Bridging the Gap Between Modelling and Computation in Bayesian Statistics

Takuo Matsubara

Thesis Submitted for the Degree of
Doctor of Philosophy

School of Mathematics, Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

July 2023

I dedicate this thesis to my loving parents and family.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. Parts of this dissertation contain work based on collaborative research and have been published or in submission, in which the extent of my contribution is outlined below:

1. Mastubara, T., Knoblauch, J., Briol, F-X., Oates, C. J. (2022) Robust Generalised Bayesian Inference for Intractable Likelihoods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84(3):997-1022.

   In this collaborative work, I led designing of the research project, theoretical analysis, and writing of the paper. The numerical experiments were led by Chris Oates. All authors contributed to each aspect of the project.

2. Mastubara, T., Knoblauch, J., Briol, F-X., Oates, C. J. (2023) Generalised Bayesian Inference for Discrete Intractable Likelihoods. *In Revision at Journal of the American Statistical Association.*

   In this collaborative work, I led designing of the research project, all analysis, and writing of the paper. The article were mainly written in collaboration of Chris Oates and myself. All authors contributed to each aspect of the project.

Other work based on collaborative research conducted during my study is mentioned but not included as parts of this dissertation:

3. Mastubara, T., Oates, C. J., Briol, F-X. (2021) The Ridgelet Prior: A Covariance Function Approach to Prior Specification for Bayesian Neural Networks. *Journal of Machine Learning Research* 22:1-57.

4. Mastubara, T., Mudd, R., Tax N., Guy, I. (2023) TCE: A Test-Based Approach to Measuring Calibration Error. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, PMLR 216:1390-1400, 2023.

Takuo Matsubara
July 2023

Acknowledgements

I am sincerely grateful to my advisor Chris Oates for his generous supports over my entire PhD odyssey and always being my best role model not only as an academic but also as a person. It was a tremendous privilege that I could pursue four years of my exhilarating PhD endeavour with Chris. I would also like to express my heartfelt gratitude to François-Xavier Briol who acted as an informal second advisor. His resourceful suggestions and advice at every important timing of my PhD study aided me in taking concrete steps toward academic career development. My profound appreciation further goes to my esteemed collaborators Jeremias Knoblauch, Richard Mudd, Niek Tax, and Ido Guy, who have given me numerous opportunities to develop my skill and knowledge. Finally, I would like to extend my gratitude to all researchers, students, and friends, with whom I have been honoured to be acquainted through my PhD study. This PhD research was funded by The Alan Turing Institute, to whom I am genuinely grateful for supporting me as Turing doctoral student throughout my PhD study and expanding the possibilities for my future.

Abstract

Models that involve intractable normalising constants represent a major computational challenge to statistical inference, since the computation of intractable normalising constants requires numerical integration of complex functions over large or possibly infinite sets, which can be impractical. In particular, Bayesian inference for intractable models demands a specially tailored algorithm to bypass evaluation of two nested intractable normalising constants originating from posterior and model simultaneously. This thesis addresses this computational challenge through the development of a novel generalised Bayesian inference approach built on a Stein discrepancy, called SD-Bayes. Generalised Bayesian inference updates prior beliefs using a loss function, rather than a likelihood, and can therefore be used to confer desirable properties to resulting generalised posteriors, such as robustness to model misspecification. In this context, the Stein discrepancy selected as the loss function circumvents evaluation of normalising constants of models and produces generalised posteriors that are accessible using standard Markov chain Monte Carlo algorithms. On a theoretical level, we show posterior consistency, asymptotic normality, and global bias-robustness of generalised posteriors. It is shown that generalised posteriors equipped with global bias-robustness demonstrate a strong insensitivity to an irrelevant outlier mixed in data, that is, a simple yet common setting of model misspecification. For intractable models in continuous domains, we derive a useful special case of the Stein discrepancy, called kernel Stein discrepancy, to be combined with SD-Bayes. The resulting SD-Bayes demonstrates strong global bias-robustness and enables fully conjugate inference for exponential family models. We provide numerical experiments on a range of intractable distributions, including applications to kernel-based exponential family models and non-Gaussian graphical models. For intractable models in discrete domains, we establish another useful special case of the Stein discrepancy, called discrete Fisher divergence, to be combined with SD-Bayes. The resulting SD-Bayes benefits from its efficient computational cost and absence of user-specified hyperparameters that can be difficult to choose in the discrete case. In addition, a new approach to calibration of generalised posteriors through optimisation is considered, independently of SD-Bayes. Applications are presented on lattice models for discrete spatial data and on multivariate models for count data, where in each case the methodology facilitates generalised Bayesian inference at efficient computational cost.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Bayesian analysis has been adopted in diverse academic and industrial fields as a framework for coherent decision-making based on observations and one's belief expressed as a model to describe a phenomenon of interest. Over the last few decades, rapid growth in computational capacity has broadened the subjects of statistical modelling to highly complex phenomena. For example, it has become increasingly ubiquitous to replace some (or all) parts of decision-making procedures with certain predictive models—even in sensitive domains such as healthcare (Topol, 2019)—and to deploy enhanced engineering systems predicated on the use of data-driven models (e.g. Girolami, 2020). An accurate description of sensitive, intricate phenomena often necessitates the use of complex models that are highly structured and high dimensional. A heavy computational burden of complex models is placed especially on inference of them, which correspondingly emphasises a computational challenge of Bayesian inference. A number of cases has emerged where Markov Chain Monte Carlo (MCMC) algorithms—the gold standard computation tools for Bayesian inference to date since the 1990s—are hard to run in realistic time, unless a gigantic computation cluster is available, or even technically impossible to apply in the first place. In the "computer" age that has been further elevated today, Bayesian statisticians appear to embark on new horizons of inference for highly complex models, as represented by such directions as approximate Bayesian computation (Marin et al., 2012) and more broadly approximate Bayesian inference (Martin et al., 2023). The aim of this thesis is to advance the quest for the emerging frontier of inference for complex models in the particular context of *intractable* models.

A considerable portion of complex models falls into a class called intractable model, that is, a model whose likelihood is analytically and computationally hard to access. To be exact, there exists two major levels of intractability: (i) the explicit form of the likelihood is not available entirely; (ii) the explicit form of the likelihood is partly available up to the normalising constant. In the former case, a model is essentially black-box with only its parameter space known and its sampling feasible. Inference in this case is often referred to as *simulator-based inference* (Cranmer et al., 2020), in which samples from the model are leveraged as an alternative means to the inaccessible likelihood. A focus of this thesis is on the latter case, in which a modeller explicitly designs a model, but it is too complex to keep the normalising constant analytical. A model p_θ typically admits a decomposition $p_\theta(x) = q_\theta(x)/Z(\theta)$ based on some non-normalised function q_θ and the normalising constant $Z(\theta) = \int_{\mathcal{X}} q_\theta(x)\mathrm{d}x$. For a modeller, the design of the model p_θ is then reduced to the

design of the function q_θ with a desideratum for the associated normalising constant $Z(\theta)$ to be analytically known. It is, however, immediately deduced that $Z(\theta)$ is only empirically known if a complex form of q_θ is necessitated to describe an intricate phenomenon. In addition, it readily becomes difficult to estimate the integral $Z(\theta)$ when q_θ is complex or x is high dimensional. Such intractable models p_θ appear in many important applications, including spatial models (Besag, 1974, 1986; Diggle, 1990), exponential random graph models (Park and Haran, 2018a), gene expression models (Jiang et al., 2021), hidden Potts models for satellite data (Moores et al., 2020), count data models (Inouye et al., 2017), and energy-based models versatile in machine learning (Lecun et al., 2006).

For most inference methodologies, the normalising constant $Z(\theta)$ is essential because $Z(\theta)$ is, in fact, a function of the parameter θ despite its designation as "constant". For example, the maximum likelihood estimator of the model p_θ requires the normalising constant $Z(\theta)$ to be explicitly available. The intractability of the model p_θ causes a critical challenge, particularly in Bayesian inference, by turning its posterior *doubly intractable* (Murray et al., 2006). Given data $\mathcal{D} = \{x_i\}_{i=1}^n$, a posterior π_n of a model p_θ is defined by

$$\pi_n(\theta) = \frac{1}{Z(\mathcal{D})} \exp\left(\sum_{i=1}^n \log p_\theta(x_i)\right) \pi(\theta) = \frac{1}{Z(\mathcal{D})} \exp\left(\sum_{i=1}^n \log q_\theta(x_i) - \log Z(\theta)\right) \pi(\theta)$$

where π denotes a prior over the parameter space Θ and $Z(\mathcal{D})$ denotes a normalising constant of the posterior itself. In general, the normalising constant $Z(\mathcal{D})$ of the posterior is intractable—i.e. not analytical and hard to compute—regardless of intractability of the model p_θ, unless any conjugate prior exists. Standard MCMC algorithms are designed to sample from the posterior π_n while circumventing an evaluation of the normalising constant $Z(\mathcal{D})$. However, if the model p_θ is intractable, the posterior π_n includes another intractable component $Z(\theta)$, the normalising constant of the model p_θ. Standard MCMC algorithms cannot be used in this setting, since they require explicit evaluation of all the components dependent on θ. It may be possible to run them in an ad-hoc manner by substituting some estimate for $Z(\theta)$ at each θ, but such an ad-hoc approach easily becomes unrealistic due to the need of an accurate estimation of $Z(\theta)$ at every iterative step of MCMC over different θ, which can be tens of thousands in total. For these reasons, Bayesian inference for intractable models typically entails a certain elegant algorithm that circumvents evaluation of both $Z(\mathcal{D})$ and $Z(\theta)$ simultaneously.

A classical approach to inference for intractable models, called *pseudo-likelihood* approach, was pioneered by Besag (1974). Certain types of model structures admit the mean field approximation of the original likelihoods, where the original model is approximated by the product of multiple conditional p.d.f.s that are all tractable. In the pseudo-likelihood approach, inference is performed using such a closed-form approximation of the intractable likelihood. In the 1990s starting with e.g. Gelfand and Smith (1990), Bayesian statistics underwent a major shift from conjugate inference to flexible inference powered by MCMC algorithms, whose original invention may be further traced back to Metropolis et al. (1953).

As entering into the age where computational capacity started blooming, advanced designs of MCMC algorithms geared towards inference of complex models were naturally driven. In the 2000s, multiple new classes of MCMC algorithms tailored to intractable models were proposed, where the high-level approach is to incorporate either sampling from models or unbiased estimation of likelihoods into the MCMC procedures (Andrieu and Roberts, 2009; Lyne et al., 2015; Murray et al., 2006; Møller et al., 2006). Despite the ingenuous convergence theories that underpin these MCMC algorithms, the need of sampling from models or unbiased estimation of likelihoods at every step of MCMC can, however, become impractical for today's ever-evolving models. The practical utility of approximate Bayesian computation based fully on simulation also came to gain attention in wide application domains (e.g. Beaumont et al., 2002). Its theoretical understanding has been one of the central topics in approximate Bayesian inference to date (Frazier et al., 2018, 2020).

In the 2010s, Bissiri et al. (2016) coined a term, *generalised Bayesian inference*, as a designation of approaches that define a pseudo-posterior using an arbitrary loss function $D_n(\theta)$ of a model p_θ in lieu of the likelihood:

$$\pi_n^D(\theta) \propto \exp\left(-\beta D_n(\theta)\right) \pi(\theta)$$

where $\beta \in (0, \infty)$ is a hyperparameter commonly used to adjust the scale of the loss $D_n(\theta)$. The standard posterior π_n is recovered by selecting a negative log-likelihood loss $D_n(\theta) = -\sum_{i=1}^n \log p_\theta(x_i)$ with $\beta = 1$. Studies on generalised Bayesian inference have been motivated independently of intractable models, and it has rarely been considered in that context. Our focus is to establish a novel Bayesian approach to intractable models based on generalised Bayesian inference. Concretely, we define a novel generalised posterior built upon a loss called *Stein discrepancy*, which has a number of appealing properties to intractable models. Strikingly, the resulting posterior is no longer doubly intractable despite the use of intractable models, and can therefore be efficiently computed by any standard MCMC algorithms. In contrast to existing works, this requires neither approximation of models nor sampling from them by the virtue of the Stein discrepancy that measures a fit of a model p_θ without knowing the normalising constant $Z(\theta)$.

Chief concerns associated with the use of complex models are not limited to just computation. The high intricacy of interested phenomena means the difficulty of modelling, by which one's model involves an increased risk of describing only some aspects of the phenomenon well and the others poorly, deviating from the ideal description of the phenomenon. Indeed, statistical modelling often deviates from the idealised approach of fine-tuned, expertly-crafted descriptions of real-world phenomena, in favour of default models fitted to a large dataset. If the default model is a good approximation to the data-generating mechanism, this strategy can be successful, but otherwise things can quickly go awry (Grünwald, 2012). Consequently, the reliability of outcomes of Bayesian inference, which has been a long-standing subject of study (Insua and Ruggeri, 2000), has become even more critical today. A better understanding of when outcomes of Bayesian

inference can be unreliable, and establishing *robust* Bayesian methodologies for that cases, are pressing issues for complex models in tandem with the computational challenge. Generalised Bayesian inference (Bissiri et al., 2016) was originally motivated in this context of model misspecification, and in particular using divergence-based loss has been shown to mitigate some of the risks involved when a model is misspecified (Jewson et al., 2018). Unlike other robust modelling strategies, these methods do *not* change the statistical model. Instead, they change how the model's parameters are scored, affecting how "good" parameter values are discerned from "bad" ones. This is a key practical advantage, as it implies that such strategies do not require precise knowledge about how the model is misspecified.

This thesis aims to contribute to the frontier of Bayesian methodologies to intractable models, adding a new line of approach based on generalised Bayesian inference and Stein discrepancy. This is the first generalised Bayesian approach considered in the context of intractable likelihoods. Stein discrepancy is a nascent class of statistical divergences, and there exists a number of different forms of Stein discrepancies to use. We begin with constructing our generalised posterior using a Stein discrepancy in the most abstract form. We then establish useful theoretical underpinnings of generalised posteriors in the aim of providing certain types of assurance for them to operate well. Subsequently, we derive novel concrete forms of the Stein discrepancy that are particularly useful for intractable models in continuous and discrete domains, respectively. The generalised posteriors resulting from these concrete Stein discrepancies are assessed both theoretically and empirically. The detailed contributions of this thesis are summarised as follows:

Chapter 3. Generalised Bayesian Inference for Intractable Models

1. We propose a novel generalised Bayesian inference framework, called *SD-Bayes*, that selects a Stein discrepancy as the loss.

2. We establish *posterior consistency* and the *Bernstein-von Mise theorem* of generalised posteriors. These results confer an appealing regularity of generalised posteriors in the limiting regime. In particular, the former implies that a generalised posterior concentrates at the same limiting point as the corresponding frequentist estimator. The latter further implies that a generalised posterior is asymptotically normal.

3. We establish a rigorous criterion of generalised posteriors to be robust against model misspecification, called *global bias-robustness*, in a simplified yet common setting where model misspecification is caused by an outlier mixed in data. Any generalised posterior with this property is equipped with a guaranteed insensitivity to outliers.

4. It is a standard practice to calibrate the scaling parameter β to adjust the scale of credible regions of generalised posteriors. We discuss a novel calibration algorithm of β based on minimisation of a statistical divergence between a generalised posterior and an approximated sampling distribution of the corresponding frequentist estimator.

Chapter 4. Case I: Approach to Continuous Intractable Models

1. We consider a particular Stein discrepancy, called the kernel Stein discrepancy (KSD), with a new widely applicable formulation derived. We propose the SD-Bayes methodology resulting from the use of KSD, called KSD-Bayes, that is (not limited to but) appealing to intractable models in continuous domains such as $\mathbb{R}^d$. It is demonstrated that KSD-Bayes is computable by any standard MCMC algorithms.

2. We show that KSD-Bayes, strikingly, achieves fully conjugate inference for intractable models in exponential family, so MCMC algorithms are not required.

3. We illustrate that KSD-Bayes satisfies all the theoretical properties established in Chapter 3. In particular, KSD-Bayes can be highly robust to an outlier mixed in data under an appropriate choice of kernel, suggested by our theoretical analysis.

4. KSD-Bayes and its robustness are assessed by four distinct experiments of continuous intractable models, including random graph models.

Chapter 5. Case II: Approach to Discrete Intractable Models

1. We derive a novel formulation of a discrete version of the Fisher divergence, called the discrete Fisher divergence (DFD), and establish that DFD is a Stein discrepancy. We then consider the SD-Bayes methodology resulting from the use of DFD, called DFD-Bayes accordingly, as an appealing approach to intractable models in discrete spaces. DFD-Bayes is computable by any standard MCMC algorithms.

2. We present a set of practical advantages of DFD over KSD that is attractive especially in the discrete case, including the computational cost and the independence of kernel.

3. We demonstrate posterior consistency and the Bernstein–von Mises theorem for DFD-Bayes. Furthermore, we establish a theoretical connection between DFD and KSD though the lens of Stein discrepancy.

4. DFD-Bayes and its difference from the other posteriors are assessed by three distinct experiments of discrete intractable models, including the multivariate count data.

The rest of this thesis is structured as follows: Chapter 2 introduces preliminary notions used in this thesis. Chapter 3 establishes SD-Bayes and provides a set of the aforementioned theoretical underpinnings. Chapter 4 concretises the SD-Bayes methodology with KSD in case of continuous intractable models, followed by Chapter 5 that considers DFD as a main choice in case of discrete intractable models. The theoretical and empirical assessments are provided in both cases. This thesis is concluded by Chapter 6 that recaps our main contributions and open avenues for future research.

Chapter 2. Preliminary

In this chapter, we briefly recap existing Bayesian methodologies for intractable models, and introduce the main tools to construct our methodology. Section 2.1 introduces main notations used in the rest of this thesis. Section 2.2 categorises existing Bayesian methodologies for intractable models into four classes. Section 2.3 describes the framework of generalised Bayesian inference, which forms the methodological basis of our approach, SD-Bayes, in Chapter 3. Section 2.4 then introduces Stein discrepancy, which plays a central role in SD-Bayes. Finally, Section 2.5 recaps the notion of reproducing kernel Hilbert space (RKHS) used to construct KSD in Chapter 4.

2.1. Setting and Notation

Let $\mathcal{X}$ be a locally compact Hausdorff space. Let $\mathcal{P}(\mathcal{X})$ denote a set of all Borel probability measures $\mathbb{P}$ on $\mathcal{X}$. A Dirac measure at x is denoted $\delta_x \in \mathcal{P}(\mathcal{X})$. If $\mathcal{X}$ is equipped with a reference measure, we abuse notation for a p.d.f. p by writing $p \in \mathcal{P}(\mathcal{X})$ to indicate that the corresponding distribution $\mathbb{P}$ is in $\mathcal{P}(\mathcal{X})$. The Euclidean norm on $\mathbb{R}^m$ is denoted $\|\cdot\|$. Let $\mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ be the set of all Borel probability measures $\mathbb{P}$ supported on $\mathbb{R}^d$, admitting a positive p.d.f. p and continuous partial derivatives $x \mapsto (\partial/\partial x_{(i)})p(x)$. For $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, denote by $L^q(\mathcal{X}, \mathbb{P})$ the Lebesgue space of measurable functions $f : \mathcal{X} \to \mathbb{R}$ s.t. $\|f\|_{L^q(\mathcal{X},\mathbb{P})} := (\int_{\mathcal{X}} |f|^q \mathrm{d}\mathbb{P})^{1/q} < \infty$ in which two elements $f, g \in L^q(\mathcal{X}, \mathbb{P})$ are identified if they are $\mathbb{P}$-almost everywhere equal. The set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ is denoted $C(\mathcal{X})$. We denote by $C_b^1(\mathbb{R}^d)$ the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $x \mapsto (\partial/\partial x_{(i)})f(x)$ are bounded and continuous on $\mathbb{R}^d$. We also denote by $C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)$ the set of bivariate functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $(x, x') \mapsto (\partial/\partial x_{(i)})(\partial/\partial x'_{(j)})f(x, x')$ are bounded and continuous on $\mathbb{R}^d \times \mathbb{R}^d$. For an arbitrary set $\mathcal{S}(\mathcal{X})$ (or $\mathcal{S}(\mathcal{X}, \mathbb{P})$) of functions $f : \mathcal{X} \to \mathbb{R}$, denote by $\mathcal{S}(\mathcal{X}; \mathbb{R}^k)$ (or $\mathcal{S}(\mathcal{X}, \mathbb{P}, \mathbb{R}^k)$) the set of $\mathbb{R}^k$-valued functions whose components belong to $\mathcal{S}(\mathcal{X})$ (or $\mathcal{S}(\mathcal{X}, \mathbb{P})$). We abbreviate a domain $\mathcal{X}$ of the Lebesgue space if $\mathcal{X}$ is clear from context, e.g., $L^q(\mathbb{P})$. Let ∇ and $\nabla\cdot$ be the gradient and the divergence operators in $\mathbb{R}^d$. For functions with multiple arguments, we sometimes use subscripts to indicate the argument to which the operator is applied, e.g., $\nabla_x f(x, y)$. For f an $\mathbb{R}^d$-valued function, $[\nabla f(x)]_{(i,j)} := (\partial/\partial x_{(i)})f_{(j)}(x)$ and $\nabla \cdot f(x) := \sum_{i=1}^d (\partial/\partial x_{(i)})f_{(i)}(x)$. For f an $\mathbb{R}^{d\times d}$-valued function, $[\nabla f(x)]_{(i,j,k)} := (\partial/\partial x_{(i)})f_{(j,k)}(x)$ and $[\nabla \cdot f(x)]_{(i)} := \sum_{j=1}^d (\partial/\partial x_{(j)})f_{(i,j)}(x)$. Hereafter $\mathcal{X}$ is used to denote a space in which data are contained. Let $\Theta \subseteq \mathbb{R}^p$ be Borel, in which a parameter of interest θ is contained.

2.2. Intractable Models and Bayesian Methodologies

A model p_θ typically admits a decomposition $p_\theta(x) = q_\theta(x)/Z(\theta)$ based on some non-normalised function q_θ and the normalising constant $Z(\theta) = \int_{\mathcal{X}} q_\theta(x)\mathrm{d}x$. A focus of this thesis is on intractable models whose non-normalised function q_θ is available but normalising constant $Z(\theta) = \int_{\mathcal{X}} q_\theta(x)\mathrm{d}x$ admits neither any analytical solution nor any sufficiently accurate approximation. Such intractability of a model p_θ causes a severe challenge in Bayesian inference due to explicit dependence of the posterior π_n on the model normalising constant $Z(\theta)$:

$$\pi_n(\theta) \propto \exp\left(\sum_{i=1}^{n} \log q_\theta(x_i) - \log Z(\theta)\right)\pi(\theta).$$

Standard MCMC algorithms require $Z(\theta)$ to be explicitly evaluated at each iterative step over different values of θ. Thus, Bayesian inference for intractable models typically entails a tailored algorithm to circumvent evaluation of the inaccessible component $Z(\theta)$. The aim of this section is to briefly review existing Bayesian methodologies for intractable models. Frequentist methodologies for intractable models are not discussed, where we refer the reader to e.g Hyvärinen (2005); Takenouchi and Kanamori (2017).

Approximate Likelihood Faced with an intractable model, a pragmatic approach is simply to employ standard Bayesian inference with a tractable approximation to the likelihood (e.g. Bhattacharyya and Atchade, 2019). A classical example of approximate likelihood is the pseudo-likelihood of Besag (1974), which replaces the joint probability mass function of the data with a product of conditional probability mass functions, each of which is sufficiently low-dimensional (or otherwise tractable enough) to permit normalising constants to be computed. Generalisations of this approach are sometimes referred to as composite likelihood (Varin et al., 2011). These approximations are usually model-specific, and analysis of the approximation error may be difficult in general (Lindsay et al., 2011).

Simulation-Based Methods One class of intractable statistical models that has been explored in detail are models for which it is possible to simulate data x conditional on the parameter θ. A well-known approach to inference in this class of models is the exchange algorithm of Møller et al. (2006) and Murray et al. (2006), which constructs a Markov chain on an extended state space for which the standard Bayesian posterior occurs as a marginal. Simulation of the Markov chain requires both exact simulation from the statistical model and evaluation of $\tilde{p}_\theta(x)$. Further methodological development has been focused on removing the requirement to evaluate $\tilde{p}_\theta(x)$, with approximate Bayesian computation (Frazier et al., 2018; Marin et al., 2012), Bayesian synthetic likelihood (Frazier et al., 2022; Price et al., 2018), MMD-Bayes (Cherief-Abdellatif and Alquier, 2020; Pacchiardi and Dutta, 2021) and the posterior bootstrap (Dellaporta et al., 2022) emerging as likelihood-free methods, which require only that data can be simulated. Unfortunately, for many statistical models

of discrete data for example, exact simulation (the state-of-the-art being e.g. Propp and Wilson, 1998) from the model is impractical.

Markov Chain-Based Methods Another pragmatic approach is to substitute exact simulations with approximate simulations, such as obtained from a Markov chain that leaves the posterior invariant. This idea has been demonstrated to work in specific instances; see Caimo and Friel (2011); Everitt (2012); Liang (2010) or the review of Park and Haran (2018b). The main drawback of these approaches, as far as this thesis is concerned, is that they require the design of a rapidly mixing Markov chain on a possibly large or infinite set. As such, these methods require bespoke implementations for each class of statistical model considered, and for many models of interest appropriate Markov chains have yet to be developed. Thus, Markov chain-based methods do not represent a general solution to intractable likelihoods.

Plugin-Based Methods The pseudo-marginal approach justifies replacing the intractable likelihood $p_\theta(x)$ with a positive unbiased estimator $\hat{p}_\theta(x)$ of the likelihood in the context of a Metropolis–Hastings algorithm (Andrieu and Roberts, 2009). The practical difficulty of this approach is to construct a positive unbiased estimator. Lyne et al. (2015) proposed the Russian roulette estimator for intractable statistical models, a simulation technique from the physics literature (Carter and Cashwell, 1975) which involves random truncation of the sum (or of an integral in the continuous context) defining the normalising constant. The Russian roulette estimator is unbiased but is not guaranteed to be positive, meaning that post hoc re-weighting of the Markov chain sample path is required. The ergodicity of Russian roulette has not, to the best of our knowledge, been theoretically studied (see the discussion in Wei and Murray, 2017). Further, the mixing time of the Markov chain is known to be sensitive to the variance of $\hat{p}_\theta(x)$, which can be large for estimators based on random truncation (especially when there is no clear a priori ordering for the summands, which can occur in the discrete context). As such, the pseudo-marginal approach does not at present represent a general computational solution to intractable likelihood.

This thesis aims to add a novel approach based on generalised Bayesian inference to these lines of existing Bayesian methodologies for intractable models.

2.3. Generalised Bayesian Inference

Consider a dataset consisting of independent random variables $\{x_i\}_{i=1}^n$ generated from a data-generating distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. Consider also a statistical model $\mathbb{P}_\theta \in \mathcal{P}(\mathcal{X})$, with the p.d.f. form p_θ, indexed by a parameter of interest $\theta \in \Theta$. The Bayesian statistician elicits a prior $\pi \in \mathcal{P}(\Theta)$, which may reflect *a priori* belief about the parameter $\theta \in \Theta$, and

determines *a posteriori* belief according to

$$\pi_n(\theta) \propto \left(\prod_{i=1}^{n} p_\theta(x_i)\right) \pi(\theta) = \exp\left(\sum_{i=1}^{n} \log p_\theta(x_i)\right) \pi(\theta). \tag{2.1}$$

If one's model is *well-specified*, i.e., there exists $\theta_0 \in \Theta$ for which $\mathbb{P} = \mathbb{P}_{\theta_0}$, the Bayesian belief updating is optimal from an information-theoretic perspective (Williams, 1980; Zellner, 1988). Simply put, it is posited in this case that the model family $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$ contains the correct data-generating mechanism $\mathbb{P}$. Optimal processing of information is a desirable property, but the assumption of adequate prior and model specification is often violated in real-world applications. If one's model is *misspecified*, i.e., there exists no such $\theta_0 \in \Theta$ that $\mathbb{P} = \mathbb{P}_{\theta_0}$, the optimality of the Bayesian belief updating is no longer guaranteed. It has been long recognised in this case that outcomes of the Bayesian belief updating can become unreliable (Grünwald, 2012). This has inspired several lines of research, including strategies for the robust specification of prior belief (Berger et al., 1994), the so-called *safe Bayes* approach (de Heide et al., 2020; Grünwald, 2011, 2012), *power posteriors* (e.g. Holmes and Walker, 2017), *coarsened posteriors* (Miller and Dunson, 2019), *bagged posteriors* (Huggins and Miller, 2020), *ρ-posteriors* (Baraud and Birgé, 2020) and Bayesian inference based on scoring rules (Giummolè et al., 2019).

