2172–2180, 2016. 113

N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-47845-2. 171

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems*, volume 23, pages 406–414, 2010. 30

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, pages 1972–1980, 2015. 3, 22, 99, 100, 105

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2606–2615, 2016. 14, 15, 23, 25, 26, 28, 29, 98, 99, 100, 101, 120, 138

Cornell University. arXiv Dataset version 68. Kaggle, 2020. URL `https://www.kaggle.com/datasets/Cornell-University/arxiv`. Accessed on 14 March 2022. 50

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, may 2020. doi: 10.1073/pnas.1912789117. 14, 171

A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)*, 79:651–676, 2017. doi: 10.1111/rssb.12183. 65

A. P. Dawid and M. Musio. Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10(2), jun 2015. doi: 10.1214/15-ba942. 31

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, sep 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. 31

S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2): 216–222, Sept. 1987. 40

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, oct 2002. doi: 10.1017/cbo9780511755347. 21, 127, 128, 142

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20:183:1–183:42, 2019. 65

A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886, oct 2015. doi: 10.1007/s00440-015-0673-1. 124

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981. ISSN 0090-5364. 36

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9694–9703, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/3ffebb08d23c609875d7177ee769a3e9-Abstract.html`. 120, 122, 123, 124

T. S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Elsevier, 1983. ISBN 978-0-12-589320-6. doi: 10.1016/B978-0-12-589320-6.50018-6. 45

T. Fernandez, N. Rivera, W. Xu, and A. Gretton. Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 3112–3122, 2020. 14

R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, jul 1925. doi: 10.1017/s0305004100009580. 31

J. Friedman and L. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979. 97

N. Friel and J. Wyse. Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308, 2012. doi: 10.1111/j.1467-9574.2011.00515.x. 27

N. Friel, A. Mira, and C. J. Oates. Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. *Bayesian Analysis*, 11(1):215–245, 2016. ISSN 1936-0975. doi: 10.1214/15-BA948. 31

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research (JMLR)*, 5:73–99, 2004. ISSN 1532-4435. doi: 10.1162/153244303768966111. 14, 20, 22

K. Fukumizu, A. Gretton, B. Schölkopf, and B. K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, volume 21, 2008. 39

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013. doi: 10.5555/2567709.2627677. 20

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, Cambridge, 2017. ISBN 978-1-139-02983-4. doi: 10.1017/9781139029834. 45, 73

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall, 1995. 25

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014. 14, 97, 98, 103, 107, 116

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, volume 28, pages 226–234, 2015. 23, 26, 28, 58, 120, 125

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of The 34th International Conference on Machine Learning*, pages 1292–1301, 2017. 15, 26, 28, 29, 38, 119, 120, 129, 131, 133, 135, 145, 164, 166

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *Annals of Applied Probability*, 2019. 23, 26, 27, 29, 120, 124, 125, 137, 149, 151, 152

F. Götze. On the rate of convergence in the multivariate CLT. *The Annals of Probability*, 19(2):724–739, 1991. ISSN 0091-1798. 26

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2006. 14, 20, 21

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pages 585–592. 2007. 20

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a. 14, 20, 21, 22, 26, 97, 98, 99

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, pages 1205–1213, 2012b. 14, 102, 105, 138

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017. 115

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002. 97

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. pages 609–616. MIT Press, Cambridge, MA, 2008. 97

N. Henze and J. Visagie. Testing for normality in any dimension based on a partial differential equation involving the moment generating function. *Annals of the Institute of Statistical Mathematics*, 72(5): 1109–1136, may 2019. doi: 10.1007/s10463-019-00720-8. 26

N. Henze, S. G. Meintanis, and B. Ebner. Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. *Communications in Statistics - Theory and Methods*, 41(9):1543–1556, may 2012. doi: 10.1080/03610926.2010.542851. 26