A particularly versatile approach to robustness to model misspecification, which encompasses most of the above, is generalised Bayesian inference (Bissiri et al., 2016); see also the earlier work of Chernozhukov and Hong (2003). This approach constructs a posterior $\pi_n^D \in \mathcal{P}(\Theta)$ using an arbitrary *loss* function $D_n : \Theta \to \mathbb{R}$ of a model $\mathbb{P}_\theta$, which may be data-dependent, and a scaling parameter $\beta > 0$, according to

$$\pi_n^D(\theta) \propto \exp\left(-\beta D_n(\theta)\right) \pi(\theta). \tag{2.2}$$

The so-called *generalised posterior* π_n^D coincides with the standard posterior π_n when the loss function is the negative log-likelihood $D_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i)$ with $\beta = 1$. As discussed in Bissiri et al. (2016); Knoblauch et al. (2022), generalised Bayesian inference is underpinned by an optimisation-centric interpretation:

$$\pi_n^D = \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\Theta)} \left\{ \beta \, \mathbb{E}_{\theta \sim \rho}\left[D_n(\theta)\right] + \mathrm{KL}(\rho \| \pi) \right\} \tag{2.3}$$

where $\mathrm{KL}(\rho\|\pi)$ denotes the Kullback–Leibler (KL) divergence between two distributions $\rho, \pi \in \mathcal{P}(\Theta)$. This perspective reveals that the standard Bayesian posterior is an implicit commitment to a particular loss function—the negative log-likelihood—and that the weighting constant β controls the influence of this loss relative to the prior π. In particular, under mild conditions, the negative log-likelihood converges to $\mathrm{KL}(\mathbb{P}\|\mathbb{P}_\theta)$ up to some constant independent of θ, which reveals that standard Bayesian posterior concentrates around the value of θ that minimizes the KL divergence between the data-generating

distribution $\mathbb{P}$ and the model $\mathbb{P}_\theta$. In the setting of misspecified models, such concentration is problematic, often leading to over-confident predictions (Bernardo and Smith, 2009).

The use of alternative, divergence-based loss functions has been demonstrated to mitigate the negative consequences of a misspecified statistical model, as pioneered in the work on α- and β-divergences in Ghosh and Basu (2016); Hooker and Vidyashankar (2014) and extended to γ-divergence in Nakagawa and Hashimoto (2020). See also Baraud and Birgé (2020). The properties of the divergence, including any potentially undesirable pathologies associated with it, determine the properties of the generalised posterior (Jewson et al., 2018; Knoblauch et al., 2022). These compelling theoretical results have led to considerable interest in generalised Bayesian inference with divergence-based loss functions, yet the divergences that have been considered to-date cannot be computed in the important setting of intractable model. The aim of this thesis is to open up a new avenue of generalised Bayesian inference for intractable models, capitalising on a nascent class of loss called Stein discrepancy that is appealing to intractable models.

Calibration It is a common practice in the context of generalised Bayesian inference to calibrate the scaling parameter $\beta > 0$ in (2.2) to adjust the gross scale of credible regions produced by generalised posteriors. If a model is misspecified, there is essentially no known optimal way to quantify uncertainty associated with the parameter. If so, it is beneficial to use a credible region that is at least approximately equipped with certain nice properties, such as frequentist coverage. Such a high-level approach was also considered in the context of non-informative prior (Robert, 2007). A typical approach to the calibration is to select β so that a credible region of a generalised posterior approximately resembles a confidence interval of its frequentist counterpart. Lyddon et al. (2019) proposed to align the trace of an asymptotic covariance of a generalised posterior with that of the frequentist counterpart. Denoting by $\widehat{V}_B$ an estimated value of the asymptotic covariance of the generalised posterior and by $\widehat{V}_F$ that of the frequentist counterpart in finite n, the value of β is then selected by $\beta = \mathrm{tr}(V_F)/\mathrm{tr}(V_B)$. The approach by Lyddon et al. (2019) is one of the most straightforward approach with theoretical explicitness. However, estimation of the asymptotic covariance can be highly unstable in high dimensional or small data applications. See also Syring and Martin (2019) for a recent review on the calibration and an alternative approach based on a stochastic sequential update algorithm.

2.4. Stein Discrepancy

In an independent line of research of generalised Bayesian inference, a Stein discrepancy was proposed in Gorham and Mackey (2015) to provide statistical divergences that are both computable and capable of providing various forms of distributional convergence control. Since its introduction, Stein discrepancy has demonstrated utility over a range of statistical applications, including hypothesis testing (Chwialkowski et al., 2016; Liu et al., 2016), parameter estimation (Barp et al., 2019), variational inference (Duncan et al., 2023;

Liu and Wang, 2016), Monte Carlo sampling (Chen et al., 2019), and post-processing of Markov chain Monte Carlo (Riabiz et al., 2022); see Anastasiou et al. (2021) for a recent review.

The approach of Stein discrepancy is based on the method of Stein (1972), which requires the identification of a linear operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{U} \to L^1(\mathcal{X}, \mathbb{Q})$, depending on a probability distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and acting on an arbitrary Banach space $\mathcal{U}$, such that

$$\mathbb{E}_{X\sim\mathbb{Q}}[\mathcal{S}_{\mathbb{Q}}[h](X)] = 0 \quad \forall h \in \mathcal{U}. \tag{2.4}$$

Such an operator $\mathcal{S}_{\mathbb{Q}}$ is called a *Stein operator* and $\mathcal{U}$ is called a *Stein set*. Given a distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, there exists infinitely many operators $\mathcal{S}_{\mathbb{Q}}$ satisfying (2.4). A convenient example is the *Langevin Stein operator* (Gorham and Mackey, 2015), defined for $\mathbb{Q} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and a Banach space $\mathcal{U}$ of differentiable functions $h : \mathbb{R}^d \to \mathbb{R}^d$, as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = h(x) \cdot \nabla \log q(x) + \nabla \cdot h(x) \tag{2.5}$$

where q is the p.d.f. of $\mathbb{Q}$. Under suitable regularity conditions on $\nabla \log q$ and $\mathcal{U}$, the Langevin Stein operator satisfies the zero-identity of (2.4) (Gorham and Mackey, 2015, Proposition 1). Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and a Stein operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{U} \to L^1(\mathcal{X}, \mathbb{Q})$ whose image is contained in $L^1(\mathcal{X}, \mathbb{P})$, the most abstract form of Stein discrepancy is defined as

$$\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) := \sup_{\|h\|_{\mathcal{U}}\leq 1} \left| \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] - \mathbb{E}_{X\sim\mathbb{Q}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] \right| = \sup_{\|h\|_{\mathcal{U}}\leq 1} \left| \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] \right|, \tag{2.6}$$

where the last equality follows directly from the zero-identity (2.4). Under mild assumptions, the Stein discrepancy defines a statistical divergence between two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$, meaning that $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) \geq 0$ with equality if and only if $\mathbb{P} = \mathbb{Q}$; see Proposition 1 and Theorem 2 in Barp et al. (2019) for example. Under slightly stronger assumptions, the Stein discrepancy provides convergence control, meaning that a sequence $(\mathbb{P}_n)_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{X})$ converges in a specified sense to $\mathbb{Q}$ whenever $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}_n) \to 0$; see Gorham and Mackey (2015, Theorem 2, Proposition 3) and Gorham and Mackey (2017, Theorem 8, Proposition 9).

In this thesis, the Stein discrepancy is used as a loss $\mathrm{SD}(\mathbb{P}_\theta\|\mathbb{P}_n)$ between a model $\mathbb{P}_\theta$ and an empirical distribution $\mathbb{P}_n$. An important property of the Stein discrepancy that we exploit in this context is that, unlike other divergences, the Stein discrepancy can often be computed without a normalising constant of $\mathbb{P}_\theta$. For example, in case of continuous domain $\mathbb{R}^d$, the Langevin Stein operators in (2.5) depend on $\mathbb{P}_\theta$ only through the log-derivative $\nabla \log p_\theta$, which can be computed even when p_θ is an intractable model. This is because $\nabla \log p_\theta(x) = \nabla p_\theta(x)/p_\theta(x)$ cancels out the normalising constant $Z(\theta)$ by the fraction. The suitability of the Stein discrepancy for use in generalised Bayesian inference has not previously been considered, and this is our focus.

In order that the Stein discrepancy is fully useful in practice, the supremum term over the unit ball of $\mathcal{U}$ in (2.6) should be efficiently computable. Remarkably, the supremum term in (2.6) is attained in closed-form by choosing an appropriate set for $\mathcal{U}$ together with a Stein operator $\mathcal{S}_{\mathbb{Q}}$. For example, the Fisher divergence, developed before the introduction of Stein discrepancy, is an example of Stein discrepancies that are available in closed-form. The Fisher divergence for distributions $\mathbb{Q}, \mathbb{P} \in \mathcal{P}_S(\mathbb{R}^d)$ is defined by

$$\mathrm{FD}^2(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}[\|\nabla \log q(X) - \nabla \log p(X)\|^2]. \tag{2.7}$$

The form of (2.7) can be further translated into a more computationally convenient form that is well-known as the *score matching* objective (Hyvärinen, 2005):

$$\mathrm{FD}^2(\mathbb{Q}\|\mathbb{P}) \stackrel{+C}{=} \mathrm{SM}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X\sim\mathbb{P}}\left[\|\nabla \log q(X)\|^2 + 2\,\mathrm{Tr}\left(\nabla^2 \log q(X)\right)\right] \tag{2.8}$$

where $\stackrel{+C}{=}$ denotes equality that holds up to a $\mathbb{Q}$-independent constant. The form of (2.8) is mostly used in practice because it is computable even when $\mathbb{P}$ is set to an empirical distribution $\mathbb{P}_n$. The Fisher divergence can be derived from (2.6) by selecting the Langevin Stein operator (2.5) for $\mathcal{S}_{\mathbb{Q}}$ and a unit ball of L^2 space of $\mathbb{R}^d$-valued functions for $\mathcal{U}$ (Barp et al., 2019). The Stein discrepancies leveraged in this thesis—KSD and DFD—are each derived in Chapter 4 for continuous intractable models and Chapter 5 for discrete intractable models.

2.5. Reproducing Kernel Hilbert Space

The construction of KSD, a special case of Stein discrepancies whose supremum term in (2.6) are available in closed-form, requires the notion of kernel and RKHS. This section contains background on kernel and the matrix-valued extension. Our main references for the matrix-valued extension are Caponnetto et al. (2008); Carmeli et al. (2006, 2010). We begin with the scalar-valued case and define a scalar-valued kernel:

Definition 1 (Scalar-valued kernel)**.** *A bivariate function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a (scalar-valued)* kernel *if*

(i) k is symmetric*; i.e. $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$,*

(ii) k is positive semi-definite*; i.e. $\sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for any number $n \in \mathbb{N}$, scalars $c_1, \ldots, c_n \in \mathbb{R}$ and points $x_1, \ldots, x_n \in \mathcal{X}$.*

One of the most fundamental facts in kernel methods is that every kernel k can be uniquely associated with some Hilbert space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$. Such a Hilbert space is called the RKHS of kernel k (Paulsen and Raghupathi, 2016, Theorem 2.14).

Definition 2 (Reproducing kernel Hilbert space)**.** *A Hilbert space $\mathcal{H}$ is called a* reproducing kernel Hilbert space *of a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if*

(i) $k(x, \cdot) \in \mathcal{H}$ *for all* $x \in \mathcal{X}$*,*

(ii) $\langle h, k(x, \cdot)\rangle_{\mathcal{H}} = h(x)$ *for all* $x \in \mathcal{X}$ *and* $h \in \mathcal{H}$*.*

Item (ii) is called the reproducing property *of* k *in* $\mathcal{H}$*.*

The unique association between a kernel k and the RKHS $\mathcal{H}$ is extremely convenient because the property of the RKHS $\mathcal{H}$ as a function space is entirely determined and analysed through the kernel k, intuitively speaking. For example, the benefit of this association is crystallised when computing a supremum over a set of functions in the RKHS $\mathcal{H}$. In several settings, such a supremum can be translated into a closed-form quantity dependent only on the kernel k (Gretton et al., 2012).

We next consider the matrix-valued extension of a scalar-valued kernel. The definitions provided above can be generalised in the form of a matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m\times m}$ for an arbitrary dimension $m \in \mathbb{N}$.

Definition 3 (Matrix-valued kernel)**.** *A bivariate function* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m\times m}$*,* $m > 1$*, is called a* (matrix-valued) kernel *if*

(i) K *is* symmetric*; i.e.* $K(x, x') = K(x', x)$ *for all* $x, x' \in \mathcal{X}$*,*

(ii) k *is* positive semi-definite*; i.e.* $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i \cdot k(x_i, x_j)c_j \geq 0$ *for any number* $n \in \mathbb{N}$*, vectors* $c_1, \ldots, c_n \in \mathbb{R}^m$ *and points* $x_1, \ldots, x_n \in \mathcal{X}$*.*

As a direct generalisation of the scalar-valued case, there exists a uniquely associated Hilbert space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}^m$ to every matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m\times m}$. Notice that the Hilbert space $\mathcal{H}$ is now a space of $\mathbb{R}^m$-valued functions on $\mathcal{X}$ and the inner produce $\langle\cdot, \cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a bivariate functional of two $\mathbb{R}^m$-valued functions. To see this Hilbert space $\mathcal{H}$ as RKHS, some additional notation is required: Let F be a $\mathbb{R}^{m\times m}$-valued function and let $F_{i,-}$ denote the vector-valued function $F_{i,-} : \mathcal{X} \to \mathbb{R}^m$ defined by the i-th row of F. Similarly, let G be a $\mathbb{R}^{m\times m}$-valued function and let $G_{-,j}$ denote the vector-valued function $G_{-,j} : \mathcal{X} \to \mathbb{R}^m$ defined by the j-th column of G. With $\mathbb{R}^m$-valued functions $f, g : \mathcal{X} \to \mathbb{R}^m$, define the symbols $\langle F, g\rangle_{\mathcal{H}}$, $\langle f, G\rangle_{\mathcal{H}}$ and $\langle F, G\rangle_{\mathcal{H}}$ by

$$\langle F, g\rangle_{\mathcal{H}} := \begin{bmatrix} \langle F_{1,-}, g\rangle_{\mathcal{H}} \\ \vdots \\ \langle F_{m,-}, g\rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^m,$$

$$\langle f, G\rangle_{\mathcal{H}} := \begin{bmatrix} \langle f, G_{-,1}\rangle_{\mathcal{H}} \\ \vdots \\ \langle f, G_{-,m}\rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^m,$$

$$\langle F, G\rangle_{\mathcal{H}} := \begin{bmatrix} \langle F_{1,-}, G_{-,1}\rangle_{\mathcal{H}} & \cdots & \langle F_{1,-}, G_{-,m}\rangle_{\mathcal{H}} \\ \vdots & & \vdots \\ \langle F_{m,-}, G_{-,1}\rangle_{\mathcal{H}} & \cdots & \langle F_{m,-}, G_{-,m}\rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^{m\times m},$$

where these are to be interpreted as compound symbols only (i.e. we are not attempting to define an inner product on matrix-valued functions). Then, the generalisation of the reproducing property (item (ii) in Definition 2) to a matrix-valued kernel K is

$$h(x) = \langle h, K(x, \cdot)\rangle_{\mathcal{H}} = \begin{bmatrix} \langle h, K_{-,1}(x, \cdot)\rangle_{\mathcal{H}} \\ \vdots \\ \langle h, K_{-,m}(x, \cdot)\rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^m$$

for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$ (Carmeli et al., 2010). The generalisation of the symmetry property (item (i) in Definition 2) is straight-forward; $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$. A Hilbert space $\mathcal{H}$ for which these two properties are satisfied is called a vector-valued RKHS that we say is *reproduced* by the matrix-valued kernel K. Matrix-valued kernels and their associated vector-valued RKHS have recently been exploited in the context of Stein's method (e.g. Barp et al., 2019; Wang et al., 2019). Since we mainly use matrix-valued kernels to construct KSD in this thesis, matrix-valued kernel and vector-valued RKHS are simply called kernel and RKHS when it is clear.

Chapter 3. Generalised Bayesian Inference for Intractable Models

Highly structured data, or data belong to a high-dimensional domain $\mathcal{X}$, are often associated with an intractable model. Moreover, the difficulty of modelling such data means that models will typically be misspecified. Thus, there is a pressing need for Bayesian methods that are both robust and compatible with intractable models. To this end, we introduce *SD-Bayes*, a generalised Bayesian procedure based on a Stein discrepancy as a loss function. There exists numerous choices of Stein discrepancies that can be used in SD-Bayes and each Stein discrepancy equips SD-Bayes with different properties. Building SD-Bayes upon a specific Stein discrepancy is deferred to Chapters 4 and 5, where we will derive concrete Stein discrepancies particularly useful for (i) models in continuous domains and (ii) models in discrete domains, respectively. In this chapter, leaving a choice of the Stein discrepancy arbitrary, we will establish useful theoretical underpinnings of generalised Bayesian procedures. The aim of this chapter is to provide certain types of assurance that generalised Bayesian procedures operate well, in advance of giving a specific shape to SD-Bayes in Chapters 4 and 5. Typically, generalised Bayesian procedures benefit from "calibration" of the posteriors to adjust their credible regions. Independently of the use of a Stein discrepancy in SD-Bayes, we will also consider a novel calibration algorithm based on a Stein discrepancy, illuminating the striking computational advantages in the context of the calibration.

This chapter is structured as follows: Section 3.1 presents SD-Bayes with a form of a Stein discrepancy kept abstract. The SD-Bayes methodology will become concrete once a specific Stein discrepancy is selected, as considered later in Chapters 4 and 5. Section 3.2 establishes asymptotic properties—*posterior consistency* and the *Bernstein–von Mises theorem*—of generalised posteriors that ensure their appealing regularities as the number of data increases. Section 3.3 establishes a condition of generalised posteriors to be robust to model misspecification caused by an outlier, formulating a qualitative criterion called *global bias-robustness* of generalised posteriors. Finally, a new approach to calibrating generalised posteriors using a Stein discrepancy is discussed in Section 3.4. Note that all the results in Sections 3.2 to 3.4 can apply to any generalised posteriors that are not limited to SD-Bayes, and may hence be of independent interest. It is also worth highlighting that the asymptotic results in Section 3.2 hold regardless of whether a model is well-specified or misspecified. The results thus cover practical cases under model misspecification where the use of generalised Bayesian inference is usually motivated.

3.1. SD-Bayes Framework

Suppose we are given a prior p.d.f. $\pi \in \mathcal{P}(\Theta)$ and a statistical model $\{\mathbb{P}_\theta \mid \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$. Let $\{x_i\}_{i=1}^n$ be independent observations generated from a population distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and let $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical measure associated to this dataset. In this context, the framework of SD-Bayes can now be defined as follows:

Definition 4 (SD-Bayes)**.** *For a model $\mathbb{P}_\theta$, select a Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ and a Stein set $\mathcal{U}$. Denote the associated Stein discrepancy by $\mathrm{SD}(\mathbb{P}_\theta\|\mathbb{P}_n)$ for a given data distribution $\mathbb{P}_n$. The SD-Bayes posterior is defined by*

$$\pi_n^D(\theta) \propto \exp\left(-\beta n \,\mathrm{SD}^\gamma(\mathbb{P}_\theta\|\mathbb{P}_n)\right)\pi(\theta) \tag{3.1}$$

where $\beta, \gamma \in (0, \infty)$.

Comparing (3.1) to (2.2) confirms that SD-Bayes is a generalised Bayesian methodology given the loss function $D_n(\theta) = n\ \mathrm{SD}^\gamma(\mathbb{P}_\theta\|\mathbb{P}_n)$, where n is the default scaling of the loss $\mathrm{SD}^\gamma(\mathbb{P}_\theta\|\mathbb{P}_n)$ to induce concentration of the posterior as n increases. There is an arbitrariness to using the γ-powered discrepancy, but $\gamma = 2$, which is our default choice, turns out to be appropriate for all Stein discrepancies considered in this thesis, ensuring that concentration of $\mathrm{SD}^2(\mathbb{P}_\theta\|\mathbb{P}_n)$ around its population value $\mathrm{SD}^2(\mathbb{P}_\theta\|\mathbb{P})$ occurs at a rate $\mathcal{O}(n^{-1/2})$ analogous to the standard Bayesian inference case. It also permits tractable computation of (3.1) and enables concrete analysis of the SD-Bayes posterior, whose details are described in Chapters 4 and 5. We hence focus on our default choice $\gamma = 2$ entirely in this thesis. A discussion of how the weighting constant β should be selected is provided in Section 3.4.

There exists appealing Stein discrepancies that attain a closed-form solution of the supremum term in (2.6). Such Stein discrepancies, that are analytically computable without the inconvenient need of approximating the supremum term, are useful in practice and capitalised in this thesis. Table 3.1 summarises three major Stein discrepancies available in closed-form. It is worth highlighting that our construction of KSD generalises the original formulations (Chwialkowski et al., 2016; Liu et al., 2016; Wang et al., 2019), accepting any arbitrary Stein operator $\mathcal{S}_{\mathbb{P}}$ and any general domain $\mathcal{X}$. Furthermore, our construction of DFD permits a wider class of discrete domains than the existing studies (Yang et al., 2018). Their details are deferred to Chapters 4 and 5 as aforementioned. A few other computable Stein discrepancies are not listed above; for example, diffusion KSD (Barp et al., 2019) is a special case of KSD under our general construction, and diffusion score matching (Barp et al., 2019) is a straightforward variant of FD replacing the Langevin Stein operator with its extension called *diffusion* Stein operator.

Next, we turn our attention to establishing several theoretical underpinnings of generalised Bayesian procedures. All the arguments in the remainder of this chapter are not limited to SD-Bayes. In what follows, $D_n(\theta)$ denotes an arbitrary loss function for

name	domain $\mathcal{X}$	Stein operator $\mathcal{S}_{\mathbb{P}}$	Stein set $\mathcal{U}$	closed form
KSD	arbitrary	arbitrary	unit ball of RKHS of kernel K	(4.1)
DFD	discrete	(5.5)	unit ball of $L^2(\mathbb{Q};\mathbb{R}^d)$	(5.2)
FD	continuous	(2.5)	unit ball of $L^2(\mathbb{Q};\mathbb{R}^d)$	(2.7)

Table 3.1 A list of major Stein discrepancies computable in closed form.

a model $\mathbb{P}_\theta$ dependent of a dataset $\{x_i\}_{i=1}^n$ and $D(\theta)$ denotes its population counterpart dependent of a data-generating distribution $\mathbb{P}$ of the dataset. These results will be applied to SD-Bayes in Chapters 4 and 5, given specific Stein discrepancies. Note that all the main results presented below are general enough to permit a loss $D_n(\theta)$ to be arbitrary, a model $\mathbb{P}_\theta$ to be misspecified, and data $\{x_i\}_{i=1}^n$ to be non-i.i.d. However, non-i.i.d. data setting causes difficulty in verifying the derived conditions in each application, and therefore non-i.i.d. data are not focused in this thesis. It is demonstrated in the subsequent chapters that i.i.d. case includes a number of interesting applications.

3.2. Posterior Consistency and Bernstein–von Mises Theorem

Posterior consistency and the Bernstein–von Mises theorem are long-studied theoretical underpinnings of Bayesian methodologies to justify them from the frequentist perspective. A posterior distribution is said to be *consistent* if it concentrates around an "optimal" parameter θ_*, that is, a minimiser of a population loss $D(\theta)$. For example, $D(\theta) = -\frac{1}{n}\mathbb{E}_{X\sim\mathbb{P}}[\log p_\theta(X)]$ in standard Bayesian inference. This guarantees that a posterior identifies the optimal parameter in the limit $n \to \infty$, placing all the probability mass at the same limiting point θ_* as the frequentist estimator $\theta_n = \arg\min_{\theta\in\Theta} D_n(\theta)$. On the other hand, the Bernstein–von Mises theorem tells us the asymptotic rate at which this convergence happens, suggesting that a re-scaled posterior by a factor of $\sqrt{n}$ is asymptotically normal around the optimal parameter θ_*. This implies that inconvenient properties of a posterior, such as multi-modality, can diminish at an asymptotic rate $\sqrt{n}$ given a sufficient amount of data. These two fundamental results ensure that a posterior is well-regulated in the limit and that Bayesian methodologies function aptly even under the circumstance where data continuously and almost infinitely increase.

Posterior consistency and the Bernstein–von Mises theorem, which were originally formulated for standard Bayesian inference, can be extended to generalised Bayesian inference of arbitrary loss $D_n(\theta)$, conferring the above underpinning from the frequentist perspective. The classical result developed for standard Bayesian inference often requires a model to be well-specified. In contract, our extension to generalised Bayesian procedures holds regardless of whether a model is well-specified or misspecified. Thus, the results in this section are practical because the use of generalised Bayesian procedures is (not only but) often considered in applications of misspecified models.

3.2.1. Posterior Consistency

First, we establish posterior consistency of generalised posteriors. The term "posterior consistency" generally refers to concentration of a posterior to a point mass at the optimal parameter θ_* in the limit $n \to \infty$. Nonetheless, it is also possible to derive a non-asymptotic rate of the concentration with respect to the data number n. Such rates, often referred to as *posterior concentration rate*, is a more precise version of posterior consistency (Rousseau, 2016). For example, posterior concentration rate in the context of generalised posteriors was considered by Cherief-Abdellatif and Alquier (2020) for a specific case where the loss is set to the maximum mean discrepancy (Gretton et al., 2012). Inspired by their analysis, we establish an analogous result for generalised posteriors with general losses.

First of all, our consistency result requires a *prior mass condition* similar to that of Cherief-Abdellatif and Alquier (2020). It specifies the amount of prior mass in a neighbourhood around the optimal parameter θ_* that is required.

Assumption 1 (Prior Mass Condition)**.** *For any $\theta_* \in \arg\inf_{\theta\in\Theta} D(\theta)$, a prior is assumed to*

1. *admit a p.d.f. π that is continuous at θ_*, with $\pi(\theta_*) > 0$;*
2. *satisfy $\int_{B_n(\alpha_1)} \pi(\theta)\mathrm{d}\theta \geq e^{-\alpha_2\sqrt{n}}$ for some constants $\alpha_1, \alpha_2 > 0$,*

where we define $B_n(\alpha_1) := \{\theta \in \Theta : |D(\theta) - D(\theta_*)| \leq \alpha_1/\sqrt{n}\}$.

This is not a strong condition because the required prior mass over $B_n(\alpha_1)$ decays exponentially as n grow. However, it often needs to be directly assumed rather than verified due to the condition depending on a population loss $D(\theta)$ that is inaccessible in practice. Our consistency result also requires a convergence rate of the empirical loss $D_n(\theta)$ to the population loss $D(\theta)$ in expectation with respect to realisation of the datasets:

Assumption 2 (Convergence Rate)**.** *At each $\theta \in \Theta$, there exists $0 < \sigma(\theta) < \infty$ s.t.*

$$\mathbb{E}_{X_1,\ldots,X_n}\left[|D_n(\theta) - D(\theta)|\right] \leq \frac{\sigma(\theta)}{\sqrt{n}}. \tag{3.2}$$

If the loss $D_n(\theta)$ is additive—i.e., $D_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(\theta, x_i)$ for some function l—and data are i.i.d., Assumption 2 holds immediately given a finite variance of l. The following proposition verifies Assumption 2 for any additive loss.

Proposition 1. *Suppose that data are i.i.d. generated from a distribution $\mathbb{P}$. Suppose that there exists $l : \Theta \times \mathcal{X} \to \mathbb{R}$ s.t. $D_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(\theta, x_i)$ and $\mathbb{V}_{X\sim\mathbb{P}}[l(\theta, X)] < \infty$ at each $\theta \in \Theta$. Then Assumption 2 holds for $\sigma(\theta) = \sqrt{\mathbb{V}_{X\sim\mathbb{P}}[l(\theta, X)]}$.*

The proof is provided in Section 3.5.1. Interestingly, the additivity does not hold for some choice of the Stein discrepancy, such as KSD whose data-dependent form corresponds

to a double summation rather than a single summation. In spite of that, it is shown in Chapter 4 that KSD satisfies Assumption 2.

For a posterior distribution, denoted by Π_n, its posterior concentration rate is often expressed in a form, given some metric d on Θ and some sequence $\{\epsilon_n\}_{n=1}^{\infty}$,

$$\mathbb{P}\left(\Pi_n \left(\{\theta \in \Theta \mid d(\theta, \theta_*) > \epsilon_n\}\right) > \epsilon \right) \to 0 \quad \text{as } n \to \infty \quad \text{for all } \epsilon > 0, \tag{3.3}$$

where the probability $\mathbb{P}$ is taken with respect to realisations of data $\{X_i\}_{i=1}^n$ used in the posterior distribution Π_n. More simply, this is equivalent to say: $\Pi_n \left(\{\theta \in \Theta \mid d(\theta, \theta_*) > \epsilon_n\}\right) \to 0$ in $\mathbb{P}$-probability. The sequence $\{\epsilon_n\}_{n=1}^{\infty}$ corresponds to the non-asymptotic concentration rate of the posterior Π_n. It follows from the Markov inequality that

$$\Pi_n \left(\{\theta \in \Theta \mid d(\theta, \theta_*) > \epsilon_n\}\right) \leq \frac{\mathbb{E}_{\theta \sim \Pi_n}\left[d(\theta, \theta_*)\right]}{\epsilon_n}. \tag{3.4}$$

Then, this inequality implies that $\Pi_n \left(\{\theta \in \Theta \mid d(\theta, \theta_*) > \epsilon_n\}\right) \to 0$ in $\mathbb{P}$-probability if the expectation $\mathbb{E}_{\theta \sim \Pi_n}\left[d(\theta, \theta_*)\right]$ decays faster than ϵ_n in $\mathbb{P}$-probability. In other words, if we find any concentration rate ϵ_n' of the expectation $\mathbb{E}_{\theta \sim \Pi_n}\left[d(\theta, \theta_*)\right]$ that holds in $\mathbb{P}$-probability, the original posterior concentration rate in (3.3) immediately holds by choosing any ϵ_n s.t. $\epsilon_n'/\epsilon_n \to 0$. Thus, it suffices to find a concentration rate of the expectation $\mathbb{E}_{\theta \sim \Pi_n}\left[d(\theta, \theta_*)\right]$ in $\mathbb{P}$-probability to show a concentration rate of the posterior Π_n. We use the former as a simplified representation of posterior consistency in this thesis. This representation of posterior consistency is now established under $d(\theta, \theta_*) = |D(\theta) - D(\theta_*)|$.

Theorem 1 (Posterior Consistency)**.** *Suppose that Assumptions 1 and 2 hold. Suppose that* $\sup_{\theta \in \Theta} \sigma(\theta) < \infty$ *for* $\sigma(\theta)$ *in 2. Then, for all* $\delta \in (0, 1]$*,*

$$\mathbb{P}\left(\int_{\Theta} |D(\theta) - D(\theta_*)| \, \pi_n^D(\theta) \mathrm{d}\theta < \frac{\alpha_1 + \alpha_2 + 2 \sup_{\theta \in \Theta} \sigma(\theta)}{\delta \sqrt{n}} \right) \geq 1 - \delta$$

where the probability is with respect to realisations of data $\{X_i\}_{i=1}^n$*.*

The proof is contained in Section 3.5.2. This result provides an additional intuitive implication of posterior consistency. By the Jensen's inequality,

$$\left| \int_{\Theta} D(\theta) \pi_n^D(\theta) - D(\theta_*) \right| \leq \int_{\Theta} |D(\theta) - D(\theta_*)| \, \pi_n^D(\theta) \tag{3.5}$$

The right-hand side converges to 0 in $\mathbb{P}$-probability by Theorem 1, so does the left-hand side because of the inequality above. This means that the expected loss $\mathbb{E}_{\theta \sim \pi_n^D}[D(\theta)]$ with respect to the posterior π_n^D converges to the minimiser $D(\theta_*)$ in $\mathbb{P}$-probability as n grows.

3.2.2. Bernstein–von Mises Theorem

Next, we derive a Bernstein–von Mises result, that is nothing but asymptotic normality of a posterior. The Bernstein-von Mises Theorem was originally formulated for standard

posteriors of well-specified models. Over the past decade, it was extended to cases of misspecified models Kleijn and van der Vaart (2012a) and further to cases of generalised posteriors Ghosh and Basu (2016); Hooker and Vidyashankar (2014); Miller (2021). The pioneering work of Hooker and Vidyashankar (2014) and Ghosh and Basu (2016) established Bernstein–von Mises results for generalised posteriors built upon α- and β-divergences. Unfortunately, the form of Stein discrepancy is rather different and alternative theoretical tools are required to tackle it. Recently, Miller (2021) introduced the most general approach to deriving Bernstein–von Mises results for essentially any generalised posteriors, demonstrating how their assumptions can be verified for several additive loss functions D_n.

The conditions of Miller (2021) can be refined into more applicable forms. We introduce the convenient conditions to draw on the argument of Miller (2021) below.

Assumption 3 (BvM Condition)**.** *Let $\Theta \subseteq \mathbb{R}^p$ be Borel. Let $H_n(\theta) := \nabla^2_\theta D_n(\theta)$. Suppose that there exists some bounded convex open set $U \subseteq \Theta$ s.t. the following hold:*

C1 D_n a.s. converges pointwise to D;

C2 D_n is r times continuously differentiable in U and $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla^r_\theta D_n(\theta)\| < \infty$ a.s. for $r = 1, 2, 3$;

C3 for all n sufficiently large, there exists $\arg\min_{\theta\in\Theta} D_n(\theta)$ in U, i.e., any $\theta_n \in \arg\min_{\theta\in\Theta} D_n(\theta)$ satisfies $\theta_n \in U$ a.s., and there exists $\theta_* \in U$ that uniquely attains $D(\theta_*) = \inf_{\theta\in\Theta} D(\theta)$.

C4 $H_n(\theta_*) \xrightarrow{a.s.} H_*$ for some nonsingular matrix H_*;

C5 π is continuous and positive at θ_*.

The existence of U implies that, for large enough n, we can essentially restrict our theoretical analysis to the bounded subset $U \subseteq \Theta$. This part of the assumption is not restrictive: it can be enforced by re-parameterising the model p_θ so that its new parameter space is bounded and convex.[1] The existence of θ_n and θ_* is more difficult to assess in practice, since the true data generating distribution is unknown. That being said, assuming their existence is common in the asymptotic analysis of Bayesian procedures (see e.g. van der Vaart, 1998, Section 10). In most cases, the condition C3 and the nonsingularity of H_* in the condition C4 have to be directly assumed due to their difficulty to guarantee a priori without knowing the true data generating distribution. The rest of the conditions can be verified (or simplified) for each loss D_n and prior π.

Before showing that Assumption 3 are sufficient for (Miller, 2021, Theorem 4), we introduce two lemmas that are useful for the main result. The first lemma is on a.s. uniform convergence of a loss D_n.

[1]For example, we can re-parameterise any unbounded parameter κ through the logistic function and define the invertible transformation $\theta = (1 + e^{-\kappa})^{-1} \in [0, 1]$.

Lemma 1 (a.s. Uniform Convergence)**.** *Suppose that the preconditions C1 and C2 in Assumption 3 holds for $r = 1$. Then D_n a.s. converges uniformly to D on U in Assumption 3.*

The proof is contained in Section 3.5.3. The second lemma is on strong consistency of an estimator θ_n that minimises a loss D_n.

Lemma 2 (Strong Consistency)**.** *Suppose that the preconditions C1, C2, and C3 in Assumption 3 holds for $r = 1$. Then, for a point $\theta_* = \arg\min_{\theta\in\Theta} D(\theta)$ and any sequence $\{\theta_n\}_{n=1}^{\infty}$ s.t. $\theta_n \in \arg\min_{\theta\in\Theta} D_n(\theta)$ for all n sufficiently large, it holds that $\theta_n \xrightarrow{a.s.} \theta_*$.*

The proof is contained Section 3.5.4.

We are in a position to present the following Bernstein–von Mises result that holds under the refined, convenient conditions provided in Assumption 3.

Theorem 2 (BvM Theorem)**.** *Suppose Assumption 3 holds. Let $(\theta_n)_{n=1}^{\infty} \subset \Theta$ be a sequence s.t. θ_n minimises D_n for all n sufficiently large. Denote by $\widetilde{\pi}_n^D$ a density on $\mathbb{R}^d$ of the random variable $\sqrt{n}(\theta - \theta_n)$, where $\theta \sim \pi_n^D$. Then*

$$\int_{\mathbb{R}^d} \left| \widetilde{\pi}_n^D(\theta) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}} \exp\left(-\frac{1}{2}\theta \cdot H_*\theta\right) \right| \mathrm{d}\theta \xrightarrow{\text{a.s.}} 0$$

where the a.s. convergence is with respect to realisations of data $\{X_i\}_{i=1}^n$.

The proof is provided in Section 3.5.5. The Bernstein–von Mises result ensures an appealing regularity of generalised posteriors to be asymptotically normal. For example, the regularity of generalised posteriors can underpin the use of Laplace's approximation to them when the number of data is sufficiently large. Beyond the scope of this thesis, intriguingly, there also exists several advanced settings where posteriors converge to non-normal asymptotic distributions. See Bochkina and Green (2014) for the case of standard Bayesian inference whose optimal parameter θ_* lies in the boundary of the closed parameter space Θ and Frazier et al. (2020) for the case of approximate Bayesian computation under model misspecification.

3.3. Bayesian Robustness to Model Misspecification by Outlier

It is crucial to formulate rigorously in what sense generalised Bayesian procedures are robust to model misspecification. In this thesis, we limit our focus to a simple yet common setting where model misspecification are caused by an outlier contaminating data. We define a property termed *global bias-robustness* that indicates a strong insensitivity of generalised posteriors to outliers mixed with data. We then derive an explicit sufficient condition for generalised posteriors with arbitrary loss D_n to satisfy global bias-robustness. Our analysis reveals that generalised posteriors under the derived condition can limit a negative influence of an extreme outlier on their inference outcomes, in contrast to a standard posterior that often fails to do so.

The notion and formulation of robustness in the context of statistics were originally formalised for frequentist estimators (Huber and Ronchetti, 2009). One of the classical concerns was robustness of estimators in the presence of outliers. A criterion called *bias-robustness* (also referred to as *gross error sensitivity*) was developed to analyse a qualitative insensitivity of frequentist estimators to an outlier mixed in data. Consider a mixture distribution $\mathbb{P}_{n,\epsilon,y} = (1-\epsilon)\mathbb{P}_n + \epsilon\delta_y$ for an empirical distribution $\mathbb{P}_n$ of any data $\{x_i\}_{i=1}^n$ and a point $y \in \mathcal{X}$, typically called the *ε-contamination model* (Huber and Ronchetti, 2009). The point y is considered to be contaminating the dataset $\{x_i\}_{i=1}^n$ and the mixture proportion ϵ controls the level of contamination. Let $T : \mathcal{P}(\mathcal{X}) \to \Theta$ be a statistical estimator viewed as a map from a given probability distribution to a quantity of interest. For example, a sample mean estimator is a map from an empirical distribution of data to its mean value. Define the *influence function* of the estimator T by

$$\mathrm{IF}(y, \mathbb{P}_n) := \frac{\mathrm{d}}{\mathrm{d}\epsilon} T(\mathbb{P}_{n,\epsilon,y})|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{T(\mathbb{P}_{n,\epsilon,y}) - T(\mathbb{P}_n)}{\epsilon}. \tag{3.6}$$

The estimator T is then said *bias-robust* if $\sup_{y\in\mathcal{X}} \| \mathrm{IF}(y, \mathbb{P}_n)\| < \infty$ (Barp et al., 2019). Intuitively speaking, if an extreme outlier is mixed with data and one's estimator is not bias-robust, values of the estimator can change drastically even by an infinitesimally small level of contamination. Our aim is to extend bias-robustness of frequentist estimators to generalised posteriors, proposing a Bayesian analogue of bias-robustness.

Several robustness properties were considered for specific choices of generalised posteriors in Ghosh and Basu (2016); Hooker and Vidyashankar (2014); Nakagawa and Hashimoto (2020). In particular, Ghosh and Basu (2016); Hooker and Vidyashankar (2014) defined an analogue of the influence function for their α, β-divergence posteriors. We define a similar quantity to Ghosh and Basu (2016) for any generalised posterior, which we call the *posterior influence function* to distinguish it from the standard influence function. In what follows, we write a loss $D_n(\theta) = D(\theta; \mathbb{P}_n)$ and its associated generalised posterior $\pi_n^D(\theta) = \pi_n^D(\theta; \mathbb{P}_n)$ to make explicit the dependence on data $\mathbb{P}_n$. Consider a generalised posterior based on a loss $D(\theta; \mathbb{P}_{n,\epsilon,y})$ dependent on contaminated data $\mathbb{P}_{n,\epsilon,y}$, denoted by $\pi_n^D(\theta; \mathbb{P}_{n,\epsilon,y})$. The posterior influence function of π_n^D is then defined by

$$\mathrm{PIF}(y, \theta, \mathbb{P}_n) := \frac{\mathrm{d}}{\mathrm{d}\epsilon} \pi_n^D(\theta; \mathbb{P}_{n,\epsilon,y})|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{\pi_n^D(\theta; \mathbb{P}_{n,\epsilon,y}) - \pi_n^D(\theta; \mathbb{P}_n)}{\epsilon}. \tag{3.7}$$

The generalised posterior π_n^D is said *globally bias-robust* if $\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\,\mathrm{PIF}(y, \theta, \mathbb{P}_n)| < \infty$. This means that change in values of the generalised posterior by the extreme contaminant y will be limited uniformly over Θ. The following lemma establishes a sufficient condition for generalised posteriors to satisfy global bias-robustness.

Lemma 3. *Let π_n^D be a generalised posterior with a loss $D(\theta; \mathbb{P}_n)$ and a prior π. Suppose $D(\theta; \mathbb{P}_n)$ is lower-bounded and $\pi(\theta)$ is upper-bounded over $\theta \in \Theta$, for any $\mathbb{P}_n$. Denote*

$D_0(y, \theta, \mathbb{P}_n) := (\mathrm{d}/\mathrm{d}\epsilon) D(\theta; \mathbb{P}_{n,\epsilon,y})|_{\epsilon=0}$. *Then* π_n^D *is globally bias-robust if, for any* $\mathbb{P}_n$,

$$1. \ \sup_{\theta \in \Theta} \sup_{y \in \mathcal{X}} |D_0(y, \theta, \mathbb{P}_n)| \, \pi(\theta) < \infty \quad \textit{and} \quad 2. \ \int_\Theta \sup_{y \in \mathcal{X}} |D_0(y, \theta, \mathbb{P}_n)| \, \pi(\theta) \mathrm{d}\theta < \infty.$$

The proof is provided in Section 3.5.6.

Note that standard Bayesian inference can easily violate the conditions of Lemma 3 in general. Indeed, when $D(\theta; \mathbb{P}_n)$ is the negative log likelihood, $D_0(y, \theta, \mathbb{P}_n)$ is derived as $D_0(y, \theta, \mathbb{P}_n) = \log p_\theta(y) - \sum_{i=1}^n \log p_\theta(x_i)$, which contains the term $\log p_\theta(y)$ that can be easily unbounded in case of $\mathcal{X} = \mathbb{R}^d$ for example. The term $\log p_\theta(y)$ can be unbounded even when a model is light-tailed, e.g., consider a log-likelihood of a normal location model. In contrast, appropriate Stein discrepancies, such as KSD, provides a degree of freedom which can be leveraged to ensure the conditions of Lemma 3. The SD-Bayes posterior resulting from such Stein discrepancies will be conferred strong robustness to model misspecification by an outlier. This will be demonstrated in Chapter 4.

3.4. Posterior Calibration via Bootstrapping and Divergence Minimisation

Selecting an appropriate value of β is critical to ensure that generalised posteriors are calibrated. To date, two existing approaches stand out. One approach was proposed in the recent review paper of Syring and Martin (2019). It consists of a new stochastic sequential update algorithm for choosing β, such that the 95% highest posterior density region coincides with a 95% confidence interval. Unfortunately, this approach leads to a large computational cost and is therefore often impractical. Another approach is due to Lyddon et al. (2019) and consists in setting β such that the trace of an asymptotic covariance matrix of a generalise posterior coincides with that of the frequentist counterpart. See also Frazier et al. (2023) who proposed a novel class of generalised posteriors whose asymptotic covariance exactly coincides with that of the frequentist counterpart. For selection of β, the algorithm by Lyddon et al. (2019) appears to be one of the most straightforward approaches if estimation of the asymptotic covariance at finite n is accurate. We will adopt the approach by Lyddon et al. (2019) when a reliable estimate of the asymptotic covariance is available. However, it will be demonstrated in Chapter 5 that estimation of the asymptotic covariance can be numerically—sometimes excessively—unstable when $\mathcal{X}$ or Θ is high dimensional. In such cases, selection of β benefits from a more efficient and stable approach that aims at the same objective to match the scale of the credible region with that of the confidence interval in an approximate sense.

We propose a novel calibration algorithm of β that exploits the advantages of a Stein discrepancy and bootstrapping. Note that the use of a Stein discrepancy in selecting β here is independent of the use of a Stein discrepancy in SD-Bayes. The calibration algorithm is applicable to generalised posteriors of any loss $D_n(\theta)$. In short, our approach consists of two steps: (i) computing minimisers of B "bootstrapped" versions of a loss $D_n(\theta)$ and (ii) estimating an optimal value of β using the closed-form expression provided

in subsequent Theorem 3. In contrast to the stochastic iterative approach of Syring and Martin (2019), step (ii) is non-iterative and exact. Additionally, computation of each minimiser in step (i) is embarrassingly parallel. In contrast to the approach of Lyddon et al. (2019), our approach relies on the bootstrap sampling distribution of the frequentist estimator, circumventing the use of the asymptotic covariance whose estimation can be highly unstable when $\mathcal{X}$ or Θ is high-dimensional. It is also worth highlighting that our approach takes a prior π into account, in another contrast to the approach of Lyddon et al. (2019) that depends only on the asymptotic covariance.

To describe our methodology in detail, we denote a generalised posterior π_n^D by $\pi_{n,\beta}^D$ making the dependence on β explicit. In step (i), B bootstrap datasets $\{x_i^{(b)}\}_{i=1}^n$ for $b = 1, \ldots, B$ are generated by sampling each $x_i^{(b)}$ uniformly with replacement from the original dataset $\{x_i\}_{i=1}^n$. We then compute a minimiser $\theta_n^{(b)} = \arg\min_{\theta\in\Theta} D_n^{(b)}(\theta)$ for the loss $D_n^{(b)}$ associated with each b-th bootstrap dataset. This leads to an empirical measure $\delta_\theta^B = \frac{1}{B}\sum_{b=1}^B \delta(\theta_n^{(b)})$ which approximates the sampling distribution of the frequentist estimator $\theta_n = \arg\min_{\theta\in\Theta} D_n(\theta)$ for the original loss D_n. In step (ii), we choose β that minimises a statistical divergence between $\pi_{n,\beta}^D$ and δ_θ^B. However, this is not straightforward, since the majority of statistical divergences (e.g. Kullback–Liebler divergence) require the normalising constant of $\pi_{n,\beta}^D$ for every β. Interestingly, this is the same computational challenge posed by intractable model. Our proposal, called the Stein posterior calibration, is therefore to employ a Stein discrepancy that circumvents evaluation of the normalising constant.

Definition 5 (Stein Posterior Calibration)**.** *Given a Stein discrepancy* SD *for probability distributions on* Θ*, the Stein posterior calibration selects* $\beta \in (0, \infty)$ *by a solution* β_* *to the following optimisation*

$$\beta_* \in \arg\min_{\beta>0} \mathrm{SD}(\pi_{n,\beta}^D \| \delta_\theta^B). \tag{3.8}$$

Strikingly, some specific choice of Stein discrepancies leads to a closed-form solution of β_* in (3.8). Selecting the score matching objective (2.8) as the Stein discrepancy in Definition 5 leads to the following form of the optimisation objective (3.8):

$$\beta_* \in \arg\min_{\beta>0} \frac{1}{n}\sum_{b=1}^B \left\|\nabla \log \pi_{n,\beta}^D\left(\theta_n^{(b)}\right)\right\|^2 + 2\,\mathrm{Tr}\left(\nabla^2 \log \pi_{n,\beta}^D\left(\theta_n^{(b)}\right)\right). \tag{3.9}$$

This allows us to establish a closed-form solution of β_* below.

Theorem 3. *Let loss* $D_n : \Theta \to \mathbb{R}$ *be twice differentiable with respect to* θ*. Suppose that* $\sum_{b=1}^B \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}(\nabla_\theta^2 D_n(\theta_n^{(b)})) > 0$ *and that there exists at least one* $\theta_n^{(b)}$ *s.t.* $\nabla_\theta D_n(\theta_n^{(b)}) \neq 0$*. Then,* β_* *in* (3.9) *admits a unique analytical solution*

$$\beta_* = \frac{\sum_{b=1}^B \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}(\nabla_\theta^2 D_n(\theta_n^{(b)}))}{\sum_{b=1}^B \|\nabla_\theta D_n(\theta_n^{(b)})\|^2}. \tag{3.10}$$

The proof is provided in Section 3.5.7.

Note that (3.10) is straightforward to compute whenever the loss D_n is amenable to automatic differentiation. This approach offers a substantial computational advantage by the closed-form solution. However, rigorous theoretical analysis of the approach, such as the asymptotic property in the limit and the advantage in high dimension, is left for future research. We hence use the approach of Lyddon et al. (2019) for theoretical explicitness if the asymptotic covariance can be reliably estimated. It will turn out that the asymptotic covariance can be stably estimated to a feasible degree in applications of Chapter 4, where β will be then selected by the approach of Lyddon et al. (2019). Our approach will be leveraged in Chapter 5, where estimation of the asymptotic covariance is exceedingly unstable in some high-dimensional applications. It will be demonstrated in Chapter 5 that our approach produces a sensible value of β with strong stability even for applications in which the approach of Lyddon et al. (2019) severely fails.

3.5. Proofs of Chapter 3

This section contains all the deferred proofs of theoretical results in Chapter 3.

3.5.1. Proof of Proposition 1

Proof. By the Jensen's inequality, the following bound holds

$$\sqrt{\left(\mathbb{E}_{X_1,\ldots,X_n}\left[|D_n(\theta) - D(\theta)|\right]\right)^2} \leq \sqrt{\mathbb{E}_{X_1,\ldots,X_n}\left[\left(D_n(\theta) - D(\theta)\right)^2\right]}.$$

Recall $D_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(\theta, x_i)$ to see that

$$\begin{aligned}\left(D_n(\theta) - D(\theta)\right)^2 &= \left(\frac{1}{n}\sum_{i=1}^n l(\theta, x_i) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(l(\theta, x_i) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right)\left(l(\theta, x_j) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right).\end{aligned}$$

Let $(*_{i,j}) := \left(l(\theta, x_i) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right)\left(l(\theta, x_j) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right)$ for better presentation. We take an expectation with respect to $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. The expectation of the term $(*_{i,j})$ is zero if $i \neq j$ because $X_1, \ldots, X_n$ are i.i.d. and $\mathbb{E}_{X_i}\left[l(\theta, X_i) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right] = 0$. Therefore, only the expectation of the term $(*_{i,i})$ for $i = 1, \ldots, n$ remains non-zero:

$$\begin{aligned}\mathbb{E}_{X_1,\ldots,X_n}\left[\left(D_n(\theta) - D(\theta)\right)^2\right] &= \mathbb{E}_{X_1,\ldots,X_n}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (*_{i,j})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{X_i}\left[\left(l(\theta, X_i) - \mathbb{E}_{X\sim\mathbb{P}}[l(\theta, X)]\right)^2\right] = \frac{1}{n}\mathbb{V}_X\left[l(\theta, X)\right].\end{aligned}$$

Plugging this equality in the first inequality at the top completes the proof. □

3.5.2. Proof of Theorem 1

Proof. The following preliminary inequality is required, which takes inspiration from Alquier et al. (2016); Cherief-Abdellatif and Alquier (2020): for all $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq D(\theta_*) + \left(\alpha_1 + \alpha_2 + \frac{2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta}\right)\frac{1}{\sqrt{n}}$$

where the probability is with respect to realisations of random data $\{X_i\}_{i=1}^n$.

First of all, we show the inequality above. By the Markov inequality and the assumption that $\mathbb{E}_{X_1,\ldots,X_n}[|D_n(\theta) - D(\theta)|] \leq \frac{\sigma(\theta)}{\sqrt{n}}$, we have a concentration inequality

$$\mathbb{P}\left(|D_n(\theta) - D(\theta)| \geq \delta\right) \leq \frac{\sigma(\theta)}{\delta\sqrt{n}} \leq \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}} \tag{3.11}$$

for each $\theta \in \Theta$, where the probability is taken with respect to random data $\{X_i\}_{i=1}^n$. Taking the complement and re-scaling δ, (3.11) is equivalent to

$$\mathbb{P}\left(|D_n(\theta) - D(\theta)| \leq \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}\right) \geq 1 - \delta. \tag{3.12}$$

Notice that by virtue of the absolute value, the following inequalities hold simultaneously with probability at least $1 - \delta$:

$$D(\theta) \leq D_n(\theta) + \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}. \tag{3.13}$$

$$D_n(\theta) \leq D(\theta) + \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}. \tag{3.14}$$

Taking an expectation with respect to the generalised posterior on both side of (3.13) yields, with probability at least $1 - \delta$,

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq \int_\Theta D_n(\theta)\pi_n^D(\theta)\mathrm{d}\theta + \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}$$

In order to apply the identity (2.3), we add the term $(1/n)\,\mathrm{KL}(\pi_n^D\|\pi) \geq 0$ in the right-hand side and see that, with probability at least $1 - \delta$,

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq \left\{\int_\Theta D_n(\theta)\pi_n^D(\theta)\mathrm{d}\theta + \frac{1}{n}\,\mathrm{KL}(\pi_n^D\|\pi)\right\} + \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}.$$

Then from the identity (2.3), the right-hand side coincides with the solution to the following variational problem over $\mathcal{P}(\Theta)$:

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\int_\Theta D_n(\theta)\rho(\theta)\mathrm{d}\theta + \frac{1}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}. \tag{3.15}$$

Plugging (3.14) in (3.15), we have with probability at least $1-\delta$,

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\int_\Theta D(\theta)\rho(\theta)\mathrm{d}\theta + \frac{1}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}. \tag{3.16}$$

Plugging the trivial bound $D(\theta) \leq D(\theta_*) + |D(\theta) - D(\theta_*)|$ into (3.16), we have

$$(3.16) \leq D(\theta_*) + \inf_{\rho\in\mathcal{P}(\Theta)}\left\{\int_\Theta |D(\theta)-D(\theta_*)|\,\rho(\theta)\mathrm{d}\theta + \frac{1}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}.$$

Notice that the infimum term can be upper bounded by any choice of $\rho \in \mathcal{P}(\Theta)$. Letting $\Pi(B_n) := \int_{B_n}\pi(\theta)\mathrm{d}\theta$, we take $\rho(\theta) = \pi(\theta)/\Pi(B_n)$ for $\theta \in B_n$ and $\rho(\theta) = 0$ for $\theta \notin B_n$. Then Assumption 1 part (2) ensures that $\int_{B_n}|D(\theta)-D(\theta_*)|\rho(\theta)\mathrm{d}\theta \leq \alpha_1/\sqrt{n}$ and that $\mathrm{KL}(\rho\|\pi) = \int_\Theta \log\left(\rho(\theta)/\pi(\theta)\right)\rho(\theta)\mathrm{d}\theta = \int_{B_n} -\log(\Pi(B_n))\pi(\theta)\mathrm{d}\theta/\Pi(B_n) = -\log\Pi(B_n) \leq \alpha_2\sqrt{n}$. Thus

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta \leq D(\theta_*) + \left(\alpha_1+\alpha_2+\frac{2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta}\right)\frac{1}{\sqrt{n}}, \tag{3.17}$$

with probability at least $1-\delta$, as claimed.