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. 2017. 99, 108

L. Hodgkinson, R. Salomone, and F. Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. Jan. 2020. 26, 30, 58, 59

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, Sept. 1948. 14, 22, 29, 54, 55, 112

M. Hoffman, F. Bach, and D. Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23, 2010. 18, 50

M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. 40

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3), jun 2008. doi: 10.1214/009053607000000677. 18

J. Huggins and L. Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems*, volume 31, pages 1899–1909, 2018. 15, 26

P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, July 2020. ISSN 13697412. doi: 10.1111/rssb.12336. 52

H. Jeffreys. *Theory of probability*. Third edition. Clarendon Press, Oxford, 1961. 27

W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, volume 29, pages 181–189. 2016. 3, 14, 15, 22, 97, 98, 99, 100, 102, 110, 138

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *Proceedings of The 34th International Conference on Machine Learning*. 2017a. 15, 102

W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, volume 30, 2017b. 3, 14, 26, 27, 98, 99, 100, 102, 106, 138

W. Jitkrittum, H. Kanagawa, P. Sangkloy, J. Hays, B. Schölkopf, and A. Gretton. Informative features for model comparison. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 26, 27, 84

O. Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, third edition, 2021. ISBN 978-3-030-61871-1; 978-3-030-61870-4. doi: 10.1007/978-3-030-61871-1. 153

L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, mar 2006. doi: https://doi.org/10.1007/s10958-006-0049-2. 40

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773–795, 1995. ISSN 0162-1459. doi: 10.1080/01621459.1995.10476572. 27

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ArXiv e-prints*, Dec. 2014. 112

D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, July 2009. 13

A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 1933. 21

J. Kowalski and X. M. Tu. *Modern applied U-statistics*. John Wiley & Sons, Inc., dec 2007. doi: 10.1002/9780470186466. 14, 22, 66

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 103, 107

M. Ledoux, I. Nourdin, and G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015. ISSN 1016-443X. doi: 10.1007/s00039-015-0312-0. 23

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, third edition, 2005. 27, 36, 101

J. N. Lim, M. Yamada, B. Schölkopf, and W. Jitkrittum. Kernel Stein tests for multiple model comparison. In *Advances in Neural Information Processing Systems*, volume 32, pages 2240–2250, 2019. 52

J. Lintusaari, M. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic Biology*, 66(1):e66–e82, 2017. 14, 97, 171

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 6316–6326, 2020. 14

Q. Liu and J. D. Lee. Black-box importance sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 952–961, 2017. 15

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 276–284, 2016. 14, 15, 23, 25, 26, 27, 28, 29, 71, 98, 99, 100, 101, 105, 120, 138, 140

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 107, 111

J. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837, 2015. 26, 98

Y. Maesono. Asymptotic mean square errors of variance estimators for U-statistics and their Edgeworth expansions. *Journal of the Japan Statistical Society*, 28(1):1–19, 1998. doi: 10.14490/jjss1995.28.1. 36

X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821. IEEE, 2017. 109, 113, 115

B. Matérn. *Spatial variation*, volume 36 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, second edition, 1986. ISBN 3-540-96365-0. doi: 10.1007/978-1-4615-7892-5. With a Swedish summary. 22, 131

T. Matsubara, J. Knoblauch, F. Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. To appear in Journal of the Royal Statistical Society. Series B, (Statistical Methodology), 2021. 15, 38

S. Meyn, R. L. Tweedie, and P. W. Glynn. Markov chains and stochastic stability. 2009. doi: 10.1017/cbo9780511626630. 65

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017. doi: 10.1561/2200000060. 20

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 0001-8678. doi: 10.2307/1428011. 20, 21, 26, 119

R. M. Neal. *Handbook of Markov chain Monte Carlo, Chapter 5*. Chapman and Hall/CRC, 1st edition, 2011. 40

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016. 97

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017. 23, 26, 28, 120

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019. 40

T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19(4):201–209, sep 1975. doi: 10.1007/bf02281970. 18