Next, we complete the main proof. It follows from (3.17) and the simple upper bound $\alpha_1+\alpha_2+\frac{2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta} \leq \frac{\alpha_1+\alpha_2+2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta}$ for $0<\delta<1$ that

$$\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta - D(\theta_*) \leq \frac{\alpha_1+\alpha_2+2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}.$$

Since θ_* minimises f, the left-hand side is non-negative. With the absolute value of the left-hand side taken, the same concentration inequality holds as follows:

$$\mathbb{P}\left(\left|\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta - D(\theta_*)\right| \leq \frac{\alpha_1+\alpha_2+2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}\right) \geq 1-\delta.$$

Finally, since $D(\theta)-D(\theta_*) \geq 0$ for any $\theta\in\Theta$ as $D(\theta_*)$ is the minimum of f, it holds that

$$\left|\int_\Theta D(\theta)\pi_n^D(\theta)\mathrm{d}\theta - D(\theta_*)\right| = \left|\int_\Theta D(\theta)-D(\theta_*)\pi_n^D(\theta)\mathrm{d}\theta\right| = \int_\Theta |D(\theta)-D(\theta_*)|\,\pi_n^D(\theta)\mathrm{d}\theta.$$

Therefore, the last concentration inequality above holds for the quantity of the right-hand side, which completes the proof. □

3.5.3. Proof of Lemma 1

Proof. Davidson (1994, Theorem 21.8) showed that $D_n \xrightarrow{a.s.} D$ uniformly on U if and only if (a) $D_n \xrightarrow{a.s.} D$ pointwise on U and (b) $\{D_n\}_{n=1}^\infty$ is strongly stochastically equicontinuous on U. The condition (a) is implied by the precondition C1 of Assumption 3 and hence the condition (b) is shown in the remainder. By Davidson (1994, Theorem 21.10), $\{D_n\}_{n=1}^\infty$ is strongly stochastically equicontinuous on U if there exists a stochastic sequence $\{\mathcal{L}_n\}_{n=1}^\infty$

independent of θ s.t.

$$|D_n(\theta) - D_n(\theta')| \leq \mathcal{L}_n \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in U \qquad \text{and} \qquad \limsup_{n\to\infty} \mathcal{L}_n < \infty \text{ a.s.}$$

Since D_n is continuously differentiable on the set U by the precondition C2 of Assumption 3 with $r = 1$, the mean value theorem yields that

$$|D_n(\theta) - D_n(\theta')| \leq \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2 \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in U.$$

Again by the precondition C2 of Assumption 3 with $r = 1$, we have $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2 < \infty$ a.s. Therefore, setting $\mathcal{L}_n = \sup_{\theta\in U} \|\nabla_\theta D_n(\theta)\|_2$ concludes the proof. □

3.5.4. Proof of Lemma 2

Proof. The strong consistency $\theta_n \overset{a.s.}{\to} \theta_*$ is shown by an argument similar to van der Vaart (1998, Theorem 5.7). First, it follows from Lemma 1 that $D_n \overset{a.s.}{\to} D$ uniformly on U under assumption. Thus, for all n sufficiently large, we can take $\delta > 0$ s.t. $|D_n(\theta) - D(\theta)| < \delta/2$ a.s. over $\theta \in U$, which in turn leads to (a) $D(\theta) < D_n(\theta) + \delta/2$ and (b) $D_n(\theta) < D(\theta) + \delta/2$ a.s. over $\theta \in U$. Then applying both (a) and (b), the following bound on $D(\theta_n)$ holds for all n sufficiently large:

$$D(\theta_n) \overset{(a)}{<} D_n(\theta_n) + \delta/2 \overset{(*)}{\leq} D_n(\theta_*) + \delta/2 \overset{(b)}{<} D(\theta_*) + \delta \quad \text{a.s.} \tag{3.18}$$

where the second inequality $(*)$ follows from the fact that θ_n is the minimiser of D_n. Since $\inf_{\theta\in\Theta} D(\theta)$ is uniquely attained at $\theta_* \in U$ by the precondition C3 of Assumption 3, we can take any $\epsilon > 0$ to see that $D(\theta) - D(\theta_*) > 0$ for all $\theta \in \Theta \setminus B_\epsilon(\theta_*)$. Given an arbitrary $\epsilon > 0$, define $B_\epsilon(\theta_*) := \{\theta \in \Theta : \|\theta - \theta_*\| < \epsilon\}$ and let $\delta = \inf_{\theta\in\Theta\setminus B_\epsilon(\theta_*)} D(\theta) - D(\theta_*) > 0$. It then follows from (3.18) with this δ plugged-in that, for all n sufficiently large,

$$D(\theta_n) < \inf_{\theta\in\Theta\setminus B_\epsilon(\theta_*)} D(\theta) \quad \text{a.s.}$$

This implies that $\theta_n \in B_\epsilon(\theta_*)$ a.s. for any $\epsilon > 0$ arbitrary small for all n sufficiently large. Therefore $\theta_n \overset{a.s.}{\to} \theta_*$ by definition of convergence. □

3.5.5. Proof of Theorem 2

Proof. We show that (Miller, 2021, Theorem 4) holds a.s. under the preconditions C1-C5 of Assumption 3. In order to apply (Miller, 2021, Theorem 4), we first extend π and D_n from Θ to $\mathbb{R}^p$ by setting $\pi(\theta) = 0$ and $D_n(\theta) = \sup_{\theta\in\Theta} |D_n(\theta)| + 1$ for all $\theta \in \mathbb{R}^p \setminus \Theta$, so that we have $\pi : \mathbb{R}^p \to \mathbb{R}$, $D_n : \mathbb{R}^p \to \mathbb{R}$ and $\pi_n^D : \mathbb{R}^p \to \mathbb{R}$. Note that, in (Miller, 2021, Theorem 4), $\{D_n\}_{n=1}^\infty$ is regarded as a sequence of deterministic functions, while here $\{D_n\}_{n=1}^\infty$ is a sequence of stochastic functions dependent on random data $\{X_i\}_{i=1}^n$. It will be shown that (Miller, 2021, Theorem 4) holds a.s. for the stochastic sequence $\{D_n\}_{n=1}^\infty$.

We hence verify that the following prerequisites (1)–(6) of (Miller, 2021, Theorem 4) hold a.s., where we define $B_\epsilon(\theta_*) := \{\theta \in \Theta : \|\theta - \theta_*\| < \epsilon\}$ and recall $H_n(\theta) = \nabla^2_\theta D_n(\theta)$ and $H_* = \lim_{n\to\infty} H_n(\theta_*)$ from Assumption 3:

1. the prior density π is continuous at θ_* and $\pi(\theta_*) > 0$.
2. $\theta_n \overset{a.s.}{\to} \theta_*$.
3. the Taylor expansion $D_n(\theta) = D_n(\theta_n) + (1/2)(\theta - \theta_n) \cdot H_n(\theta_n)(\theta - \theta_n) + r_n(\theta - \theta_n)$ holds on U a.s. where r_n is the reminder term.
4. the remainder r_n of the Taylor expansion satisfies that $|r_n(\theta)| \le C\|\theta\|_2^3,\ \forall\theta \in B_\epsilon(0)$ a.s. for all n sufficiently large and some $\epsilon > 0$.
5. $H_n(\theta_n) \overset{a.s.}{\to} H_*$, $H_n(\theta_n)$ is symmetric for all n sufficiently large and H_* is positive definite.
6. $\liminf_{n\to\infty} \left(\inf_{\theta\in\mathbb{R}^p\setminus\mathcal{B}_\epsilon(\theta_n)} D_n(\theta) - D_n(\theta_n) \right) > 0$ a.s. for any $\epsilon > 0$.

Part (1): The precondition C5 of Assumption 3.

Part (2): Lemma 2.

Part (3): From the precondition C2 of Assumption 3, D_n is 3 times continuously differentiable over U. Noting that $\nabla_\theta D_n(\theta) = 0$ at a minimiser θ_n of D_n, the Taylor expansion of D_n around the minimiser θ_n gives that $D_n(\theta) = D_n(\theta_n) + \frac{1}{2}(\theta - \theta_n) \cdot H_n(\theta_n)(\theta - \theta_n) + r_n(\theta - \theta_n)$ where r_n is the remainder of the Taylor expansion.

Part (4): Since r_n is the remainder of the Taylor expansion, we have an upper bound

$$|r_n(\theta - \theta_n)| \le \frac{1}{6} \sup_{\theta\in U} \|\nabla^3_\theta D_n(\theta)\|_2 \|\theta - \theta_n\|_2^3, \quad \forall\theta \in U.$$

The precondition C2 of Assumption 3 guarantees that $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla^3_\theta D_n(\theta)\|_2 < \infty$ a.s. It is thus possible to take some positive constant C s.t. $(1/6)\sup_{\theta\in U} \|\nabla^3_\theta D_n(\theta)\|_2 \le C$ a.s. for all n sufficiently large. For all n sufficiently large, there exists some open ϵ-neighbour $B_\epsilon(\theta_n)$ contained in the open set U since $\theta_n \in U$. Combining these two facts concludes that

$$|r_n(\theta - \theta_n)| \le C\|\theta - \theta_n\|_2^3, \quad \forall\theta \in B_\epsilon(\theta_n) \quad \Longrightarrow \quad |r_n(\theta)| \le C\|\theta\|_2^3, \quad \forall\theta \in B_\epsilon(0)$$

holds for some $\epsilon > 0$.

Part (5): We first show that $\|H_n(\theta_n) - H_*\|_2 \overset{a.s.}{\to} 0$. By the triangle inequality,

$$\|H_n(\theta_n) - H_*\|_2 \le \|H_n(\theta_n) - H_n(\theta_*)\|_2 + \|H_n(\theta_*) - H_*\|_2 .$$

For the first term, it follows from the mean value theorem that

$$\|H_n(\theta_n) - H_n(\theta_*)\|_2 \leq \sup_{\theta \in U} \|\nabla_\theta H_n(\theta)\|_2 \|\theta_n - \theta_*\|_2 = \sup_{\theta \in U} \|\nabla_\theta^3 D_n(\theta)\|_2 \|\theta_n - \theta_*\|_2.$$

The precondition C2 of Assumption 3 guarantees that $\limsup_{n\to\infty} \sup_{\theta\in U} \|\nabla_\theta^3 D_n(\theta)\|_2 < \infty$ a.s. It is thus possible to take some positive constant C' s.t. $\|H_n(\theta_n) - H_n(\theta_*)\|_2 \leq C'\|\theta_n - \theta_*\|_2$ for all n sufficiently large. Then we have $\|H_n(\theta_n) - H_n(\theta_*)\|_2 \overset{a.s.}{\to} 0$ by the preceding part (2) $\theta_n \overset{a.s.}{\to} \theta_*$. For the second term, it is directly implied by the precondition C4 of Assumption 3 that $\|H_n(\theta_*) - H_*\|_2 \overset{a.s.}{\to} 0$. Combining these two facts concludes that $\|H_n(\theta_n) - H_*\|_2 \overset{a.s.}{\to} 0$. We next show that $H_n(\theta_n)$ is symmetric. The (i,j) entry of $H_n(\theta) = \nabla_\theta^2 D_n(\theta)$ is given by the partial derivative $(\partial^2/\partial\theta_i\partial\theta_j)D_n(\theta)$ with respect to i-th and j-th entry of θ. Since D_n is twice continuously differentiable by the precondition C2 of Assumption 3, the Schwartz's theorem implies that the commutation $(\partial^2/\partial\theta_i\partial\theta_j)D_n(\theta) = (\partial^2/\partial\theta_j\partial\theta_i)D_n(\theta)$ holds and therefore $H_n(\theta)$ is symmetric for any $\theta \in \Theta$. Finally, we show positive definiteness of H_*. For all n sufficiently large, $H_n(\theta_n)$ is positive semi-definite by the fact that θ_n is the minimiser of D_n and accordingly the limit H_* is positive semi-definite. Then H_* is positive definite since H_* is nonsingular by the precondition C4 of Assumption 3.

Part (6): It holds for any sequence $a_n, b_n \in \mathbb{R}$ that $\liminf_{n\to\infty}(a_n - b_n) \geq \liminf_{n\to\infty} a_n + \liminf_{n\to\infty}(-b_n)$. Furthermore, from the property that $\liminf_{n\to\infty}(-b_n) = -\limsup_{n\to\infty} b_n$, we have $\liminf_{n\to\infty}(a_n - b_n) \geq \liminf_{n\to\infty} a_n - \limsup_{n\to\infty} b_n$. Applying this, we have

$$\liminf_{n\to\infty} \left(\inf_{\theta\in\mathbb{R}^p \setminus B_\epsilon(\theta_n)} D_n(\theta) - D_n(\theta_n) \right) = \underbrace{\liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^p\setminus B_\epsilon(\theta_n)} D_n(\theta)}_{=:(*_1)} - \underbrace{\limsup_{n\to\infty} D_n(\theta_n)}_{=:(*_2)}.$$

For the first term $(*_1)$, it is obvious from the way of extending D_n from Θ to $\mathbb{R}^p$ that

$$(*_1) = \liminf_{n\to\infty} \inf_{\theta\in\mathbb{R}^p\setminus B_\epsilon(\theta_n)} D_n(\theta) \geq \liminf_{n\to\infty} \inf_{\theta\in\Theta\setminus B_\epsilon(\theta_n)} D_n(\theta) \quad \text{a.s.}$$

For any set $A \subset \mathbb{R}^p$ and function $g : \mathbb{R}^p \to \mathbb{R}$, define $\inf_{\theta\in A\setminus B_\epsilon(\theta_n)} g(\theta) := \sup_{\theta\in A} g(\theta)$ if $A \setminus B_\epsilon(\theta_n)$ is empty. Decomposing Θ into two sets U and $\Theta \setminus U$ leads to

$$(*_1) \geq \liminf_{n\to\infty} \inf_{\theta\in\Theta\setminus B_\epsilon(\theta_n)} D_n(\theta) \geq \min\left(\underbrace{\liminf_{n\to\infty} \inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta)}_{=:(*_{11})}, \underbrace{\liminf_{n\to\infty} \inf_{\theta\in\Theta\setminus(U\cup B_\epsilon(\theta_n))} D_n(\theta)}_{=:(*_{12})} \right) \text{ a.s.}$$

For the term $(*_{11})$, since $D_n \overset{a.s.}{\to} D$ uniformly on U by Lemma 1 and $\theta_n \overset{a.s.}{\to} \theta_*$ by the preceding part (2),

$$(*_{11}) = \liminf_{n\to\infty} \inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta) = \lim_{n\to\infty} \inf_{\theta\in U\setminus B_\epsilon(\theta_n)} D_n(\theta) = \inf_{\theta\in U\setminus B_\epsilon(\theta_*)} D(\theta) \quad \text{a.s.}$$

For the term $(*_{12})$, since the global minimiser θ_n of D_n is contained in U a.s. for all n sufficiently large by the precondition C3 of Assumption 3,

$$(*_{12}) = \liminf_{n\to\infty} \inf_{\theta\in\Theta\setminus(U\cup B_\epsilon(\theta_n))} D_n(\theta) > \liminf_{n\to\infty} \inf_{\theta\in U} D_n(\theta) = \inf_{\theta\in U} D(\theta) = D(\theta_*) \quad \text{a.s.}$$

where the second equality follows from the a.s. uniform convergence of D_n on U by Lemma 1. For the second term $(*_2)$, again since $D_n \overset{a.s.}{\to} D$ uniformly on U and $\theta_n \overset{a.s.}{\to} \theta_*$, we have

$$(*_2) = \limsup_{n\to\infty} D_n(\theta_n) = \lim_{n\to\infty} D_n(\theta_n) = D(\theta_*) \quad \text{a.s.}$$

The original term $(*_1)-(*_2)$ is lower bounded by $(*_1)-(*_2) \geq \min((*_{11})-(*_2), (*_{12})-(*_2))$ a.s., and both the term $(*_{11}) - (*_2)$ and $(*_{12}) - (*_2)$ are then further lower bounded by

$$(*_{11}) - (*_2) = \inf_{\theta\in U\setminus B_\epsilon(\theta_*)} D(\theta) - D(\theta_*) > 0 \quad \text{and} \quad (*_{12}) - (*_2) > D(\theta_*) - D(\theta_*) = 0 \quad \text{a.s.},$$

where the first inequality follows from the precondition C3 of Assumption 3 indicating that $\inf_{\theta\in\Theta} D(\theta)$ is uniquely attained at $\theta_* \in U$. Therefore, we have $(*_1) - (*_2) \geq \min((*_{11}) - (*_2), (*_{12}) - (*_2)) > 0$ a.s., which concludes the proof. □

3.5.6. Proof of Lemma 3

Proof. First of all, it holds as demonstrated in (17) of Ghosh and Basu (2016) that

$$\mathrm{PIF}(y, \theta, \mathbb{P}_n) = \beta n \pi_n^D(\theta) \left(-D_0(y, \theta, \mathbb{P}_n) + \int_\Theta D_0(y, \theta', \mathbb{P}_n) \pi_n^D(\theta') \mathrm{d}\theta' \right).$$

By Jensen's inequality, we have an upper bounded

$$\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\,\mathrm{PIF}(y, \theta, \mathbb{P}_n)| \leq \beta n \sup_{\theta\in\Theta} \pi_n^D(\theta) \left(\sup_{y\in\mathcal{X}} |D_0(y, \theta, \mathbb{P}_n)| + \int_\Theta \sup_{y\in\mathcal{X}} |D_0(y, \theta', \mathbb{P}_n)|\, \pi_n^D(\theta') \mathrm{d}\theta' \right).$$

Recall that $\pi_n^D(\theta) = \pi(\theta)\exp(-\beta n D(\theta; \mathbb{P}_n))/Z$ where $0 < Z < \infty$ is the normalising constant. An upper bound $\pi_n^D(\theta) \leq \pi(\theta)\exp(-\beta n \inf_{\theta\in\Theta} D(\theta; \mathbb{P}_n))/Z =: C\pi(\theta)$ holds for some $0 < C < \infty$, since $D(\theta; \mathbb{P}_n)$ is lower bounded by assumption and n is fixed. From this upper bound, we have

$$\begin{aligned}
&\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\,\mathrm{PIF}(y, \theta, \mathbb{P}_n)| \leq \beta n C \sup_{\theta\in\Theta} \pi(\theta) \left(\sup_{y\in\mathcal{X}} |D_0(y, \theta, \mathbb{P}_n)| + C \int_\Theta \sup_{y\in\mathcal{X}} |D_0(y, \theta', \mathbb{P}_n)|\, \pi(\theta') \mathrm{d}\theta' \right) \\
&\leq \beta n C \sup_{\theta\in\Theta} \left(\pi(\theta) \sup_{y\in\mathcal{X}} |D_0(y, \theta, \mathbb{P}_n)| \right) + \beta n C^2 \left(\sup_{\theta\in\Theta} \pi(\theta) \right) \int_\Theta \sup_{y\in\mathcal{X}} |D_0(y, \theta', \mathbb{P}_n)|\, \pi(\theta') \mathrm{d}\theta'.
\end{aligned}$$

Since $\sup_{\theta\in\Theta} \pi(\theta)$ in the second term is finite by assumption, the conditions 1 and 2 in the statement are sufficient for $\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\,\mathrm{PIF}(y, \theta, \mathbb{P}_n)| < \infty$, as claimed. □

3.5.7. Proof of Theorem 3

Proof. We first calculate the Fisher divergence between the generalised posterior π_n^D and an empirical distribution δ_θ^B of the bootstrap minimisers $\{\theta_n^{(b)}\}_{b=1}^B$, and then minimise it as a function of the weighting constant β. Recall that the score-matching divergence (Hyvärinen, 2005) is given by

$$\mathrm{D}(\pi_n^D \| \delta_\theta^B) = \frac{1}{B}\sum_{b=1}^{B} \left\| \nabla_\theta \log \pi_n^D(\theta_n^{(b)}) \right\|^2 + 2\,\mathrm{Tr}\left(\nabla_\theta^2 \log \pi_n^D(\theta_n^{(b)}) \right).$$

The score function of π_n^D is given by $\nabla_\theta \log \pi_n^D(\theta) = -\beta \nabla_\theta D_n(\theta) + \nabla_\theta \log \pi(\theta)$, which is independent of the normalising constant of π_n^D. Similarly, the second derivative is $\nabla_\theta^2 \log \pi_n^D(\theta) = -\beta \nabla_\theta^2 D_n(\theta) + \nabla_\theta^2 \log \pi(\theta)$. Therefore, the derivative terms in the Fisher divergence is written as

$$\begin{aligned}
\nabla_\theta \log \pi_n^D(\theta_n^{(b)}) &= \beta^2 \|\nabla_\theta D_n(\theta_n^{(b)})\|^2 - 2\beta \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \|\nabla_\theta \log \pi(\theta_n^{(b)})\|^2, \\
\nabla_\theta^2 \log \pi_n^D(\theta_n^{(b)}) &= -\beta\,\mathrm{Tr}\left(\nabla_\theta^2 D_n(\theta_n^{(b)}) \right) + \mathrm{Tr}\left(\nabla_\theta^2 \log \pi(\theta_n^{(b)}) \right).
\end{aligned}$$

Now consider minimising the Fisher divergence $\mathrm{D}(\pi_n^D \| \delta_\theta^B)$ with respect to the weighting constant β. Plugging the derivative terms in the Fisher divergence, we have

$$\mathrm{D}(\pi_n^D \| \delta_\theta^B) = \frac{1}{B}\sum_{b=1}^{B} \beta^2 \|\nabla_\theta D_n(\theta_n^{(b)})\|^2 - 2\beta \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) - 2\beta\,\mathrm{Tr}\left(\nabla_\theta^2 D_n(\theta_n^{(b)}) \right) + C$$

where we denote any term independent of β by C in this proof. Exchanging the order of the summation and the constant β, the Fisher divergence turns out to be a quadratic function of β:

$$\begin{aligned}
\mathrm{D}(\pi_n^D \| \delta_\theta^B) &= \beta^2 \underbrace{\frac{1}{B}\sum_{b=1}^{B} \|\nabla_\theta D_n(\theta_n^{(b)})\|^2}_{=(a)} - 2\beta \underbrace{\frac{1}{B}\sum_{b=1}^{B} \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}\left(\nabla_\theta^2 D_n(\theta_n^{(b)}) \right)}_{=(b)} + C \\
&= a\beta^2 - 2b\beta + C = a\left(\beta - \frac{b}{a} \right)^2 - \frac{b^2}{4a^2} + C
\end{aligned}$$

where the last equality follows from completing the square. Therefore, the Fisher divergence $\mathrm{D}(\pi_n^D \| \delta_\theta^B)$ is minimised at $\beta_* = b/a$, that is,

$$\beta_* = \frac{\sum_{b=1}^{B} \nabla_\theta D_n(\theta_n^{(b)}) \cdot \nabla_\theta \log \pi(\theta_n^{(b)}) + \mathrm{Tr}\left(\nabla_\theta^2 D_n(\theta_n^{(b)}) \right)}{\sum_{b=1}^{B} \|\nabla_\theta D_n(\theta_n^{(b)})\|^2},$$

as claimed, where the denominator and numerator are positive immediately from the first and second assumption respectively, which assures that $\beta_* > 0$. □

Chapter 4. Case I: Approach to Continuous Intractable Models

In this chapter, we consider the SD-Bayes approach to intractable models in continuous domains, such as $\mathcal{X} = \mathbb{R}^d$. Specifically, the methodology is developed for a particular Stein discrepancy called *kernel Stein discrepancy* (KSD), and we call the resulting generalised Bayesian approach *KSD-Bayes*. This is the first generalised Bayesian approach to inference for models that involve an intractable likelihood. It is shown that KSD-Bayes (1) provides robustness to model misspecification; (2) produces a generalised posterior that is tractable for any standard MCMC algorithms, or even closed-form when an appropriate conjugate prior (which we identify) is used together with an exponential family model; (3) satisfies several desirable theoretical properties, including a Bernstein–von Mises result which holds irrespective of whether the model is correctly specified. These results appear to represent a compelling case for the use of KSD-Bayes as an alternative to standard Bayesian inference with intractable likelihood. However, KSD-Bayes is no panacea and caution must be taken to avoid certain pathologies of KSD-Bayes, which we highlight in Section 4.2.3.

The chapter is structured as follows: Section 4.1 provides the construction of KSD. In addition, guidance for a choice of kernel to use in KSD-Bayes is contained. Section 4.2 presents the KSD-Bayes methodology, including the fully conjugate inference achieved by KSD-Bayes for exponential family models and the computational aspect of the non-conjugate inference. Section 4.3 establishes asymptotic properties and global bias-robustness of KSD-Bayes. Empirical assessment of KSD-Bayes with four distinct experiments are outlined in Section 4.4. Finally, Section 4.6 contains all deferred proofs of theoretical results presented in this chapter.

4.1. Kernel Stein Discrepancy

Compared to other Stein discrepancies, KSDs are attractive because they enable the supremum in (2.6) to be be explicitly computed. To define KSD, we require the concept of a (matrix-valued) *kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$; the precise definition was introduced in Chapter 2. For our purposes in this section, it suffices to point out that any kernel K has a uniquely associated Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}^d$, called a vector-valued reproducing kernel Hilbert space (RKHS). This RKHS constitutes the Stein set in KSD, and we therefore denote this RKHS as $\mathcal{H}$. The associated norm and inner product will respectively be denoted $\|\cdot\|_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

4.1.1. Construction

Let $\mathcal{S}_{\mathbb{Q}}$ be a Stein operator and denote the action of $\mathcal{S}_{\mathbb{Q}}$ on both the first and second argument[1] of a kernel K as $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K$. The following result is a generalisation of the original construction of KSD (Chwialkowski et al., 2016; Liu et al., 2016; Wang et al., 2019) to general Stein operators $\mathcal{S}_Q$ and general domains $\mathcal{X}$.

Assumption 4. *Let $\mathcal{H}$ be a RKHS with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$. For $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, let $\mathcal{S}_{\mathbb{Q}}$ be a Stein operator with domain $\mathcal{H}$. For each fixed $x \in \mathcal{X}$, we assume $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$. Further, we assume that $\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X)\right] < \infty$.*

Proposition 2 (Closed Form of Stein Discrepancy)**.** *Under Assumption 4, we have*

$$\mathrm{SD}^2(\mathbb{Q}\|\mathbb{P}) = \mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X')\right] \tag{4.1}$$

where X and X' are independent.

The proof is contained in Section 4.6.1. Proposition 2 shows that the supremum term in (2.6) is attained in closed-form by the expectation of $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(x,x')$, which is the definition of the KSD. In our context, KSD is used for inference of a parametric model $\mathbb{P}_\theta$. It is clear from Proposition 2 that KSD between between a parametric model $\mathbb{Q} = \mathbb{P}_\theta$ and an empirical distribution $\mathbb{P} = \mathbb{P}_n$ is available in closed form as

$$\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j). \tag{4.2}$$

The explicit form of $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K$ depends on the Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$. We advocate the default use of the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in (2.5) for the case $\mathcal{X} = \mathbb{R}^d$, which leads to

$$\begin{aligned}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') = {}& \nabla\log p_\theta(x)\cdot K(x,x')\nabla\log p_\theta(x') + \nabla_x\cdot(\nabla_{x'}\cdot K(x,x')) \\ & + \nabla\log p_\theta(x)\cdot(\nabla_{x'}\cdot K(x,x')) + (\nabla_x\cdot K(x,x'))\cdot\nabla\log p_\theta(x')\end{aligned} \tag{4.3}$$

where p_θ is a p.d.f. for $\mathbb{P}_\theta \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. Clearly, this expression is straightforward to evaluate whenever we have access to derivatives of the kernel and the log density. If the derivatives are analytically tedious, the expression above is amenable to the use of automatic differentiation tools (Baydin et al., 2018) in practice. For maximum clarity, the vector calculus notation is expanded as follows:

$$\begin{aligned}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') = {}& \sum_{i,j=1}^{d}\frac{\partial}{\partial x_{(i)}}\log p_\theta(x)\left[K(x,x')\right]_{(i,j)}\frac{\partial}{\partial x_{(j)}}\log p_\theta(x) + \frac{\partial^2}{\partial x_{(i)}\partial x'_{(j)}}\left[K(x,x')\right]_{(i,j)} \\ & + \frac{\partial}{\partial x_{(i)}}\log p_\theta(x)\frac{\partial}{\partial x'_{(j)}}\left[K(x,x')\right]_{(i,j)} + \frac{\partial}{\partial x'_{(j)}}\log p_\theta(x')\frac{\partial}{\partial x_{(i)}}\left[K(x,x')\right]_{(i,j)}\end{aligned}$$

[1] More precisely, denoting the j-th column of $K(x,x') \in \mathbb{R}^{d\times d}$ by $K_{-,j}(x,x') \in \mathbb{R}^d$, we define $\mathcal{S}_{\mathbb{Q}}K(x,x') := [\mathcal{S}_{\mathbb{Q}}K_{-,1}(x,x'),\ldots,\mathcal{S}_{\mathbb{Q}}K_{-,d}(x,x')] \in \mathbb{R}^d$ where $\mathcal{S}_{\mathbb{Q}}K_{-,j}(x,x') := \mathcal{S}_{\mathbb{Q}}[K_{-,j}(\cdot,x')](x)$ is an action of $\mathcal{S}_{\mathbb{Q}}$ for the $\mathbb{R}^d$-valued function $K_{-,j}(\cdot,x')$ at each $x' \in \mathcal{X}$. We further define $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(x,x') := \mathcal{S}_{\mathbb{Q}}[\,\mathcal{S}_{\mathbb{Q}}K(x,\cdot)\,](x')$ as an action of $\mathcal{S}_{\mathbb{Q}}$ for the $\mathbb{R}^d$-valued function $\mathcal{S}_{\mathbb{Q}}K(x,\cdot)$ at each $x \in \mathcal{X}$.

where we recall that K is a matrix-valued kernel.

Note that it is straightforward to verify the assumption that $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional for each fixed $x \in \mathcal{X}$ once the form of $\mathcal{S}_{\mathbb{Q}}$ is specified; see Section 4.6.2 to verify this for the case of the Langevin Stein operator.

4.1.2. Recommended Choice of Kernel Function

Any methodology based on RKHS entails an appropriate choice of kernel K that determines underlying properties of the methodology. Conveniently, for Euclidean domains $\mathcal{X} = \mathbb{R}^d$, there exists guiding results that motivate a specific choice of kernel K in KSD-Bayes. To use KSD in SD-Bayes, we advocate the default use of a kernel of the form

$$K(x, x') = \frac{M(x)M(x')^\top}{(1 + (x - x')^\top \Sigma^{-1}(x - x'))^\gamma}, \tag{4.4}$$

where $\gamma \in (0, 1)$ is a constant, Σ is a positive definite matrix, and $M \in C_b^1(\mathbb{R}^d; \mathbb{R}^{d\times d})$ will be called a matrix-valued *weighting function*[2]. For $M(x) = I_d$, (4.4) is called an *inverse multi-quadratic* (IMQ) kernel. A pair of the IMQ kernel K and the Langevin Stein operator $\mathcal{S}_{\mathbb{Q}}$ has several appealing properties in the context of KSD. Firstly, under mild conditions on $\mathbb{Q}$, the convergence $\mathrm{KSD}(\mathbb{Q}||\mathbb{P}_n) \to 0$ implies that $\mathbb{P}_n$ converges weakly to $\mathbb{Q}$ (Chen et al., 2019, Theorem 4). This convergence control ensures that small values of $\mathrm{KSD}(\mathbb{P}_\theta||\mathbb{P}_n)$ imply similarity between $\mathbb{P}_\theta$ and $\mathbb{P}_n$ in the topology of weak convergence, so that minimising KSD is meaningful. Note that other common kernels (e.g., Gaussian or Matérn kernels) fail to provide such convergence control (Gorham and Mackey, 2017, Theorem 6). Secondly, and on a more practical level, the combination of Stein operator and IMQ kernel, with $\gamma = 1/2$, was found to work well in previous studies (Chen et al., 2019; Riabiz et al., 2021); we therefore also recommend $\gamma = 1/2$ as a default.

The weighting function $M(x)$ facilitates a trade-off between efficiency and robustness of inference. Recall that Section 3.3 established the global bias-robustness property of generalised posteriors. Subsequently Section 4.3 in this chapter derives a condition for KSD-Bayes to be globally bias-robust that depends on the choice of the weighting function $M(x)$. If global bias robustness is *not* required, then we recommend setting $M(x) = I_d$ as a default, which enjoys the aforementioned properties of KSD. If global bias-robustness *is* required, then we recommend selecting $M(x)$ such that the derived condition in subsequent Theorem 4 are satisfied; see the worked examples in Section 4.4.

The theoretical analysis of Section 4.3 assumed that K is fixed, but in our experiments we follow standard practice in the kernel methods community and recommend a data-adaptive choice of the matrix Σ for better performance especially in high dimensional cases. All experiments we report used the ℓ_1-regularised sample covariance matrix estimator of

[2]The use of a non-constant weighting function is equivalent to replacing the Langevin Stein operator with a *diffusion* Stein operator whose *diffusion matrix* is $M(x)$; see Gorham et al. (2019).

Ollila and Raninen (2019). The sensitivity of KSD-Bayes to the choice of kernel parameters is investigated in Section A.1.

4.2. KSD-Bayes Methodology

We select KSD in SD-Bayes as a default choice of Stein discrepancy for continuous intractable likelihoods. The resulting generalised posterior will be referred to as the *KSD-Bayes* posterior. KSD is particularly attractive for SD-Bayes since it enables the generalised posterior in Definition 6 to be explicitly computed.

Definition 6 (KSD-Bayes)**.** *Given a Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ and a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$, select* $\mathrm{KSD}^2(\mathbb{P}_\theta\|\cdot)$ *for the Stein discrepancy in SD-Bayes. We define the resulting generalised posterior as the KSD-Bayes posterior:*

$$\pi_n^D(\theta) \propto \exp\left(-\beta n \,\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n)\right)\pi(\theta) \tag{4.5}$$

where π is a prior p.d.f. over $\theta \in \Theta$.

This is clearly a special case of SD-Bayes given $\mathrm{SD}^\gamma(\mathbb{P}_\theta\|\mathbb{P}_n) = \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n)$. Whether KSD-Bayes is reasonable or not hinges crucially on whether KSD is a meaningful way to quantify the difference between the empirical distribution $\mathbb{P}_n$ and the parametric model $\mathbb{P}_\theta$. As mentioned in Section 4.1.2, sufficient conditions for convergence control have been established for the Langevin Stein operator, under which the convergence of $\mathrm{KSD}(\mathbb{P}_\theta\|\mathbb{P}_n)$ implies the weak convergence of $\mathbb{P}_n$ to $\mathbb{P}_\theta$ (Gorham and Mackey, 2017, Theorem 8). This provides some preliminary assurance that KSD-Bayes may work. We present formal theoretical guarantees based on posterior consistency, BvM theorem, and global bias-robustness in Section 4.3. The specific choices of K for use in KSD-Bayes in Section 4.1.2 was motivated by these theoretical results.

4.2.1. Conjugate Inference for Exponential Family Models

The generalised posterior can be exactly computed in the case of an natural exponential family model when a conjugate prior is used. Let $\eta : \Theta \to \mathbb{R}^k$ and $t : \mathcal{X} \to \mathbb{R}^k$ be any sufficient statistic for some $k \in \mathbb{N}$ and let $a : \Theta \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}$. An exponential family model has p.m.f. or p.d.f. (with respect to an appropriate reference measure on $\mathcal{X}$) of the form

$$p_\theta(x) = \exp(\eta(\theta)\cdot t(x) - a(\theta) + b(x)). \tag{4.6}$$

This includes a wide range of distributions with an intractable normalisation constant $\exp(a(\theta))$, used in statistical applications such as random graph estimation (Yang et al., 2015), spin glass models (Besag, 1974) and the kernel exponential family model (Canu and Smola, 2006). The model in (4.6) is called *natural* when the canonical parametrisation $\eta(\theta) = \theta$ is employed.

Proposition 3. *Consider $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in (2.5), where $\mathbb{P}_\theta$ is the exponential family in (4.6), and a kernel $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Given a prior p.d.f. π, the KSD-Bayes posterior has a p.d.f. of the form*

$$\pi_n^D(\theta) \propto \exp\left(-\beta n\{\eta(\theta)\cdot\Lambda_n\eta(\theta) + \eta(\theta)\cdot\nu_n\}\right)\pi(\theta),$$

where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as

$$\Lambda_n := \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i)\cdot K(x_i,x_j)\nabla t(x_j),$$
$$\nu_n := \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i)\cdot(\nabla_{x_j}\cdot K(x_i,x_j)) + \nabla t(x_j)\cdot(\nabla_{x_i}\cdot K(x_i,x_j)) + 2\nabla t(x_i)\cdot K(x_i,x_j)\nabla b(x_j).$$

For a natural exponential family under $\eta(\theta) = \theta$, the prior $\pi(\theta) \propto \exp(-\frac{1}{2}(\theta-\mu)\cdot\Sigma^{-1}(\theta-\mu))$ leads to a generalised posterior of the closed-form

$$\pi_n^D(\theta) \propto \exp\left(-\frac{1}{2}(\theta-\mu_n)\cdot\Sigma_n^{-1}(\theta-\mu_n)\right),$$

where $\Sigma_n^{-1} := \Sigma^{-1} + 2\beta n\Lambda_n$ and $\mu_n := \Sigma_n^{-1}(\Sigma^{-1}\mu - \nu_n)$.

The proof is in Section 4.6.3. That the Gaussian distribution will be conjugate in KSD-Bayes, even in the presence of intractable model, is remarkable and notably different from the classical Bayesian case, albeit at a $\mathcal{O}(n^2)$ computational cost of Λ_n and ν_n. Strategies to further reduce this computational cost are discussed in Section 4.2.2. It is well known that certain minimum discrepancy estimators, such as the *score matching estimator* (Hyvärinen, 2005) and the *minimum KSD estimator* (Barp et al., 2019), have closed forms in the case of an exponential family models; it is similar reasoning that has led us to Proposition 3.

4.2.2. Non-Conjugate Inference and Computation

To access the generalised posterior in the non-conjugate case, essentially any existing MCMC algorithms for *tractable* distributions can be used. For example, the Gaussian form of the data-dependent term in Proposition 3 suggests that elliptical slice sampling may work well when the natural parametrisation of the exponential family is employed (Murray et al., 2010). Efficient gradient-based samplers, such as the Langevin Monte Carlo algorithm, can also be used whenever the gradient of (4.5) is available. The per-iteration computational cost appears to be $\mathcal{O}(n^2)$ since, for each state θ visited along the sample path, the KSD in (4.2) must be evaluated. However, various strategies enable this computational cost to be mitigated. For concreteness of the discussion that follows, we consider the Langevin Stein

operator, for which

$$\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) \overset{+C}{=} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \nabla \log p_\theta(x_i) \cdot K(x_i, x_j) \nabla \log p_\theta(x_j) + \nabla \log p_\theta(x_i) \cdot (\nabla_{x_j} \cdot K(x_i, x_j)) + (\nabla_{x_i} \cdot K(x_i, x_j)) \cdot \nabla \log p_\theta(x_j)$$

where the equality holds up to a θ-independent constant.

Memoisation: The above expression depends on θ only through the terms $\{\nabla \log p_\theta(x_i)\}_{i=1}^n$, of which there are $\mathcal{O}(n)$, while all other terms involving K, of which there are $\mathcal{O}(n^2)$, can be computed once and memoised. The double summation still necessitates $\mathcal{O}(n^2)$ computational cost but this operation is *embarrassingly parallel*.

Finite rank kernel: Computational cost can be reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ using a finite rank kernel. A useful and important example is the rank one kernel $K(x, x') = I_d$, which reduces the KSD in (4.2) to

$$(4.2) \overset{+C}{=} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \log p_\theta(x_i) \right\|^2$$

and is closely related to divergences used in *score matching* (Hyvärinen, 2005). Random finite rank approximations of the kernel can also considered in this context (Huggins and Mackey, 2018).

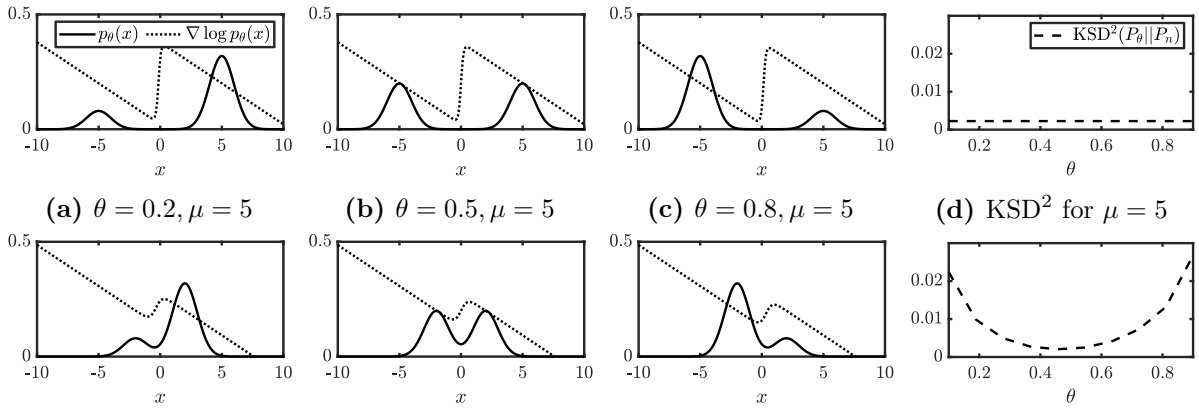
Stochastic approximation: The construction of low-cost unbiased estimators for (4.2) is straight-forward via sampling *mini-batches* from the dataset. This enables a variety of exact and approximate algorithms for posterior approximation to be exploited (e.g. Ma et al., 2015). Alternatively, Gorham et al. (2020); Huggins and Mackey (2018) argued for stochastic approximations of KSD that could be used.

4.2.3. Limitations of KSD-Bayes

A divergence $\mathrm{D}(\mathbb{Q}||\mathbb{P})$ induces an information geometry (Amari, 1997), encoding a particular sense in which $\mathbb{Q}$ can be considered to differ from $\mathbb{P}$. As such, all divergence exhibit *pathologies*, meaning that certain characteristics that distinguish $\mathbb{Q}$ from $\mathbb{P}$ are less easily detected. A documented pathology of gradient-based discrepancies, including the KSD, is their insensitivity to the existence of high-probability regions which are well-separated; see Gorham et al. (Section 5.1 2019) and Wenliang and Kanagawa (2021). To see this, consider a Gaussian mixture model

$$p_\theta(x) = \frac{\theta}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) + \frac{(1-\theta)}{\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2}\right) \tag{4.7}$$

where $\theta \in [0, 1]$ specifies the mixture ratio and $\mu \in \mathbb{R}$ controls the separation between the two components. If the two components are well-separated i.e. $\mu \gg 1$, the gradient $\nabla \log p_\theta$ becomes insensitive to θ and hence a gradient-based divergence such as KSD will



(a) $\theta = 0.2, \mu = 5$ **(b)** $\theta = 0.5, \mu = 5$ **(c)** $\theta = 0.8, \mu = 5$ **(d)** KSD^2 for $\mu = 5$

(e) $\theta = 0.2, \mu = 2$ **(f)** $\theta = 0.5, \mu = 2$ **(g)** $\theta = 0.8, \mu = 2$ **(h)** KSD^2 for $\mu = 2$

Figure 4.1 Illustrating the insensitivity to mixture proportions of KSD. Panels (a-c,e-g) display the density function $p_\theta(x)$ from (4.7) together with the gradient $\nabla \log p_\theta(x)$, the latter rescaled to fit onto the same plot. Panels (d,h) display the discrepancy $\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$, where $\mathbb{P}_n$ is an empirical distribution of $n = 1000$ samples from the model with $\theta = 0.5$.

be insensitive to θ, as demonstrated in Figure 4.1. For this reason, caution is warranted when gradient-based discrepancies are used. However, in practice direct inspection of the dataset and knowledge of how $\mathbb{P}_\theta$ is parametrised can be used to ascertain whether either distribution is multi-modal. Our applications in Section 4.4 are not expected to be multi-modal (with the exception of the kernel exponential family in Section 4.4.3 which was selected to demonstrate the insensitivity to mixing proportions of KSD-Bayes).

A second limitation of KSD-Bayes is non-invariance to a change of coordinates in the dataset. This is a limitation of loss-based estimators in general. In Section 4.1.2 we recommend a data-adaptive choice of kernel, which serves to provide approximate invariance to affine transformations of the dataset. As usual in statistical analyses, we recommend *post-hoc* assessment of the sensitivity of inferences to perturbations of the dataset. A third limitation of KSD-Bayes is the loss of efficiency that can occur in settings where the data are high-dimensional. Sliced versions of KSD have been proposed to address the curse of dimension for KSD (Gong et al., 2021), but to limit scope we do not consider the combination of sliced discepancies and KSD-Bayes in this work.

Despite these limitations, KSD-Bayes represents a flexible and effective procedure for generalised Bayesian inference in the context of an intractable likelihood. Our attention turns next to theoretical analysis of KSD-Bayes.

4.3. Theoretical Assessment

This section contains a comprehensive theoretical treatment of KSD-Bayes. The main results are posterior consistency and the Bernstein–von Mises theorem in Section 4.3.2, and global bias-robustness of the generalised posterior in Section 4.3.3. In obtaining these results we have developed novel intermediate results concerning an important V-statistic estimator for KSD; these are anticipated to be of independent interest, so we present these in Section 4.3.1. Note that all theory is valid for the misspecified regime where $\mathbb{P}$ need not

be an element of $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Moreover, the results in Section 4.3.1 and Section 4.3.2 hold for general data domains $\mathcal{X}$. For the entirety of this section we set $\beta = 1$, with all results for $\beta \neq 1$ immediately recovered by replacing K with βK. We use the following standing assumptions and additional notations.

Standing Assumptions: The dataset $\{x_i\}_{i=1}^n$ consists of independent samples generated from $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, with empirical distribution denoted $\mathbb{P}_n := (1/n)\sum_{i=1}^n \delta_{x_i}$. The set $\Theta \subseteq \mathbb{R}^p$ is open, convex and bounded[3]. Assumption 4 holds with $\mathbb{Q} = \mathbb{P}_\theta$ for every $\theta \in \Theta$.

Notation: For shorthand, let ∂^1, ∂^2 and ∂^3 denote the partial derivatives $(\partial/\partial\theta_{(h)})$, $(\partial^2/\partial\theta_{(h)}\partial\theta_{(k)})$ and $(\partial^3/\partial\theta_{(h)}\partial\theta_{(k)}\partial\theta_{(l)})$ for $h, k, l \in \{1, \ldots, p\}$, where to reduce notation the indices (h, k, l) are left implicit.

4.3.1. Minimum KSD Estimators

First we present novel analysis of the V-statistic $\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$. A related U-statistic estimator of KSD was analysed in Barp et al. (2019) but this is only an estimate of $\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})$, rendering it unsuitable for generalised Bayesian inference, which requires losses to be lower-bounded (Jewson et al., 2018). Furthermore, our results for the V-statistic do not depend on a specific form of $\mathcal{S}_{\mathbb{P}_\theta}$, in contrast to Barp et al. (2019) who considered the *diffusion* Stein operator, and may hence be of independent interest.

Despite the bias present in a V-statistic, our standing assumptions are sufficient to derive the following consistency result:

Lemma 4 (a.s. Pointwise Convergence)**.** *For each $\theta \in \Theta$,*

$$\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}) \xrightarrow{a.s.} 0.$$

The proof is contained in Section 4.6.4. If we impose further regularity, we can obtain a uniform convergence result. It will be convenient to introduce a collection of assumptions that are indexed by $r_{\max} \in \{0, 1, 2, \ldots\}$, as follows:

Assumption 5 ($r_{\max}$)**.** *For all integers $0 \leq r \leq r_{\max}$, the following conditions hold:*

(1) the map $\theta \mapsto \partial^r \mathcal{S}_{\mathbb{P}_\theta}[h](x)$ exists and is continuous, for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$;

(2) the map $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_\theta})[h](x)$ is a continuous linear functional on $\mathcal{H}$, for each $x \in \mathcal{X}$;

(3) $\mathbb{E}_{X\sim\mathbb{P}}[\sup_{\theta\in\Theta}((\partial^r \mathcal{S}_{\mathbb{P}_\theta})(\partial^r \mathcal{S}_{\mathbb{P}_\theta})K(X, X))] < \infty$,

where $(\partial^0 \mathcal{S}_{\mathbb{P}_\theta}) := \mathcal{S}_{\mathbb{P}_\theta}$; note that (2) with $r = 0$ is implied from Standing Assumption 2.

In the expression above, the first and second operations of $(\partial^r \mathcal{S}_{\mathbb{P}_\theta})$ are applied, respectively, to the first and second argument of K, as with $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x, x)$. These assumptions become

[3]The assumption that Θ is bounded is used only to simplify the statement of our results. For the case where Θ is not bounded, it is sufficient for Assumptions 5 and 6 to hold on an open, convex and bounded subset $U \subset \Theta$. Then Lemmas 5 and 6 hold on the bounded subset U, and all the other results hold on Θ.

concrete when considering a specific Stein operator. In the case of the Langevin Stein operator,

$$\partial^r \mathcal{S}_{\mathbb{P}_\theta}[h](x) = \partial^r \Big(\nabla_x \log p_\theta(x) \cdot h(x)\Big) + \partial^r \Big(\nabla_x \cdot h(x)\Big) = \Big(\partial^r \nabla_x \log p_\theta(x)\Big) \cdot h(x). \quad (4.8)$$

The operator $\partial^r \mathcal{S}_{\mathbb{P}_\theta}$ in (4.8) is therefore well-defined and $\theta \mapsto \partial^r \mathcal{S}_{\mathbb{P}_\theta}[h](x)$ is continuous whenever $\theta \mapsto \nabla_x \log p_\theta(x)$ is r-times continuously differentiable over Θ. For each fixed $x \in \mathcal{X}$, it is clear that $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_\theta})[h](x)$ is a continuous linear functional on $\mathcal{H}$. Then the term $(\partial^r \mathcal{S}_{\mathbb{P}_\theta})(\partial^r \mathcal{S}_{\mathbb{P}_\theta})K(x,x)$ in the final part of Assumption 5 takes the explicit form

$$(\partial^r \mathcal{S}_{\mathbb{P}_\theta})(\partial^r \mathcal{S}_{\mathbb{P}_\theta})K(x,x) = \Big(\partial^r \nabla_x \log p_\theta(x)\Big) \cdot K(x,x) \Big(\partial^r \nabla_x \log p_\theta(x)\Big) \quad (4.9)$$

where the regularity of (4.9) depends on K and $\mathbb{P}_\theta$. The uniform convergence result is presented as follows.

Lemma 5 (a.s. Uniform Convergence)**.** *Suppose Assumption 5 ($r_{\max} = 1$) holds. Then*

$$\sup_{\theta \in \Theta} \Big| \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}) \Big| \xrightarrow{a.s.} 0.$$

The proof is contained in Section 4.6.5.

Our next results concern consistency and asymptotic normality of an estimator θ_n that minimises the V-statistic in (4.2).

Assumption 6. *There exist minimisers θ_n of $\mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P}_n)$ for all sufficiently large $n \in \mathbb{N}$, and there exists a unique $\theta_* = \inf_{\theta \in \Theta} \mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P})$.*

Lemma 6 (Strong Consistency)**.** *Suppose Assumptions 5 ($r_{\max} = 1$) and 6 hold. Then*

$$\theta_n \xrightarrow{a.s.} \theta_*.$$

The proof is contained in Section 4.6.6. For the well-specified case where $\exists \theta_0$ s.t. $\mathbb{P}_{\theta_0} = \mathbb{P}$, the uniqueness of θ_* holds automatically if KSD is a proper divergence i.e. $\mathrm{KSD}(\mathbb{P} \| \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$. For example, if the preconditions of Barp et al. (2019, Proposition 1) are satisfied and the parametrisation $\theta \mapsto \mathbb{P}_\theta$ is injective, the minimum is uniquely attained. Let $H_* := \nabla_\theta^2 \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})|_{\theta=\theta_*}$ and $J_* := \mathbb{E}_{X \sim \mathbb{P}}[S(X,\theta_*) S(X,\theta_*)^\top]$, where we define the column vector $S(x,\theta) := \mathbb{E}_{X \sim \mathbb{P}}[\nabla_\theta (\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x,X))]$. Asymptotic normality of θ_n can be established if further regularity is imposed:

Lemma 7 (Asymptotic Normality)**.** *Suppose Assumptions 5 ($r_{\max} = 3$) and 6 hold. If H_* is non-singular,*

$$\sqrt{n}\,(\theta_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, H_*^{-1} J_* H_*^{-1})$$

where $\xrightarrow{d}$ denotes the convergence in distribution.

The proof is contained in Section 4.6.7. Our main theoretical results on KSD-Bayes are presented next.

4.3.2. Posterior Consistency and Bernstein–von Mises

Armed with the results of Section 4.3.1, we now establish posterior consistency and the Bernstein–von Mises theorem of KSD-Bayes. In Section 3.2, we derived these for generalised posteriors of an abstract class. We apply the results in Section 3.2 for KSD-Bayes by verifying the provided preconditions.

It was established in Section 3.2 that posterior consistency of generalised posteriors holds under Assumption 1 (prior mass condition) and Assumption 2 (convergence rate). In most cases, Assumption 1 has to be directly posited due to the difficulty to the pre-diagnosis although it is not restrictive. We verify Assumption 2 for $D_n(\theta) = \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$ and $D(\theta) = \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})$. We previously showed in Section 3.2 that, if a loss is additive, Assumption 2 can be verified by Proposition 1. KSD is no longer an additive loss because of the "double summation" but it nonetheless elegantly satisfies Assumption 2 as follows.

Lemma 8. *For each $\theta \in \Theta$, with $\sigma(\theta) = 4\mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X))]$,*

$$\mathbb{E}_{X_1,\ldots,X_n}\left[\left|\,\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})\right|\right] \leq \frac{\sigma(\theta)}{\sqrt{n}}. \tag{4.10}$$

The proof is contained in Section 4.6.8. We establish posterior consistency of KSD-Bayes:

Proposition 4. *Suppose Assumption 1 holds under $D(\theta) = \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})$ and $\sigma(\theta) := 4\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]$ is bounded in Θ. Then, for all $\delta \in (0,1]$,*

$$\mathbb{P}\left(\int_\Theta \left|\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}) - \mathrm{KSD}^2(\mathbb{P}_{\theta_*}\|\mathbb{P})\right| \pi_n^D(\theta)\mathrm{d}\theta > \delta\right) \leq \frac{\alpha_1+\alpha_2+2\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}$$

where the probability is with respect to realisations of the dataset $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

Proof. The proof is a direct consequence of Theorem 1 that holds under Assumption 1 and Assumption 2. The former is assumed and the latter holds as Lemma 8. □

Next, we derive a Bernstein–von Mises result of KSD-Bayes built on the general result of Theorem 2. The following result is established by showing that Assumption 5 for KSD-Bayes implies all the conditions of Theorem 2.