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989. 15, 25

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2015. 112

R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, volume 29, pages 496–504. 2016. 26

R. Rehurek and P. Sojka. Gensim–Python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011. 50

G. Reinert and N. Ross. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *Annals of Applied Probability*, 29, 2019. 26, 58

M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal thinning of MCMC output. To appear in Journal of the Royal Statistical Society. Series B, (Statistical Methodology), 2021. 15, 38

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(none), jan 2004. doi: 10.1214/154957804100000024. 65

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341, dec 1996. 40

P. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society B*, 67(4):515–530, 2005. 97

N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011. 23, 26

S. T. Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632. The MIT Press, 1997. 25

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29, June 2016. 98

S. M. Schennach and D. Wilhelm. A simple parametric model selection test. *Journal of the American Statistical Association*, 112(520):1663–1674, jul 2017. doi: 10.1080/01621459.2016.1224716. 36

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461 – 464, 1978. doi: 10.1214/aos/1176344136. 27

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009. 57, 112

S. Shao, P. E. Jacob, J. Ding, and V. Tarokh. Bayesian model comparison with the Hyvärinen score: computation and consistency. *Journal of the American Statistical Association*, 114(528):1826–1837, mar 2019. doi: 10.1080/01621459.2018.1518237. 31

J. Shi, Y. Zhou, J. Hwang, M. K. Titsias, and L. Mackey. Gradient estimation with discrete Stein operators. Feb. 2022. 26, 30, 58, 59

C.-J. Simon-Gabriel, A. Barp, B. Schölkopf, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. *under review*, June 2020. 14

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, jun 1948. doi: 10.1214/aoms/1177730256. 21

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007. 20, 99

L. F. South, M. Riabiz, O. Teymur, and C. J. Oates. Postprocessing of MCMC. *Annual Review of Statistics and Its Application*, 9(1), nov 2021. doi: 10.1146/annurev-statistics-040220-091727. 29

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research (JMLR)*, 11: 1517–1561, 2010. ISSN 1532-4435. 14, 19, 20, 21, 22

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011. 19, 23, 29, 99

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6, 2012. doi: 10.1214/12-ejs722. 21, 22, 69

A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. *Advances in Neural Information Processing Systems*, 30, May 2017. 98

C. Stein. 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 (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, 1972. 14, 23, 26

C. Stein. *Approximate computation of expectations*, volume 7 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986. ISBN 0-940600-08-0. 26

M. L. Stein. *Interpolation of spatial data*. Springer Series in Statistics. Springer-Verlag, New York, 1999. ISBN 0-387-98629-4. doi: 10.1007/978-1-4612-1494-6. Some theory for Kriging. 22, 131

I. Steinwart and A. Christmann. *Support vector machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. 17, 18, 19, 28, 58, 94

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. 2016. 14, 102, 138

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 107, 108

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004. 97

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005. 98

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B, (Statistical Methodology)*, 61(3):611–622, 1999. 25

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, oct 2000. doi: https://doi.org/10.1017/CBO9780511802256. 35, 89, 90, 91, 94, 95

C. Villani. *Optimal transport: old and new*. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. 40, 127

R. von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18(3):309–348, sep 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730385. 14, 22

F.-Y. Wang. Exponential contraction in Wasserstein distances for diffusion semigroups with negative curvature. *Potential Analysis*, 53(3):1123–1144, feb 2020. doi: 10.1007/s11118-019-09800-z. 124

S. Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013. 27

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, dec 2004. doi: 10.1017/cbo9780511617539. 121, 147

W. Xu and T. Matsuda. A Stein goodness-of-fit test for directional distributions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 320–330, 2020. 39, 52

M. Yamada, D. Wu, Y.-H. H. Tsai, I. Takeuchi, R. Salakhutdinov, and K. Fukumizu. Post selection inference with incomplete maximum mean discrepancy estimator. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9*, 2019. 98

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5561–5570, 2018. 14, 15, 25, 26, 30, 58