Proposition 5. *Suppose Assumption 5 ($r_{\max} = 3$) and 6 holds. Suppose that a prior π is positive and continuous at θ_*. Denote by $\widetilde{\pi}_n^D$ a density on $\mathbb{R}^p$ of a random variable $\sqrt{n}(\theta-\theta_n)$ for $\theta \sim \pi_n^D$. If $H_* := \nabla_\theta^2 \mathrm{KSD}(\mathbb{P}_\theta\|\mathbb{P})|_{\theta=\theta_*}$ is nonsingular,*

$$\int_{\mathbb{R}^p}\left|\hat{\pi}_n^D(\theta) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}}\exp\left(-\frac{1}{2}\theta\cdot H_*\theta\right)\right|\mathrm{d}\theta \xrightarrow{a.s.} 0,$$

where the a.s. convergence is with respect to realisations of the dataset $\{x_i\}_{i=1}^n$.

Proof. We verify the precondition C1–C5 of Assumption 3. The precondition C1 follows from Lemma 4, C3 from the standing assumption and Assumption 6, C5 from part (1)

of Assumption 1. The precondition C2 is used for Lemma 5 and proven in intermediate Lemma 10 in subsequent Section 4.6. Finally, the precondition C4 is used for Lemma 7 and proven in intermediate Lemma 11 in subsequent Section 4.6. □

These positive results are encouraging, as they indicate the limitations of KSD-Bayes described in Section 4.2.3 are at worst a finite sample size effect. However, we note that the asymptotic precision matrix H_* from Proposition 5 differs to the precision matrix $H_* J_*^{-1} H_*$ of the minimum KSD estimator from Lemma 7; This is analogous to the case of standard Bayesian inference under model misspecification, where Bayesian credible sets can have asymptotically incorrect frequentist coverage if the statistical model is misspecified (Kleijn and van der Vaart, 2012b; Müller, 2013).

Remark 1. *The analysis in Sections 4.3.1 and 4.3.2 covers general domains $\mathcal{X}$ and Stein operators $\mathcal{S}_{\mathbb{P}}$. Henceforth, in this chapter, we restrict attention to $\mathcal{X} = \mathbb{R}^d$.*

4.3.3. Global Bias-Robustness of KSD-Bayes

An important property of KSD-Bayes is that, through a suitable choice of kernel K, the generalised posterior can be made robust to contamination in the dataset. Consider the *ε-contamination model* $\mathbb{P}_{n,\epsilon,y} = (1-\epsilon)\mathbb{P}_n + \epsilon\delta_y$, where $y \in \mathcal{X}$ and $\epsilon \in [0,1]$ (Huber and Ronchetti, 2009), where the datum y is considered to be contaminating the dataset $\{x_i\}_{i=1}^n$. In Section 3.3, we defined the global bias-robustness property of generalised posteriors that indicates a strong insensitivity to outliers in the dataset; see Lemma 3 for the condition to satisfy global bias-robustness. This global bias-robustness of KSD-Bayes will now be established.

Note again that standard Bayesian inference does not satisfy the conditions of Lemma 3 in general when $\mathcal{X} = \mathbb{R}^d$. Indeed, when $D(\theta)$ is the negative log likelihood, the quantity $D_0(y, \theta, \mathbb{P}_n)$ defined in Lemma 3 is given by $D_0(y, \theta, \mathbb{P}_n) = \log p_\theta(y) - \sum_{i=1}^n \log p_\theta(x_i)$, where the term $\log p_\theta(y)$ can easily violate the condition of Lemma 3 for a large class of models, including a light-tailed one such as a normal location model. In contrast, the kernel K in KSD-Bayes provides a degree of freedom which can be leveraged to ensure that the conditions of Lemma 3 *are* satisfied. The specific form of $D_0(y, \theta, \mathbb{P}_n)$ for KSD-Bayes is derived as

$$D_0(y, \theta, \mathbb{P}_n) = 2\mathbb{E}_{X\sim\mathbb{P}_n}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta} K(X, y)] - 2\mathbb{E}_{X,X'\sim\mathbb{P}_n}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta} K(X, X')].$$

See Section 4.6.9 for the derivation. This enables us to derive sufficient conditions on K for global bias-robustness of KSD-Bayes, which we now present.

Theorem 4 (Global Bias-Robustness of KSD-Bayes)**.** *For each $\theta \in \Theta$, let $\mathbb{P}_\theta \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and let $\mathcal{S}_{\mathbb{P}_\theta}$ denote the Langevin Stein operator in* (2.5)*. Let $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$.*

Suppose that π is bounded over Θ. If there exists a function $\gamma : \Theta \to \mathbb{R}$ such that

$$\sup_{y \in \mathbb{R}^d} \left(\nabla_y \log p_\theta(y) \cdot K(y,y) \nabla_y \log p_\theta(y) \right) \leq \gamma(\theta) \tag{4.11}$$

and $\sup_{\theta \in \Theta} |\pi(\theta)\gamma(\theta)| < \infty$ and $\int_\Theta \pi(\theta)\gamma(\theta) \mathrm{d}\theta < \infty$, then KSD-Bayes is globally bias-robust.

The proof is contained in Section 4.6.10. The preconditions of Theorem 4 can be satisfied through an appropriate choice of kernel K. In Section 4.1.2 we recommended the choice of kernel K of the form

$$K(x, x') = \frac{M(x)M(x')^\top}{(1 + (x - x')^\top \Sigma^{-1}(x - x'))^\gamma}.$$

Then the main quantity to use in checking the condition turns out to be

$$\sup_{y \in \mathbb{R}^d} \left(\nabla_y \log p_\theta(y) \cdot K(y,y) \nabla_y \log p_\theta(y) \right) = \sup_{y \in \mathbb{R}^d} \| M(y) \nabla_y \log p_\theta(y) \|^2.$$

It can be observed from the equation above that it suffices to ensure global bias-robustness of KSD-Bayes if we choose a function M that decays fast enough to cancel the growth of $\nabla \log p_\theta$. The difference in performance of robust and non-robust instances of KSD-Bayes is explored in detail in Section 4.4. A comparison of KSD-Bayes to existing robust generalised Bayesian methodologies for tractable likelihood can be found in Section A.4.

This completes our methodological and theoretical development, and next we turn to empirical performance assessment.

4.4. Empirical Assessment

In this section four distinct experiments are presented. The first experiment, in Section 4.4.1, concerns a normal location model, allowing the standard posterior and our generalised posterior to be compared and confirming our robustness results are meaningful. Section 4.4.2 presents a two-dimensional precision estimation problem, where standard Bayesian computation is challenging but computation with KSD-Bayes is trivial. Then, Section 4.4.3 presents a 25-dimensional kernel exponential family model, and Section 4.4.4 presents a 66-dimensional exponential graphical model. The kernel exponential family model allows us to explore a multi-modal dataset and to understand the potential limitations of KSD-Bayes in that context (c.f. Section 4.2.3). For all experiments, the default settings of kernel K in Section 4.1.2 were used. The approach of Lyddon et al. (2019) was adopted to select the weight β. For a well-specified normal location model in Section 4.4.1, the asymptotic variance of the KSD-Bayes posterior with $\beta = 1$ is theoretically never smaller than that of the standard posterior. This provides a heuristic motivation to restrict β to $(0, 1]$. We used this restriction as a safeguard against over-confidence of the KSD-Bayes posterior in the experiments in this section. The full detail of the selection of β

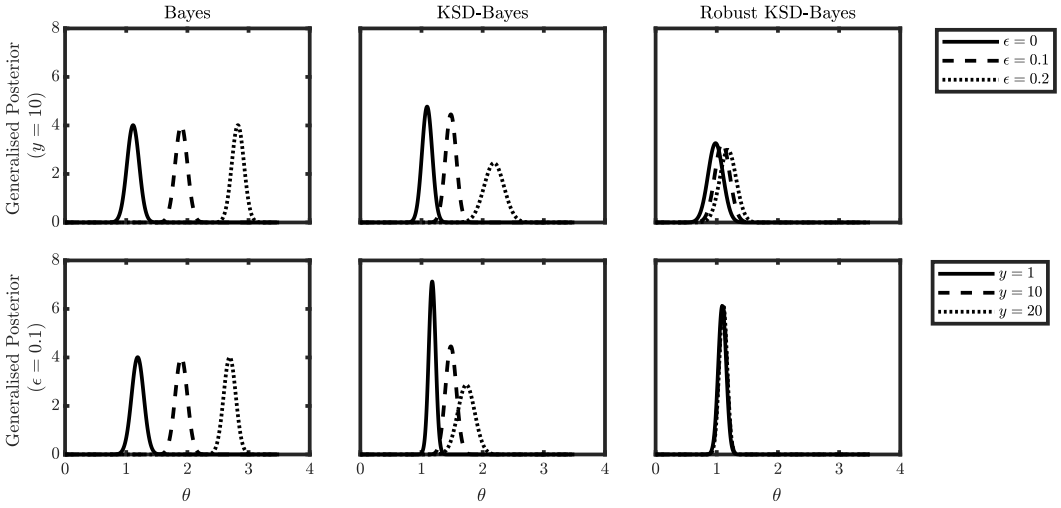


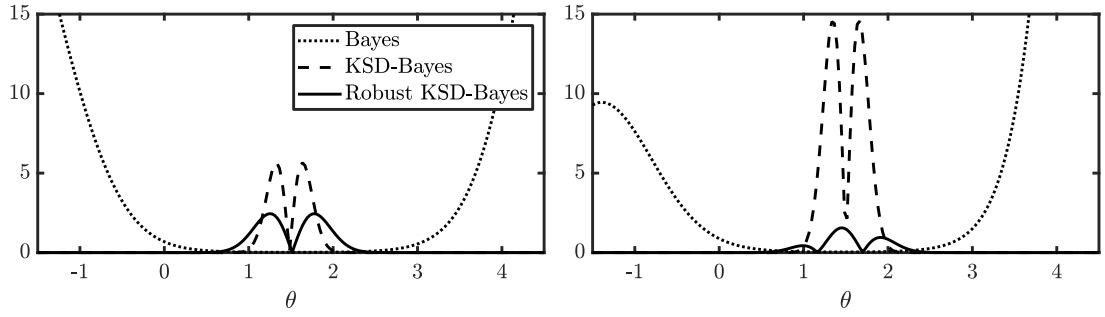
Figure 4.2 Posteriors and generalised posteriors for the normal location model. The true parameter value is $\theta = 1$, while a proportion ϵ of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. In the top row $y = 10$ is fixed and $\epsilon \in \{0, 0.1, 0.2\}$ are considered, while in the bottom row $\epsilon = 0.1$ is fixed and $y \in \{1, 10, 20\}$ are considered.

is provided in Section A.5. Source code to reproduce these experiments can be downloaded from https://github.com/takuomatsubara/KSD-Bayes.

4.4.1. Normal Location Model

For illustrative purposes, we first consider fitting a normal location model $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$ to a dataset $\{x_i\}_{i=1}^n$. Our aim is to illustrate the robustness properties of KSD-Bayes, and we therefore generated the dataset using a contaminated data-generating model where, for each index $i = 1, \ldots, n$ independently, with probability $1 - \epsilon$ the datum x_i was drawn from $\mathbb{P}_\theta$ with "true" parameter $\theta = 1$, otherwise x_i was drawn from $\mathbb{P}_y = \mathcal{N}(y, 1)$, so that y and ϵ control, respectively, the nature and extent of the contamination in the dataset. The task is to make inferences for θ based on a contaminated dataset of size $n = 100$. The prior on θ was $\mathcal{N}(0, 1)$.

The standard Bayesian posterior is depicted in the leftmost panels of Figure 4.2, for varying ϵ (top row) and varying y (bottom row). Straightforward calculation shows that the expected posterior mean is $\frac{n}{n+1}[\theta + \epsilon(y - \theta)]$, which increases linearly as either y or ϵ are increased, with the other fixed. This behaviour is evident in the leftmost panels of Figure 4.2. The generalised posterior from KSD-Bayes is depicted in the central panels of Figure 4.2. This generalised posterior is slightly less sensitive to contamination compared to the standard posterior. Moreover, the variance slightly increases whenever either ϵ or y are increased, as a result of estimating β. In the rightmost panels of Figure 4.2 we display the robust generalised posterior using the weighting function $M(x) = (1+x^2)^{-1/2}$, intended to bound the influence of large values in the dataset. This choice of $M(x)$ vanishes just fast enough as $|x| \to \infty$ to ensure that the bias-robustness conditions of Theorem 4 are satisfied; see Section A.3. The effect is clear from the bottom right panel of Figure 4.2, where even for $y = 20$ (and ϵ fixed to a small value, $\epsilon = 0.1$) the robust generalised



(a) $\theta \mapsto |\,\mathrm{PIF}(y = 2.0, \theta, \mathbb{P}_n)|$ **(b)** $\theta \mapsto |\,\mathrm{PIF}(y = 20, \theta, \mathbb{P}_n)|$

Figure 4.3 Posterior influence function for the normal location model.

posterior remains centred close to the true value $\theta = 1$. While our theoretical results relate to y and do not guarantee robustness when ϵ is increased, the top right panel in Figure 4.2 suggests that the robust generalised posterior is indeed robust in this regime as well. Figure 4.3 displays the posterior influence function (3.7) for this normal location model. This reveals that the standard Bayesian posterior is not bias-robust, since the tails of the posterior are highly sensitive to the contaminant y. In contrast, the tails of the generalised posterior are insensitive to the contaminant. This appears to be the case for both weighting functions, despite only one weighting function satisfying the conditions of Theorem 4.

4.4.2. Precision Parameters in an Intractable Likelihood Model

Our second experiment is a toy model due to Liu et al. (2019); an exponential family model $p_\theta(x) = \exp(\theta \cdot t(x) - a(\theta) + b(x))$ where $\theta \in \mathbb{R}^2$ are parameters to be inferred and $x \in \mathbb{R}^5$. The model specification is completed with

$$t(x) = (\tanh(x_{(4)}), \tanh(x_{(5)})), \quad b(x) = -0.5 \textstyle\sum_{i=1}^{5} x_{(i)}^2 + 0.6 x_{(1)} x_{(2)} + 0.2 \sum_{i=3}^{5} x_{(1)} x_{(i)}.$$

Despite the apparent simplicity of this model, the term $a(\theta)$, which determines the normalisation constant, is analytically intractable and exact simulation from this data-generating model is not straightforward (excluding the case $\theta = 0$). In sharp contrast, the generalised posterior produced by KSD-Bayes is available in closed form for this model. Our aim here is to assess robustness of the generalised posterior, focusing on the setting where y is fixed and ϵ is increased, since this is the regime for which our theoretical results do *not* hold. A dataset of size $n = 500$ was generated from the model $\mathbb{P}_\theta$ with true parameter $\theta = (0, 0)$, so that $\mathbb{P}_\theta$ has the form $\mathcal{N}(0, \Sigma)$ and can be exactly sampled. Each datum x_i was, with probability ϵ, shifted to $x_i + y$ where $y = (10, \ldots, 10)$. The prior on θ was $\mathcal{N}(0, 10^2 I)$.

The left column in Figure 4.4 displays the standard posterior[4], which is seen to be sensitive to contamination in the dataset, in much the same way observed for the normal

[4]To obtain these results, the intractable normalisation constant was approximated using a numerical cubature method. To do this, we recognise that $p_\theta(x) = \mathcal{N}(x; 0, \Sigma) r_\theta(x) / C_\theta$ where $r_\theta(x) =$

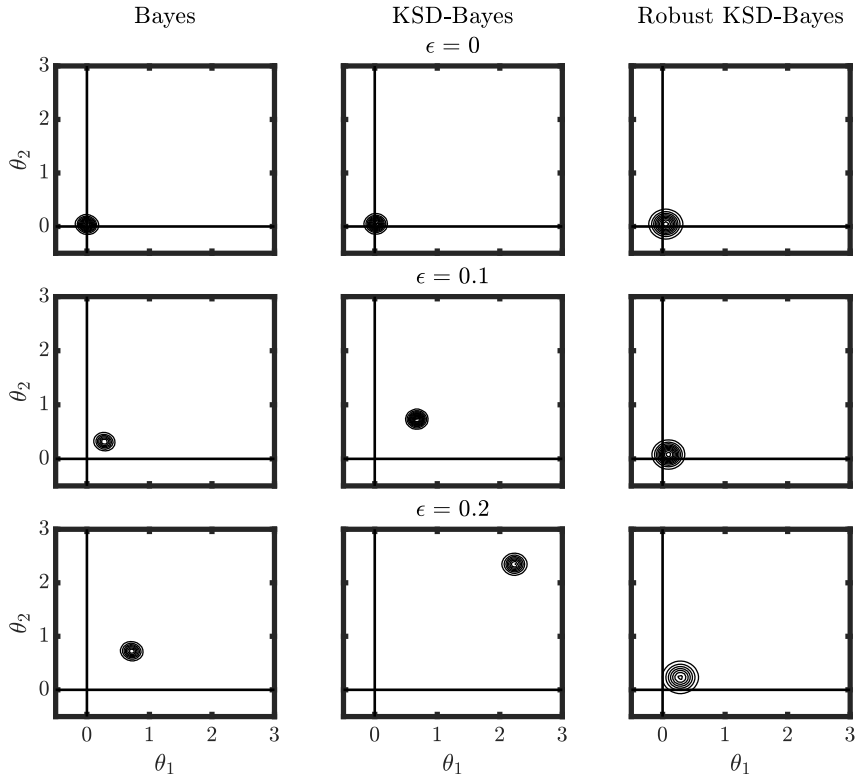


Figure 4.4 Posteriors and generalised posteriors for the Liu et al. (2019) model. The true parameter value is $\theta = 0$, while a proportion ϵ of the data were contaminated by being shifted by an amount $y = (10, 10)$.

location model in Section 4.4.1. The generalised posterior with $M(x) = I_d$ is depicted in the middle column of Figure 4.4, and is seen to be *more* sensitive to contamination compared to the standard Bayesian posterior, in that the mean moves further from 0 as ϵ is increased. Finally, in the right column of Figure 4.4 we display the robust generalised posterior obtained with weighting function

$$M(x) = \text{diag}\left((1 + x_{(1)}^2 + \cdots + x_{(5)}^2)^{-1/2}, (1 + x_{(1)}^2 + x_{(2)}^2)^{-1/2}, \ldots, (1 + x_{(1)}^2 + x_{(5)}^2)^{-1/2}\right),$$

which ensures the criteria for bias-robustness in Theorem 4 are satisfied. From the figure, we observe that the robust generalised posterior remains centred close to the data-generating value $\theta = 0$, even for the largest contamination proportion considered ($\epsilon = 0.2$), with a variance that increases as ϵ is increased. At $\epsilon = 0$, the spread of the robust generalised posterior is almost twice that of the standard posterior, which reflects the trade-off between robustness and efficiency.

4.4.3. Robust Nonparametric Density Estimation

Our third experiment concerns density estimation using the kernel exponential family, and explores the performance of KSD-Bayes when the dataset is multi-modal (c.f. Section 4.2.3).

$\exp(\theta_1 \tanh(x_4) + \theta_2 \tanh(x_5))$. Then $C_\theta = \int r_\theta(x) \mathrm{d}\mathcal{N}(x; 0, \Sigma)$, which was approximated using (polynomial order 10) Gauss-Hermite cubature in 2D.

Let q denote a reference p.d.f. on $\mathbb{R}^d$, and let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel. The *kernel exponential family* model (Canu and Smola, 2006)

$$p_\theta(x) \propto q(x) \exp(\langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}(\kappa)}) \tag{4.12}$$

is parametrised by f, an element of the RKHS $\mathcal{H}(\kappa)$. The implicit normalisation constant of (4.12), if it exists, is typically an intractable function of f. There appears to be no Bayesian or generalised Bayesian treatment of (4.12) in the literature, which may be due to intractability of the likelihood. As the theory in this paper is finite-dimensional, we consider a finite-rank approximation of elements in $\mathcal{H}(\kappa)$ of the form $f(x) = \sum_{i=1}^{p} \theta_{(i)}\phi_{(i)}(x)$, with coefficients $\theta_{(i)} \in \mathbb{R}$ and basis functions $\phi_{(i)} \in \mathcal{H}(\kappa)$, where we will take θ to be $p = 25$ dimensional. Finite rank approximations have previously been considered for frequentist learning of kernel exponential families in Strathmann et al. (2015); Sutherland et al. (2018). In our case, the finite rank approximation ensures that any prior we induce on f via a prior on the coefficients $\theta_{(i)}$ will be supported on $\mathcal{H}(\kappa)$. If one is interested in a well-defined limit as $p \to \infty$ then one will need to ensure a.s. convergence of the sum in this limit. If the ϕ_i are orthonormal in $\mathcal{H}(\kappa)$, and if the $\theta_{(i)}$ are a priori independent, then $\mathbb{E}[\|f\|^2_{\mathcal{H}(\kappa)}] = \sum_{i=1}^{p} \mathbb{E}[\theta^2_{(i)}]$ so a sufficient condition, for example, is $\mathbb{E}[\theta^2_{(i)}] = O(n^{-1-\delta})$ for some $\delta > 0$.

Our interest is in the performance of KSD-Bayes applied to a multi-modal dataset, and to explore these we considered the *galaxy data* of Postman et al. (1986); Roeder (1990), comprising $n = 82$ velocities in km/sec of galaxies from 6 well-separated conic sections of a survey of the *Corona Borealis*. The data were whitened prior to computation, but results are reported with the original scale restored. For the kernel exponential family, we use $q(x) = \mathcal{N}(0, 3^2)$ and the kernel $\kappa(x, y) = \exp(-(x - y)^2/2)$, which ensures that (4.12) is normalisable due to Proposition 2 of Wenliang et al. (2019). For basis functions we use $\phi_{(i+1)}(x) = (x^i/\sqrt{i!}) \exp(-x^2/2)$, $i = 0, \ldots, 24$, which are orthonormal in $\mathcal{H}(\kappa)$ (Steinwart et al., 2006). For our prior, we let $\theta_{(i)} \sim \mathcal{N}(0, 10^2 i^{-1.1})$, which is weakly informative within the constraint of having a well-defined $p \to \infty$ limit. Our contamination model replaces a proportion ϵ of the dataset with values independently drawn from $\mathcal{N}(y, 0.1^2)$, with $y = 5$, shown as black bars in the top row of Figure 4.5.

The generalised posterior with $M(x) = 1$ is displayed in the second row of Figure 4.5, with the bottom row presenting a robust generalised posterior based on the weighting function $M(x) = (1 + x^2)^{-1/2}$, which ensures the conditions of Theorem 4 are satisfied. The results we present are for fixed y and increasing ϵ, since this regime is *not* covered by Theorem 4. The generalised posterior mean is a uni-modal density, which we attribute to the insensitivity of KSD to mixture proportions discussed in Section 4.2.3, but multi-modal densities are evident in sampled output. Our results indicate that the robust weighting function reduces sensitivity to contamination in the dataset (note how the mass in the central mode of the generalised posterior decreases when $\epsilon = 0.2$, when the identity weighting function is used). Whether this insensitivity of KSD to well-separated regions in

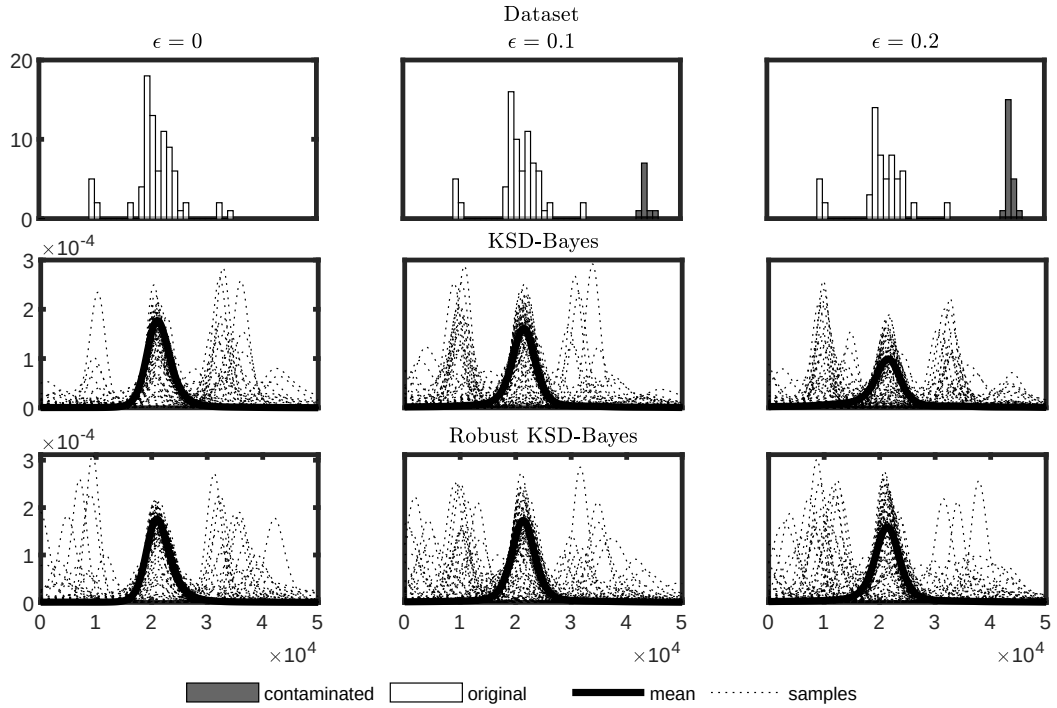


Figure 4.5 Generalised posteriors for the kernel exponential family model. A proportion ϵ of the data (top row) were contaminated. Samples from the generalised posteriors correspond to probability density functions, shown as dotted curves.

the dataset is desirable or not will depend on the application, but in this case it happens to be beneficial.

4.4.4. Network Inference with Exponential Graphical Models

Our final example concerns an exponential graphical model, representing negative conditional relationships among a collection of random variables $W = (W_1, \ldots, W_d)$, described in Yang et al. (2015, Sec. 2.5). The likelihood function is

$$p_{W|\theta}(w|\theta) \propto \exp\Big(-\sum_i \theta_{(i)} w_{(i)} - \sum_{i<j} \theta_{(i,j)} w_{(i)} w_{(j)} \Big), \tag{4.13}$$

where $w \in (0, \infty)^d$ and $\theta_{(i)} > 0, \theta_{(i,j)} \geq 0$. The total number of parameters is $p = d(d+1)/2$. Simulation from this model is challenging and the normalisation constant is an intractable integral, so in what follows a standard Bayesian analysis is not attempted. Our aim is to fit (4.13) to a protein kinase dataset, mimicking an experiment presented by Yu et al. (2016) in the score-matching context. This dataset, originating in Sachs et al. (2005), consists of quantitative measurements of $d = 11$ phosphorylated proteins and phospholipids, simultaneously measured from single cells using a fluorescence-activated cell sorter, so the parameter θ is 66-dimensional. Nine stimulatory or inhibitory interventional conditions were combined to give a total of $7,466$ cells in the dataset. The data were square-root transformed and samples containing values greater than 10 standard deviations from their mean were judged to be *bona fide* outliers and were removed. The remaining dataset

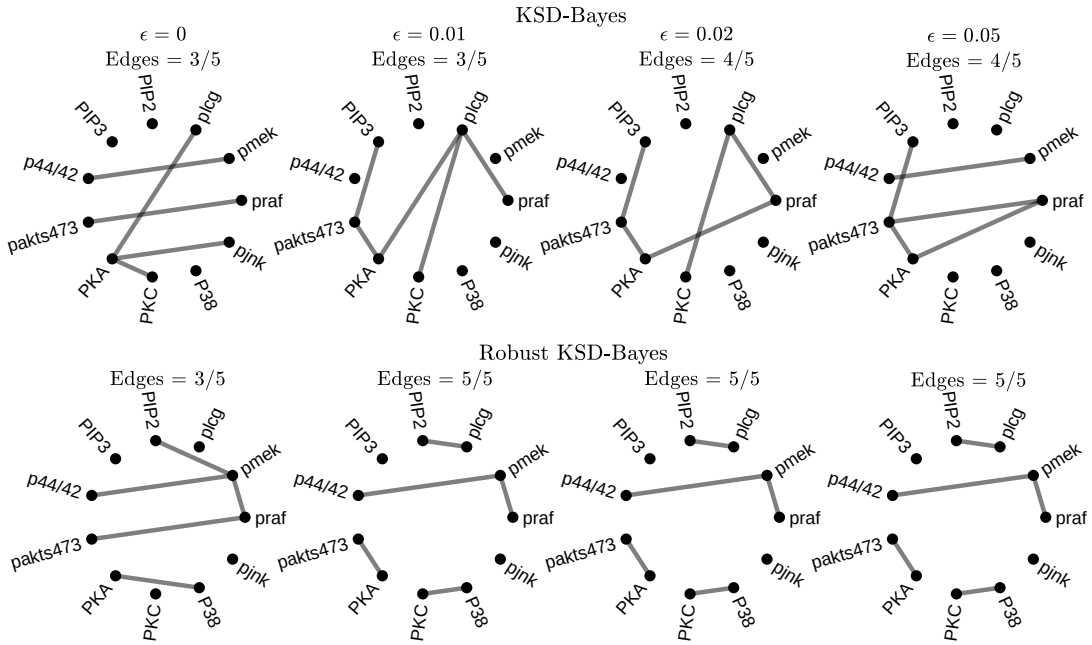


Figure 4.6 Exponential graphical model; estimated protein signalling networks as a function of the proportion ϵ of contamination in the dataset.

of size $n = 7,449$ was normalised to have unit standard deviation. In most cases the measurement reflects the activation state of the kinases, and scientific interest lies in the mechanisms that underpin their interaction[5]. These mechanisms are often summarised as a *protein signalling network*, whose nodes are the d proteins and whose edges correspond to the pairs of proteins that interact. An important statistical challenge is to *estimate* a protein signalling network from such a dataset (Oates, 2013). However, it is known that existing approaches to *network inference* are non-robust, in a general sense, with community challenges regularly highlighting the different conclusions drawn by different estimators applied to an identical dataset (Hill et al., 2016). Our interest is in whether networks estimated using KSD-Bayes are robust.

For our experiment, the variables $w_{(i)}$ were re-parametrised as $x_{(i)} := \log(w_{(i)})$, in order that they are unconstrained and $\mathbb{P}_\theta \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. For the contamination model, a proportion ϵ of the data were replaced with the fixed value $y = (10, \ldots, 10) \in \mathbb{R}^d$. Parameters were *a priori* independent with $\theta_{(i)} \sim \mathcal{N}_{\mathrm{T}}(0, 1)$, $\theta_{(i,j)} \sim \mathcal{N}_{\mathrm{T}}(0, 1)$, where $\mathcal{N}_{\mathrm{T}}$ is the Gaussian distribution truncated to the positive orthant of $\mathbb{R}^p$. This prior is conjugate to the likelihood, as explained in Section 4.2.1, and allows the generalised posterior to be exactly computed. Generalised posteriors were produced both without and with the exponential weighting function $[M(x)]_{(i,i)} = \exp(-x_{(i)})$, the latter aiming to reduce sensitivity to large values in the dataset and coinciding with the identity weighting function at $x = 0$. From these, protein signalling networks were estimated using the s most significant edges, defined as the s largest values of $\bar{\theta}_{(i,j)}/\sigma_{(i,j)}$, where the generalised posterior marginal for $\theta_{(i,j)}$ is

[5]There is no scientific basis to expect only negative conditional dependencies in the dataset; in this sense the model is likely to be misspecified. Our interest is in assessing the robustness properties of KSD-Bayes only, and no scientific conclusions will be drawn using this model.

$\mathcal{N}_{\mathrm{T}}(\bar{\theta}_{(i,j)}, \sigma^2_{(i,j)})$. Results are shown in Figure 4.6; to optimise visualisation we report results for $s = 5$, though for other values of s similar conclusions hold. It is interesting to observe little agreement between the networks returned when the identity weighting function is used, which may reflect the difficulty of the network inference task. Reduced sensitivity to ϵ was observed when the exponential weighting function was used. In Figure 4.6 we report the number of edges that are consistent with the network reported in Sachs et al. (2005, Fig. 3A); the use of the exponential weighting function resulted in more edges being consistent with this benchmark network.

4.5. Concluding Remark

In this chapter, we proposed KSD-Bayes, that is, the SD-Bayes methodology resulting from the use of KSD. Strikingly, KSD-Bayes is computable by any standard MCMC algorithms even if intractable models are used, further admitting fully conjugate inference for exponential family models. Moreover, an appropriate choice of kernel, provided in Section 4.1.2, confers strong robustness against outliers on KSD-Bayes. The simultaneous attainment of the computational efficiency and robustness makes KSD-Bayes a distinctive Bayesian methodology for intractable models. The robustness cannot be achieved if, for example, the score matching objective (2.8) is used in SD-Bayes as an alternative Stein discrepancy that is available in closed form. If the loss D_n is the score matching objective (2.8), the quantity $D_0(y, \theta, \mathbb{P}_n)$ defined in Lemma 3 is given by $D_0(y, \theta, \mathbb{P}_n) = \mathrm{SM}^2(\mathbb{P}_\theta \| \delta_y) + \mathrm{SM}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$, where the term $\mathrm{SM}^2(\mathbb{P}_\theta \| \delta_y)$ can easily violate the condition of Lemma 3 for a large class of models. It is straightforward to verify that with a normal location model. On the other hand, the degree of freedom in the form of KSD introduced by a choice of kernel creates room for the condition in Lemma 3 to be satisfied for any given model. The computational cost of KSD is $\mathcal{O}(n^2)$ because the data-dependent form corresponds to a double summation, in contrast to e.g. the score matching objective whose data-dependent form corresponds to a single summation. However, if the conjugate inference of KSD-Bayes is available, the quadratic cost $\mathcal{O}(n^2)$ occurs only once when computing a mean and covariance of the KSD-Bayes posterior in the Gaussian form. The conjugate inference significantly suppresses the computational cost of KSD-Bayes compared to cases where MCMC algorithms are required.

4.6. Proofs of Chapter 4

This section contains all the deferred proofs of theoretical results in Chapter 4. The proofs of posterior consistency and the Bernstein–von Mises theorem of KSD-Bayes are placed in the main text since they are immediate from Theorems 1 and 2 in Chapter 3 provided the results in Section 4.3.1. The proofs of other results and useful lemmas are contained in this section. Set $D_n(\theta) = \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$ and $D(\theta) = \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})$ throughout this section.

4.6.1. Proof of Proposition 2

The following properties of the Stein operator $\mathcal{S}_{\mathbb{Q}}$ will be useful:

Lemma 9. *Under Assumption 4, we have, for all $x, x' \in \mathcal{X}$ and $h \in \mathcal{U}$,*

(i) $\mathcal{S}_{\mathbb{Q}} K(x, \cdot) \in \mathcal{H}$,

(ii) $\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot)\rangle_{\mathcal{H}}$,

(iii) $|\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x')| \leq \sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x)}\sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x', x')}$.

Proof. First of all, since $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$ for each fixed $x \in \mathcal{X}$ by assumption, from the Riesz representation theorem (Steinwart and Christmann, 2008, Theorem A.5.12) there exists a *representer* $g_x \in \mathcal{H}$ for each fixed $x \in \mathcal{X}$ s.t.

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h, g_x\rangle_{\mathcal{H}}.$$

Second of all, the reproducing property $h(x') = \langle h(\cdot), K(\cdot, x')\rangle_{\mathcal{H}}$ holds for any $h \in \mathcal{H}$, where we recall that the inner product between $h \in \mathcal{H}$ and a matrix-valued function $K(x, \cdot)$ is defined in Section 2.5. By the reproducing property, for all $x, x' \in \mathcal{X}$,

$$g_x(x') = \langle g_x, K(\cdot, x')\rangle_{\mathcal{H}} = \mathcal{S}_{\mathbb{Q}}\left[K(\cdot, x')\right](x) = \mathcal{S}_{\mathbb{Q}} K(x, x'). \tag{4.14}$$

In particular, $\mathcal{S}_{\mathbb{Q}} K(x, \cdot) \in \mathcal{H}$ since $g_x \in \mathcal{H}$, establishing item (i). Based on these two observations, we can rewrite $\mathcal{S}_{\mathbb{Q}}[h](x)$ at each fixed $x \in \mathcal{X}$ as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h, g_x\rangle_{\mathcal{H}} = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot)\rangle_{\mathcal{H}}, \tag{4.15}$$

establishing item (ii). We now apply (4.15) with $h(\cdot) = \mathcal{S}_{\mathbb{Q}} K(x', \cdot)$ to deduce that

$$\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x', x) = \mathcal{S}_{\mathbb{Q}}\left[\mathcal{S}_{\mathbb{Q}} K(x', \cdot)\right](x) = \langle \mathcal{S}_{\mathbb{Q}} K(x', \cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot)\rangle_{\mathcal{H}}. \tag{4.16}$$

Applying the Cauchy-Schwarz inequality,

$$|\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x')| = |\langle \mathcal{S}_{\mathbb{Q}} K(x, \cdot), \mathcal{S}_{\mathbb{Q}} K(x', \cdot)\rangle_{\mathcal{H}}| \leq \|\mathcal{S}_{\mathbb{Q}} K(x, \cdot)\|_{\mathcal{H}} \|\mathcal{S}_{\mathbb{Q}} K(x', \cdot)\|_{\mathcal{H}}.$$

Here for each $x \in \mathcal{X}$ the norm term can computed using (4.16):

$$\|\mathcal{S}_{\mathbb{Q}} K(x, \cdot)\|_{\mathcal{H}} = \sqrt{\langle \mathcal{S}_{\mathbb{Q}} K(x, \cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot)\rangle_{\mathcal{H}}} = \sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x)}$$

Therefore for all $x, x' \in \mathcal{X}$ we have

$$|\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x')| \leq \sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x, x)}\sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x', x')},$$

establishing item (iii). □

We now move on to the main proof of Proposition 2.

Proof. From item (ii) of Lemma 9, for each $x \in \mathcal{X}$, $h \in \mathcal{H}$, we have

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot)\rangle_{\mathcal{H}}.$$

Taking the expectation of both sides,

$$\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] = \mathbb{E}_{X\sim\mathbb{P}}\left[\langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(X, \cdot)\rangle_{\mathcal{H}}\right] = \left\langle h(\cdot), \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X, \cdot)\right]\right\rangle_{\mathcal{H}}. \tag{4.17}$$

Here since the inner product is continuous liner operator, the expectation and inner product can be exchanged if the function $x \mapsto \mathcal{S}_{\mathbb{Q}} K(x, \cdot)$ is *Bochner* $\mathbb{P}$*-integrable* (Steinwart and Christmann, 2008, A.32). This is indeed the case, since from item (ii) of Lemma 9 again, and Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{X\sim\mathbb{P}}\left[\|\mathcal{S}_{\mathbb{Q}} K(X, \cdot)\|_{\mathcal{H}}\right] &= \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{\langle \mathcal{S}_{\mathbb{Q}} K(X, \cdot), \mathcal{S}_{\mathbb{Q}} K(X, \cdot)\rangle_{\mathcal{H}}}\right] \\ &= \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X, X)}\right] \leq \sqrt{\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X, X)\right]} < \infty\end{aligned}$$

where the last term is finite by Assumption 4. A standard argument based on the Cauchy–Schwarz inequality gives

$$\begin{aligned}\sup_{\|h\|_{\mathcal{H}}\leq 1}\left|\left\langle h(\cdot), \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X, \cdot)\right]\right\rangle_{\mathcal{H}}\right| &= \left\|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X, \cdot)\right]\right\|_{\mathcal{H}} \\ &= \sqrt{\left\langle \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X, \cdot)\right], \mathbb{E}_{X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X', \cdot)\right]\right\rangle_{\mathcal{H}}} \\ &= \sqrt{\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\left\langle \mathcal{S}_{\mathbb{Q}} K(X, \cdot), \mathcal{S}_{\mathbb{Q}} K(X', \cdot)\right\rangle_{\mathcal{H}}\right]} \\ &= \sqrt{\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X, X')\right]}\end{aligned} \tag{4.18}$$

where X and X' are independent, and we again appeal to Bochner $\mathbb{P}$-integrability to interchange expectation and inner product. Thus from (4.17) and (4.18) we have

$$\mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) = \left(\sup_{\|h\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right]\right|\right)^2 = \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X, X')\right],$$

as claimed. □

4.6.2. Assumption 4 for the Langevin Stein operator

We demonstrate how to verify the assumption that $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$ for each fixed $x \in \mathcal{X}$ in the case where $\mathcal{S}_{\mathbb{Q}}$ is the Langevin Stein operator (2.5) for $\mathbb{Q} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. Since a linear functional is continuous if and only if it is bounded, we aim to show that, for each fixed $x \in \mathcal{X}$, there exist a constant C_x s.t. $|\mathcal{S}_{\mathbb{Q}}[h](x)| \leq C_x\|h\|_{\mathcal{H}}$

for all $h \in \mathcal{H}$. For each fixed $x \in \mathbb{R}^d$, the Langevin Stein operator $\mathcal{S}_{\mathbb{Q}}$ is given as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \nabla \log q(x) \cdot h(x) + \nabla \cdot h(x).$$

From the reproducing property $h(x) = \langle h, K(x, \cdot)\rangle_{\mathcal{H}}$ for any $h \in \mathcal{H}$, we have

$$\begin{aligned}\mathcal{S}_{\mathbb{Q}}[h](x) &= \nabla \log q(x) \cdot \langle h, K(x,\cdot)\rangle_{\mathcal{H}} + \nabla_x \cdot \langle h, K(x,\cdot)\rangle_{\mathcal{H}} \\ &= \langle h, K(x,\cdot)\nabla \log q(x)\rangle_{\mathcal{H}} + \langle h, \nabla_x \cdot K(x,\cdot)\rangle_{\mathcal{H}}\end{aligned}$$

where the order of inner product and other operators is exchangeable by the continuity of $\langle h, \cdot\rangle_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}$ (Steinwart and Christmann, 2008, Corollary 4.36). Then by the Cauchy–Schwarz inequality,

$$\begin{aligned}|\mathcal{S}_{\mathbb{Q}}[h](x)| &\le \Big(\|K(x,\cdot)\nabla \log q(x)\|_{\mathcal{H}} + \|\nabla_x \cdot K(x,\cdot)\|_{\mathcal{H}}\Big)\|h\|_{\mathcal{H}} \\ &= \left(\sqrt{\nabla \log q(x) \cdot K(x,x)\nabla \log q(x)} + \sqrt{\nabla \cdot (\nabla \cdot K(x,x))}\right)\|h\|_{\mathcal{H}} =: C_x\|h\|_{\mathcal{H}}.\end{aligned}$$

where the first and second gradient of $\nabla \cdot (\nabla \cdot K(x,x))$ are taken each with respect to the first and second argument of K. For the constant C_x to exist, it is sufficient to require that $\nabla \log q(x)$, $K(x,x)$ and $\nabla \cdot (\nabla \cdot K(x,x))$ exist. This is the case when, for example, $\mathbb{Q} \in \mathcal{P}_{\mathcal{S}}(\mathbb{R}^d)$ and $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$, as assumed in Gorham and Mackey (2017).

4.6.3. Proof of Proposition 3

Proof. From (4.3), $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K$ is given by

$$\begin{aligned}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \overset{+C}{=} &\underbrace{\nabla \log p_\theta(x) \cdot K(x,x')\nabla \log p_\theta(x')}_{(*_1)} \\ &+ \underbrace{\nabla \log p_\theta(x) \cdot (\nabla_{x'} \cdot K(x,x'))}_{(*_2)} + \underbrace{\nabla \log p_\theta(x') \cdot (\nabla_x \cdot K(x,x'))}_{(*_3)},\end{aligned}$$

where $\overset{+C}{=}$ indicates equality up to an additive term that is θ-independent. The exponential family model in (4.6) satisfies $\nabla \log p_\theta(x) = \nabla t(x)\eta(\theta) + \nabla b(x)$. For term $(*_1)$, we have

$$\begin{aligned}\sum_{i=1}^n\sum_{j=1}^n (*_1) = &\sum_{i=1}^n\sum_{j=1}^n (\nabla t(x_i)\eta(\theta)) \cdot K(x_i,x_j)\nabla t(x_j)\eta(\theta) + \nabla b(x_i) \cdot K(x_i,x_j)\nabla t(x_j)\eta(\theta) \\ &+ (\nabla t(x_i)\eta(\theta)) \cdot K(x_i,x_j)\nabla b(x_j) + \nabla b(x_i) \cdot K(x_i,x_j)\nabla b(x_j) \\ \overset{+C}{=} &\ \eta(\theta) \cdot \left(\sum_{i=1}^n\sum_{j=1}^n \nabla t(x_i)^\top K(x_i,x_j)\nabla t(x_j)\right)\eta(\theta) \\ &+ \eta(\theta) \cdot \left(2\sum_{i=1}^n\sum_{j=1}^n \nabla t(x_i)^\top K(x_i,x_j)\nabla b(x_j)\right) \qquad (4.19)\end{aligned}$$

where the last equality follows from symmetry of K. For terms $(*_2)$ and $(*_3)$,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}(*_2) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\nabla t(x_i)\eta(\theta))\cdot(\nabla_{x'}\cdot K(x_i,x_j)) + \nabla b(x_i)\cdot(\nabla_{x'}\cdot K(x_i,x_j)) \\ &\overset{+C}{=} \eta(\theta)\cdot\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\nabla t(x_i)^\top(\nabla_{x'}\cdot K(x_i,x_j))\right), \qquad (4.20)\\ \sum_{i=1}^{n}\sum_{j=1}^{n}(*_3) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\nabla t(x_i)\eta(\theta))\cdot(\nabla_{x}\cdot K(x_i,x_j)) + \nabla b(x_j)\cdot(\nabla_{x}\cdot K(x_i,x_j)) \\ &\overset{+C}{=} \eta(\theta)\cdot\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\nabla t(x_j)^\top(\nabla_{x}\cdot K(x_i,x_j))\right). \qquad (4.21)\end{aligned}$$

From Equation (4.2), the KSD-Bayes posterior is

$$\pi_n^D(\theta) \propto \pi(\theta)\exp\left(-\beta n\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(*_1)+(*_2)+(*_3)\right\}\right),$$

so we may collect together terms in Equations (4.19) to (4.21) to obtain the expressions in Proposition 3. □

4.6.4. Proof of Lemma 4 (a.s. Pointwise Convergence)

Proof. Decomposing the double summation of $D_n(\theta)$ into the diagonal term ($i=j$) and non-diagonal term ($i\neq j$),

$$\begin{aligned}D_n(\theta) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}(x_i,x_j) \\ &= \frac{1}{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_i)}_{(*_a)} + \frac{n-1}{n}\underbrace{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_i)}_{(*_b)}.\end{aligned}$$

Fix $\theta\in\Theta$. From the strong law of large number (Durrett, 2010, Theorem 2.5.10),

$$(*_a) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_i) \xrightarrow{a.s.} \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right],$$

provided that $\mathbb{E}_{X\sim\mathbb{P}}\left[|\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)|\right]<\infty$. From the positivity of $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x)$, we have $\mathbb{E}_{X\sim\mathbb{P}}\left[|\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)|\right] = \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]$, which has been assumed to exist. The form of (b) is called an *unbiased statistic* (or *U-statistic* for short) and Hoeffding (1961) proved the strong law of large numbers

$$(*_b) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right],$$

whenever $\mathbb{E}_{X,X'\sim\mathbb{P}}\left[|\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')|\right] < \infty$. From item (iii) of Lemma 9 and Jensen's inequality, we have $\mathbb{E}_{X,X'\sim\mathbb{P}}\left[|\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')|\right] \leq \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]$ where the right hand side is again assumed to exist. Therefore, since $1/n \to 0$ and $(n-1)/n \to 1$,

$$D_n(\theta) = \frac{1}{n}(*_a) + \frac{n-1}{n}(*_b) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right] = D(\theta),$$

where the argument holds for each fixed $\theta \in \Theta$. □

4.6.5. Proof of Lemma 5 (a.s. Uniform Convergence)

Similarly to ∇^2_θ, we let $\nabla^3_\theta := \nabla_\theta \otimes \nabla_\theta \otimes \nabla_\theta$ denote the tensor product $\otimes$ where each component is given by $\partial^3_{h,k,l}$. We first see in the following lemma that Assumption 5 implies the precondition C2 of Assumption 3. The proof of the following lemma uses technical results presented later in Section A.6.

Lemma 10 (Derivatives a.s. Bounded)**.** *Suppose Assumption 5 ($r_{max} = 3$) holds. Then* $\limsup_{n\to\infty}\sup_{\theta\in\Theta}\left\|\nabla^r_\theta D_n(\theta)\right\|_2 < \infty$ *a.s. for* $r = 1, 2, 3$*. If instead Assumption 5 ($r_{max} = 1$) holds, then the result holds for* $r = 1$*.*

Proof. First of all, for finite n we have

$$\nabla^r_\theta D_n(\theta) = \nabla^r_\theta \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \nabla^r_\theta\Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j)\Big).$$

From the triangle inequality and Lemma 17, we further have

$$\sup_{\theta\in\Theta}\left\|\nabla^r_\theta D_n(\theta)\right\|_2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \sup_{\theta\in\Theta}\left\|\nabla^r_\theta\Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j)\Big)\right\|_2 \leq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n M^r(x_i,x_j).$$

It follows from Lemma 19 that $(1/n^2)\sum_{i=1}^n\sum_{j=1}^n M^r(x_i,x_j) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X,X')] < \infty$. Therefore, a.s. $\limsup_{n\to\infty}\sup_{\theta\in\Theta}\left\|\nabla^r_\theta D_n(\theta)\right\|_2 < \infty$. Inspection of the proof reveals that the argument still holds for $r = 1$ if Assumption 5 ($r_{max} = 1$) holds instead. □

Now we move on to the main proof.

Proof. It directly follows from Lemma 1, which holds under the precondition C1 and C2 of Assumption 3 for $r = 1$. The precondition C1 follows from Lemma 4 and C2 for $r = 1$ follows from Lemma 10. □

4.6.6. Proof of Lemma 6 (Strong Consistency)

Proof. It directly follows from Lemma 2, which holds under the preconditions C1, C2, and C3 in Assumption 3 for $r = 1$. The precondition C1 follows from Lemma 4, C2 for $r = 1$ follows from Lemma 10, and C3 is assumed. □

4.6.7. Proof of Lemma 7 (Asymptotic Normality)

We first introduce the two following lemmas that facilitate the main proof. The proof of the following lemma uses technical results presented later in Section A.6.

Lemma 11 (A.S. Convergence of Derivatives)**.** *Suppose Assumption 5 ($r_{max} = 3$) and 6 hold. Then we have $\nabla_\theta^r D_n(\theta_*) \xrightarrow{a.s.} \nabla_\theta^r D(\theta_*)$ for $r = 1, 2, 3$.*

Proof. The argument is analogous to that used to prove Lemma 4, based on the decomposition

$$\nabla_\theta^r D_n(\theta) = \nabla_\theta^r \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x_i, x_j) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \nabla_\theta^r \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x_i, x_j) \Big).$$

Let $F(x, x') := \nabla_\theta^r \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big)$ to see that

$$\nabla_\theta^r D_n(\theta) = \frac{1}{n} \underbrace{\frac{1}{n} \sum_{i=1}^n F(x_i, x_i)}_{(*_1)} + \frac{n-1}{n} \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n F(x_i, x_j)}_{(*_2)}.$$

It follows from the strong law of large number (Durrett, 2010, Theorem 2.5.10) that $(*_1) \xrightarrow{a.s.} \mathbb{E}_{X \sim \mathbb{P}}[F(X, X)]$ provided $E_{X \sim \mathbb{P}}[\|F(X, X)\|_2] < \infty$. Similarly, it follows from the strong law of large number for U-statistics (Hoeffding, 1961) that $(*_2) \xrightarrow{a.s.} \mathbb{E}_{X, X' \sim \mathbb{P}}[F(X, X')]$ provided $E_{X, X' \sim \mathbb{P}}[\|F(X, X')\|_2] < \infty$. Both the required conditions holds by Lemma 18 and the fact that $\|F(x, x')\|_2 \leq \sup_{\theta \in \Theta} \|\nabla_\theta^r (\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x'))\|_2 \leq M^r(x, x')$ from Lemma 17. Thus

$$\nabla_\theta^r D_n(\theta) \xrightarrow{a.s.} \mathbb{E}_{X, X' \sim \mathbb{P}}[F(X, X')] = \mathbb{E}_{X, X' \sim \mathbb{P}}[\nabla_\theta^r \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x_i, x_j) \Big)].$$

Since $\mathbb{E}_{X, X' \sim \mathbb{P}}[\|F(X, X')\|_2] < \infty$, we may apply the dominated convergence theorem to interchange expectation and differentiation:

$$\mathbb{E}_{X, X' \sim \mathbb{P}}[\nabla_\theta^r (\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(X, X'))] = \nabla_\theta^r \mathbb{E}_{X, X' \sim \mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(X, X')] = \nabla_\theta^r D(\theta).$$

Therefore, setting $\theta = \theta_*$, we conclude that $\nabla_\theta^r D_n(\theta_*) \xrightarrow{a.s.} \nabla_\theta^r D(\theta_*)$. □

Lemma 12 (Moment Condition for Asymptotic Normality)**.** *Suppose that Assumption 5 ($r_{max} = 3$) holds. Let $F(x, x') := \nabla_\theta (\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x'))$ for any fixed $\theta \in \Theta$. Then we have $\mathbb{E}_{X, X' \sim \mathbb{P}} \left[\|F(X, X')\|_2^2 \right] < \infty$ and $\mathbb{E}_{X \sim \mathbb{P}} \left[\|F(X, X)\|_2 \right] < \infty$.*

Proof. First of all, it follows from Lemma 17 that for any $x, x' \in \mathcal{X}$,

$$\|F(x, x')\|_2 \leq \sup_{\theta \in \Theta} \|\nabla_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big)\|_2 \leq M^1(x, x').$$

Thus for the first moment we have $\mathbb{E}_{X \sim \mathbb{P}} \left[\|F(X, X)\|_2 \right] \leq \mathbb{E}_{X \sim \mathbb{P}}[M^1(X, X)] < \infty$ from Lemma 18. For the second moment, $\mathbb{E}_{X, X' \sim \mathbb{P}} \left[\|F(X, X')\|_2^2 \right] \leq \mathbb{E}_{X, X' \sim \mathbb{P}} \left[M^1(X, X')^2 \right] =: (*)$.

By definition,

$$(*) = \mathbb{E}_{X,X' \sim \mathbb{P}} \left[\left(m^1(X) m^0(X') + m^0(X) m^1(X') \right)^2 \right] = 4 \mathbb{E}_{X \sim \mathbb{P}} \left[m^1(X)^2 \right] \mathbb{E}_{X \sim \mathbb{P}} \left[m^0(X)^2 \right].$$

Each of these latter expectations is finite by Lemma 18, which completes the proof. □

Now we move on to the main proof.

Proof. It was assumed that, for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, the map $\theta \mapsto \mathcal{S}_{\mathbb{P}_\theta}[h](x)$ is three times continuously differentiable, from which it follows that f_n is three times continuously differentiable as well. Since θ_n minimises f_n for all sufficiently large n, we have $\nabla_\theta D_n(\theta_n) = 0$. Hence a second order Taylor expansion around θ_* yields

$$0 = \nabla_\theta D_n(\theta_n) = \nabla_\theta D_n(\theta_*) + \nabla_\theta^2 D_n(\theta_*)(\theta_n - \theta_*) + (\theta_n - \theta_*) \cdot \nabla_\theta^3 D_n(\theta_n')(\theta_n - \theta_*)$$

where $\theta_n' = \alpha \theta_* + (1 - \alpha)\theta_n$ for some $\alpha \in [0, 1]$. By transposing the terms properly and scaling the both side by $\sqrt{n}$, we have

$$\sqrt{n}(\theta_n - \theta_*) = \Big(\underbrace{\nabla_\theta^2 D_n(\theta_*)}_{(*_1)} + \underbrace{(\theta_n - \theta_*) \cdot \nabla_\theta^3 D_n(\theta_n')}_{(*_2)} \Big)^{-1} \Big(- \underbrace{\sqrt{n} \nabla_\theta D_n(\theta_*)}_{(*_3)} \Big).$$

In the remainder, we show the convergence of $(*_1)$, $(*_2)$ and $(*_3)$, and apply the Slutsky's theorem to see the convergence in distribution of $\sqrt{n}(\theta - \theta_n)$.

Term $(*_1)$**:** First of all, by the triangle inequality,

$$\left\| \nabla_\theta^2 D_n(\theta_n) - \nabla_\theta^2 D(\theta_*) \right\|_2 \leq \underbrace{\left\| \nabla_\theta^2 D_n(\theta_n) - \nabla_\theta^2 D_n(\theta_*) \right\|_2}_{(**_1)} + \underbrace{\left\| \nabla_\theta^2 D_n(\theta_*) - \nabla_\theta^2 D(\theta_*) \right\|_2}_{(**_2)}.$$

By the mean value theorem applied to $(**_1)$ and Lemma 10 (i.e. $\lim_{n \to \infty} \sup_{\theta \in \Theta} \|\nabla_\theta^3 D_n(\theta)\|_2 < \infty$ a.s.), there a.s. exists a constant $0 < C < \infty$ s.t., for all sufficiently large n,

$$(**_1) = \left\| \nabla_\theta^2 D_n(\theta_n) - \nabla_\theta^2 D_n(\theta_*) \right\|_2 \leq \sup_{\theta \in \Theta} \|\nabla_\theta^3 D_n(\theta)\|_2 \|\theta_n - \theta_*\|_2 \leq C \|\theta_n - \theta_*\|_2.$$

Then applying Lemma 6 (i.e. $\|\theta_n - \theta_*\|_2 \overset{a.s.}{\to} 0$), we have $(**_1) \overset{a.s.}{\to} 0$. Further the preceding Lemma 11 implied that $(**_2) \overset{a.s.}{\to} 0$. Therefore, we conclude that $\nabla_\theta^2 D_n(\theta_n) \overset{a.s.}{\to} \nabla_\theta^2 D(\theta_*) = H_*$, where the Hessian H_* is semi positive definite since θ_* is the minimiser of D from Assumption 6.

Term $(*_2)$**:** From the Cauchy–Schwarz inequality and auxiliary result Lemma 10,

$$\begin{aligned} \limsup_{n \to \infty} \left\| (\theta_n - \theta_*) \cdot \nabla_\theta^3 D_n(\theta_n') \right\|_2 &\leq \limsup_{n \to \infty} \sup_{\theta \in \Theta} \|\nabla_\theta^3 D_n(\theta)\|_2 \|\theta_n - \theta_*\|_2 \\ &\leq \underbrace{\limsup_{n \to \infty} \sup_{\theta \in \Theta} \|\nabla_\theta^3 D_n(\theta)\|_2}_{< \infty \text{ a.s.}} \times \limsup_{n \to \infty} \|\theta_n - \theta_*\|_2 \end{aligned}$$

Since Lemma 6 implies that $\|\theta_n - \theta_*\|_2 \xrightarrow{a.s.} 0$, we have $(*_2) \xrightarrow{a.s.} 0$.

Term $(*_3)$**:** Let $F(x, x') := \nabla_\theta(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x, x'))|_{\theta=\theta_*} \in \mathbb{R}^p$ and recall that $S(x, \theta_*) = \mathbb{E}_{X\sim\mathbb{P}}[F(x, X)] \in \mathbb{R}^p$. Then

$$\begin{aligned}\sqrt{n}\nabla_\theta D_n(\theta_*) &= \sqrt{n}\left(\frac{1}{n^2}\sum_{i=1}^n F(x_i, x_i) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j\neq i}^n F(x_i, x_j)\right) \\ &= \frac{1}{\sqrt{n}}\underbrace{\frac{1}{n}\sum_{i=1}^n F(x_i, x_i)}_{(*_a)} + \frac{n-1}{n}\underbrace{\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n F(x_i, x_j)}_{(*_b)}.\end{aligned}$$

First, it follows from the strong law of large number (Durrett, 2010, Theorem 2.5.10) that $(*_a) \xrightarrow{a.s.} \mathbb{E}_{X\sim\mathbb{P}}[F(X, X)]$ whenever $\mathbb{E}_{X\sim\mathbb{P}}[\|F(X, X)\|_2] < \infty$. Second, since $(*_b)$ is a U-statistic multiplied by $\sqrt{n}$, it follows from van der Vaart (1998, Theorem 12.3) that $(*_b) \xrightarrow{p} (1/\sqrt{n})\sum_{i=1}^n S(x_i, \theta_*)$ whenever $\mathbb{E}_{X,X'\sim\mathbb{P}}[\|F(X, X')\|_2^2] < \infty$. (Here $\xrightarrow{p}$ denotes convergence in probability.) Both the required conditions indeed hold from the auxiliary result Lemma 12. Thus we have

$$\sqrt{n}\nabla_\theta D_n(\theta_*) = \frac{1}{\sqrt{n}}(*_a) + \frac{n-1}{n}(*_b) \xrightarrow{p} \frac{1}{\sqrt{n}}\sum_{i=1}^n S(x_i, \theta_*).$$

This convergence in probability implies that $\sqrt{n}\nabla_\theta D_n(\theta_*)$ and $(1/\sqrt{n})\sum_{i=1}^n S(x_i, \theta_*)$ converge in distribution to the same limit. Therefore we may apply the central limit theorem for $(1/\sqrt{n})\sum_{i=1}^n S(x_i, \theta_*)$ to obtain the asymptotic distribution of $\sqrt{n}\nabla_\theta D_n(\theta_*)$. Again from van der Vaart (1998, Theorem 12.3), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n S(x_i, \theta_*) \xrightarrow{d} \mathcal{N}(0, J_*), \qquad J_* = \mathbb{E}_{X\sim\mathbb{P}}\left[S(X, \theta_*)S(X, \theta_*)^\top\right]$$

whenever $\mathbb{E}_{X,X'\sim\mathbb{P}}[\|F(X, X')\|_2^2] < \infty$, which implies the existence of the covariance matrix J_*. Hence $\sqrt{n}\nabla_\theta D_n(\theta_*) \xrightarrow{d} \mathcal{N}(0, J_*)$.

Collecting together these results, we have shown that

$$(*_1) \xrightarrow{a.s.} H_*, \qquad (*_2) \xrightarrow{a.s.} 0, \qquad (*_3) \xrightarrow{d} \mathcal{N}(0, J_*).$$

Since H_* is guaranteed to be at least positive semi-definite, it is in fact strictly positive definite if H_* is non-singular, as we assumed. Finally, Slutsky's theorem allows us to conclude that $\sqrt{n}(\theta - \theta_n) \xrightarrow{d} \mathcal{N}(0, H_*^{-1}J_*H_*^{-1})$ as claimed. □

4.6.8. Proof of Lemma 8

Proof. Since $|a^2 - b^2| = |(a+b)(a-b)| = (a+b)|a-b|$ for all $a, b \in [0, \infty)$, we have an equality

$$\underbrace{\left|\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})\right|}_{=:(*)} = \underbrace{\left(\mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P}_n) + \mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P})\right)}_{=:(*_1)} \underbrace{\left|\mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P}_n) - \mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P})\right|}_{=:(*_2)}.$$

In what follows $\mathbb{E}$ denotes the expectation $\mathbb{E}_{X_1,\ldots,X_n}$ with respect to the dataset $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$ for better presentation. Applying the Cauchy–Schwarz inequality, we have

$$\mathbb{E}[(*)] = \mathbb{E}[(*_1)(*_2)] \leq \sqrt{\mathbb{E}[(*_1)^2]}\sqrt{\mathbb{E}[(*_2)^2]}. \tag{4.22}$$

To conclude the proof, we bound the two expectations on the right hand side.

Bounding $\mathbb{E}[(*_1)^2]$: From the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$,

$$\mathbb{E}[(*_1)^2] \leq 2\mathbb{E}\left[\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) + \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})\right] = 2\Big(\mathbb{E}\left[\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)\right] + \mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P})\Big).$$

The preconditions of Lemma 9 holds due to Standing Assumption 2. Thus, from Lemma 9 part (iii), together with Jensen's inequality, we have the two bounds $\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) \leq (1/n)\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i, x_i)$ and $\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}) \leq \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X, X)]$. Plugging these into the previous inequality, and exploiting independence of x_i and x_j whenever $i \neq j$, we have

$$\begin{aligned}\mathbb{E}[(*_1)^2] &\leq 2\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X_i, X_i)\right] + \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X, X)]\right) \\ &= 2\Big(\mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X, X)] + \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X, X)]\Big) = 4\sigma(\theta),\end{aligned}$$

where existence of $\sigma(\theta)$ for all $\theta \in \Theta$ is ensured by Standing Assumption 2.

Bounding $\mathbb{E}[(*_2)^2]$: Recall that $\mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P}_n) = \sup_{\|h\|_\mathcal{H} \leq 1}\left|\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_\theta}[h](x_i)\right|$ by original definition of the Stein discrepancy, where $\mathcal{H}$ is set to the RKHS. From the fact $|\sup_x |f(x)| - \sup_y |g(y)|| \leq \sup_x |f(x) - g(x)|$ for functions f and g, the term $(*_2)$ is upper bounded by

$$\begin{aligned}(*_2) &= \left|\sup_{\|h\|_\mathcal{H}\leq 1}\left|\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_\theta}[h](x_i)\right| - \sup_{\|h\|_\mathcal{H}\leq 1}\left|\mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}[h](X)]\right|\right| \\ &\leq \sup_{\|h\|_\mathcal{H}\leq 1}\left|\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_\theta}[h](x_i) - \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}[u](X)]\right| = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_{X\sim\mathbb{P}}[f(X)]\right|.\end{aligned}$$

where $\mathcal{F} := \{\mathcal{S}_{\mathbb{P}_\theta}[h] \mid \|h\|_\mathcal{H} \leq 1\}$. We can see from this expression that standard arguments in the context of Rademacher complexity theory can be applied. Noting that $|\cdot|^2$ is a

convex function, Proposition 4.11 in Wainwright (2019) gives that

$$\mathbb{E}\left[(*_2)^2\right] \leq \mathbb{E}\left[\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{X\sim\mathbb{P}}[f(X)]\right|\right)^2\right] \leq \mathbb{E}\mathbb{E}_\epsilon\left[2^2\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right)^2\right]$$

where $\{\epsilon_i\}_{i=1}^n$ are independent random variables taking values in $\{-1,+1\}$ with equiprobability $1/2$ and $\mathbb{E}_\epsilon$ is the expectation over $\{\epsilon_i\}_{i=1}^n$. From the essentially same derivation as Proposition 2, the following equality holds:

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right| &= \sup_{\|h\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathcal{S}_{\mathbb{P}_\theta}[h](x_i)\right| = \sup_{\|h\|_{\mathcal{H}}\leq 1}\left|\left\langle h, \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathcal{S}_{\mathbb{P}_\theta}K(x_i,\cdot)\right\rangle_{\mathcal{H}}\right| \\ &= \left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \mathcal{S}_{\mathbb{P}_\theta}K(x_i,\cdot)\right\|_{\mathcal{H}} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j)}.\end{aligned}$$

Plugging this equality into the upper bound of $\mathbb{E}\left[(*_2)^2\right]$, we have

$$\begin{aligned}\mathbb{E}\left[(*_2)^2\right] &\leq 4\mathbb{E}\mathbb{E}_\epsilon\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X_i,X_j)\right] \\ &= 4\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}[K(X_i,X_i)]\right] = \frac{4}{n}\mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)] = \frac{4\sigma(\theta)}{n}.\end{aligned}$$

Finally, combining both the bounds complete the proof. □

4.6.9. The Form of $D_0(y,\theta,\mathbb{P}_n)$ for KSD

The following lemma clarifies the form of $D_0(y,\theta,\mathbb{P}_n)$ for KSD:

Lemma 13. *For $D(\theta;\mathbb{P}_{n,\epsilon,y}) = \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_{n,\epsilon,y})$, we have*

$$D_0(y,\theta,\mathbb{P}_n) = 2\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,y)\right] - 2\mathbb{E}_{X,X'\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right]. \tag{4.23}$$

Proof. From the definition of the ϵ-contamination model as a mixture model, and using the symmetry of K, we have

$$\begin{aligned}\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_{n,\epsilon,y}) &= \mathbb{E}_{X,X'\sim\mathbb{P}_{n,\epsilon,y}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right] \\ &= (1-\epsilon)^2\mathbb{E}_{X,X'\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right] + 2(1-\epsilon)\epsilon\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,y)\right] \\ &\quad + \epsilon^2\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y).\end{aligned}$$

Direct differentiation then yields

$$\begin{aligned}D_0(y,\theta,\mathbb{P}_n) &= \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_{n,\epsilon,y})\Big|_{\epsilon=0} \\ &= 2\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,y)\right] - 2\mathbb{E}_{X,X'\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\right],\end{aligned}$$

as claimed. □

4.6.10. Proof of Theorem 4

Proof. From Lemma 3 with $\mathcal{X} = \mathbb{R}^d$, it is sufficient to show that

$$\text{(i)} \ \sup_{\theta\in\Theta}\left(\pi(\theta)\sup_{y\in\mathbb{R}^d}|D_0(y,\theta,\mathbb{P}_n)|\right) < \infty \quad \text{and} \quad \text{(ii)} \ \int_\Theta \sup_{y\in\mathbb{R}^d}|D_0(y,\theta,\mathbb{P}_n)|\,\pi(\theta)\mathrm{d}\theta < \infty.$$

To establish (i) and (ii) we exploit the expression for $D_0(y,\theta,\mathbb{P}_n)$ in Lemma 13. This furnishes us with the bound

$$\left|D_0(y,\theta,\mathbb{P}_n)\right| \le 2\mathbb{E}_{X\sim\mathbb{P}_n}\Big[\underbrace{|\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,y)|}_{=:(*_1)}\Big] + 2\underbrace{\mathbb{E}_{X,X'\sim\mathbb{P}_n}\Big[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X')\Big]}_{=:(*_2)}. \tag{4.24}$$

From Lemma 9, $(*_1) \le \sqrt{\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y)}\sqrt{\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)}$ and $(*_2) \le \mathbb{E}_{X\sim\mathbb{P}_n}[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)]$. Plugging these bounds into (4.24) and using Jensen's inequality gives

$$(4.24) \le 2\sqrt{\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y)}\sqrt{\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]} + 2\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]. \tag{4.25}$$

Now, observing that

$$\mathbb{E}_{X\sim\mathbb{P}_n}\Big[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\Big] \le \mathbb{E}_{X\sim\mathbb{P}_n}\Big[\sup_{y\in\mathbb{R}^d}(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y))\Big] = \sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y) \tag{4.26}$$

and taking a supremum over y in (4.25), we obtain the bound

$$\sup_{y\in\mathbb{R}^d}|D_0(y,\theta,\mathbb{P}_n)| \le 4\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y). \tag{4.27}$$

Therefore, from (4.27), it suffices to verify the conditions

$$\text{(I)} \ \sup_{\theta\in\Theta}\left(\pi(\theta)\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y)\right) < \infty \quad \text{and} \quad \text{(II)} \ \int_\Theta \sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y)\pi(\theta)\mathrm{d}\theta < \infty,$$

which imply the original conditions (i) and (ii). To this end, in the remainder we (a) exploit the specific form of $\mathcal{S}_{\mathbb{P}_\theta}$ to derive an explicit upper bound on $\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y)$, then (b) verify the conditions (I) and (II) based on this upper bound.

Part (a): By the reproducing property of K, the definition of the diffusion Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$, and the fact $(a_1+a_2)^2 \le 2(a_1^2+a_2^2)$ for $a_1, a_2 \in \mathbb{R}$, we have the bound

$$\begin{aligned}\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(y,y) = \|\mathcal{S}_{\mathbb{P}_\theta}K(y,\cdot)\|_{\mathcal{H}}^2 &= \|\nabla_y \log p_\theta(y)\cdot K(y,\cdot) + \nabla_y \cdot K(y,\cdot)\|_{\mathcal{H}}^2 \\ &\le 2\|\nabla_y \log p_\theta(y)\cdot K(y,\cdot)\|_{\mathcal{H}}^2 + 2\,\|\nabla_y \cdot K(y,\cdot)\|_{\mathcal{H}}^2.\end{aligned}$$

For the first term, the reproducing property of K gives that

$$\|\nabla_y \log p_\theta(y) \cdot K(y,\cdot)\|_{\mathcal{H}}^2 = \nabla_y \log p_\theta(y) \cdot K(y,y) \nabla_y \log p_\theta(y),$$

while for the second term, and letting $R(x,x') := \nabla_x \cdot (\nabla_{x'} \cdot K(x,x'))$, the reproducing property gives that

$$\|\nabla_y K(y,\cdot)\|_{\mathcal{H}}^2 = \left\langle \nabla_y \cdot K(y,\cdot), \nabla_y \cdot K(y,\cdot) \right\rangle_{\mathcal{H}} = R(y,y).$$

Thus, taking the supremum with respect to $y \in \mathbb{R}^d$ yields the upper bound,

$$\sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(y,y) \leq 2 \sup_{y\in\mathbb{R}^d} \left(\nabla_y \log p_\theta(y) \cdot K(y,y) \nabla_y \log p_\theta(y) \right) + 2 \sup_{y\in\mathbb{R}^d} R(y,y).$$

Since $K \in C_b^{1\times 1}(\mathbb{R}^d \times \mathbb{R}^d)$ by assumption, it follows that $C_{MK} := \sup_{y\in\mathbb{R}^d} R(y,y) < \infty$. Thus, we have arrived at

$$\sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(y,y) \leq 2\gamma(\theta) + 2C_{MK}, \tag{4.28}$$

where $\gamma(\theta)$ was defined in the statement of Theorem 4.

Part (b): Now we are in a position to verify conditions (I) and (II). For condition (I), we use (4.28) to obtain

$$\sup_{\theta\in\Theta} \left(\pi(\theta) \sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(y,y) \right) \leq 2 \sup_{\theta\in\Theta} \pi(\theta)\gamma(\theta) + 2C_{MK} \sup_{\theta\in\Theta} \pi(\theta)$$

which is finite by assumption. Similarly, for condition (II), we use (4.28) to obtain

$$\int_\Theta \sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(y,y) \pi(\theta) \mathrm{d}\theta \leq 2 \int_\Theta \pi(\theta)\gamma(\theta)\mathrm{d}\theta + 2C_{MK} \int_\Theta \pi(\theta)\mathrm{d}\theta,$$

which is also finite by assumption. This completes the proof. □

Chapter 5. Case II: Approach to Discrete Intractable Models

In this chapter, we turn our attention to the SD-Bayes approach to intractable models in discrete domains $\mathcal{X}$. We present *DFD-Bayes*, the first generalised Bayesian inference approach tailored to inference for discrete intractable models. The approach is based on a novel discrete extension of the Fisher divergence, termed *discrete Fisher divergence* (DFD), which is a special case of Stein discrepancy. DFD-Bayes achieves several properties that render it particularly attractive for discrete intractable models. First, independence of DFD-Bayes on user-specified hyperparameters, such as kernel in KSD, is appealing in discrete domains $\mathcal{X}$, where a natural choice of kernel often does not exist for given $\mathcal{X}$ or can be highly impractical due to the computational cost. For example, one natural choice of kernel in finite cardinality domains is the heat kernel, which originates in spectral graph theory (Chung and Graham, 1997), but evaluation of the heat kernel requires a $O(D^3)$ cost for $D = \text{card}(\mathcal{X})$. Second, DFD-Bayes enjoys efficient computation at cost $O(nd)$ linear in the size of the dataset, in contrast to quadratic cost $O(n^2 d)$ of KSD-Bayes. Applications of discrete intractable likelihoods often involve a relatively high dimensional domain $\mathcal{X}$ and requires a relatively large volume of data for reliable inference. DFD-Bayes offers the improved computational capacity to handle a large amount of data.

This chapter is structured as follows: Section 5.1 presents the construction of DFD, verifying that it is indeed a Stein discrepancy. Our formulation is a generalisation of the existing discrete extension of the Fisher divergence, in which $\mathcal{X}$ is no longer restricted to finite cardinality or one dimensional sets. DFD-Bayes is introduced in Section 5.2 with detail discussions on its computational advantages. Section 5.3 derives posterior consistency and the Bernstein–von Mises theorem of DFD-Bayes. We also establish a theoretical connection between DFD-bayes and KSD-Bayes in this section. Section 5.4 contains empirical assessment of DFD-Bayes based on three distinct intractable models in discrete domains. Finally, Section 5.6 contains all deferred proofs of theoretical results in Chapter 5.

Notations Development of DFD entails an operator acting on each coordinate of the input variable. In Chapter 5 only, we denote the input variable by $\boldsymbol{x}$ in bold to distinguish the i-th coordinate of arbitrary input $\boldsymbol{x}$, denoted by x_i, from the i-th point $\boldsymbol{x}_i$ of dataset $\{\boldsymbol{x}_i\}_{i=1}^n$. In discrete domain $\mathcal{X}$, we identify probability distributions on $\mathcal{X}$ with their probability mass functions with respect to the counting measure on $\mathcal{X}$. Thus, probability distributions in Chapter 5 are denoted by lower case letters, e.g. q.

5.1. Discrete Fisher Divergence

The Fisher divergence underpins several frequentist estimators for intractable models, most notably score matching (Hyvärinen, 2005), and has been used in the context of Bayesian model selection (Dawid and Musio, 2015) for example. Classically, it is defined for sufficiently regular densities p and q on continuous domains such as $\mathbb{R}^d$ by

$$\mathrm{FD}^2(p\|q) = \mathbb{E}_{X\sim q}[\|\nabla \log p(X) - \nabla \log q(X)\|^2] \tag{5.1}$$

where ∇ denotes the gradient operator in $\mathbb{R}^d$. Its main advantage is that it can be computed without knowledge of the normalising constant[1] of p and, furthermore, expectations with respect to p are not required. In Chapter 4, we considered the use of KSD in SD-Bayes because it achieves the robustness, which the Fisher divergence cannot, and the computational cost can be significantly suppressed when the conjugate inference is available, as discussed in Section 4.5. However, the use of KSD becomes less appealing in the discrete case because the conjugate inference is no longer available and a natural choice of kernel is often impractical to compute, as discussed later. The Fisher divergence was extended to discrete domains in Lyu (2009); Xu et al. (2022). The existing work focuses on domains $\mathcal{X}$ of finite cardinality or one-dimensional models. A technical contribution in this section, which may be of independent interest, is to present an extension of Fisher divergence to discrete domains that can be a countably infinite set in multiple dimensions.

5.1.1. Discrete Domain and Difference Operators

First, we introduce a discrete domain $\mathcal{X}$ to be considered and a discrete analogue of the gradient operator in $\mathbb{R}^d$ used to develop DFD.

Standing Assumption 1. *Let $\mathcal{X} = S_1 \times \cdots \times S_d$, where for each $i = 1, \ldots, d$ there is an order isomorphism $S_i \cong I_i \subseteq \mathbb{Z}$, and $d \in \mathbb{N}$.*

Simply put, S_i is an ordered set of any states that we can assign indices of counting numbers. This setting is general enough to include diverse data types, such as multivariate count data, or network data with a fixed vertex set. For any set $S \cong I \subseteq \mathbb{Z}$, precisely one of the following must hold: (i) no smallest or largest elements of S exist; (ii) both a smallest element, $s_{\min}$, and a largest element, $s_{\max}$, exist; (iii) only $s_{\min}$ exists; (iv) only $s_{\max}$ exists. Without loss of generality, we will identify the case (iv) with (iii) by reversing the ordering of S. In addition, it will be useful to extend the domains S_i to include an additional state (not part of the ordering), denoted $\star$, and to this end we let $S_i^\star = S_i \cup \{\star\}$ and $\mathcal{X}^\star = S_1^\star \times \cdots \times S_d^\star$. A function $h : \mathcal{X} \to \mathbb{R}$ extends to a function $h : \mathcal{X}^\star \to \mathbb{R}$ by setting $h(\boldsymbol{x}) = 0$ whenever any of the coordinates of $\boldsymbol{x}$ are equal to $\star$. We define increment and decrement rules of the coordinates of $\boldsymbol{x}$.

[1]The Fisher divergence depends only on $\nabla \log p$, equal to the ratio $(\nabla p)/p$, meaning it is sufficient to know p up to a normalising constant.

Definition 7. *Let $S \cong I \subseteq \mathbb{Z}$. For consecutive elements $r < s < t$ in S we let $s^- := r$ and $s^+ := t$. If both $s_{\min}$ and $s_{\max}$ exist, we let $s^-_{\min} := s_{\max}$ and $s^+_{\max} := s_{\min}$ or, if only $s_{\min}$ exists, we let $s^-_{\min} := \star$ and $\star^+ = s_{\min}$. For $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$, define $\boldsymbol{x}^{i+} := (x_1, \ldots, x_i^+, \ldots, x_d)$ and $\boldsymbol{x}^{i-} := (x_1, \ldots, x_i^-, \ldots, x_d)$.*

Simply put, this ensures that each element s has both a preceding and proceeding element, so that increments and decrements are well-defined. The above structure can be exploited to define an operator for $\mathcal{X}$ that is analogous to the gradient operators for $\mathbb{R}^d$:

Definition 8. *For $h : \mathcal{X} \to \mathbb{R}$, define the backward difference operator by*

$$\nabla^- h(\boldsymbol{x}) := \left[h(\boldsymbol{x}) - h(\boldsymbol{x}^{1-}),\ \cdots,\ h(\boldsymbol{x}) - h(\boldsymbol{x}^{d-})\right]^\top \in \mathbb{R}^d.$$

This difference operator plays a central role in DFD, which we formally define next.

5.1.2. Construction

We construct DFD, a divergence applicable to probability measures in discrete domains $\mathcal{X}$, based on the definitions in Section 5.1.1. The divergence satisfies the requirements of a proper local scoring rule and thus complements existing scoring rule methodology developed in the finite domain context in Dawid et al. (2012). Recall that $L^p(q, \mathbb{R}^d)$ denotes the Lebesgue space of measurable functions $f : \mathcal{X} \to \mathbb{R}^d$ such that $\sum_{i=1}^d \mathbb{E}_{X\sim q}[|f_i(X)|^p] < \infty$, in which two elements $f, g \in L^p(q, \mathbb{R}^d)$ are identified if they are q-almost everywhere equal. Values of $f \in L^p(q, \mathbb{R}^d)$ in a measure zero domain of q i.e. $\{\boldsymbol{x} \in \mathcal{X} \mid q(\boldsymbol{x}) = 0\}$ are arbitrary and not involved in the Lebesgue integral with respect to q (Rudin, 1987, Remark 1.37, p.29). In what follows, it is sufficient for functions $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$ to be well-defined in the support of q.

Definition 9 (Discrete Fisher Divergence)**.** *Let p and q be probability distributions on $\mathcal{X}$, such that $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. The* discrete Fisher divergence *is defined as*

$$\mathrm{DFD}^2(p\|q) := \mathbb{E}_{X\sim q}\left[\left\|\frac{\nabla^- p(X)}{p(X)} - \frac{\nabla^- q(X)}{q(X)}\right\|^2\right]. \tag{5.2}$$

The choice of a Euclidean norm in (5.2) is not critical and other norms could be employed, but for expository purposes the standard Euclidean norm will be used. Proposition 6 justifies the name 'divergence' and offers an alternative computable formula for (5.2).

Proposition 6. *The discrete Fisher divergence satisfies $\mathrm{DFD}^2(p\|q) \geq 0$ for any p, q, with equality if and only if $p = q$. Furthermore, if $p(\boldsymbol{x}^{j+}) > 0$ for all $\boldsymbol{x}$ and $j = 1, \ldots, d$ in the support of q, it admits the following alternative formula*

$$\mathrm{DFD}^2(p\|q) = \mathbb{E}_{X\sim q}\left[\sum_{j=1}^d \left(\frac{p(X^{j-})}{p(X)}\right)^2 - 2\left(\frac{p(X)}{p(X^{j+})}\right)\right] + C(q), \tag{5.3}$$

where the term $C(q) := \mathbb{E}_{X\sim q}[\sum_{j=1}^{d} 1 + (1 - q(X^{j-})/q(X))^2]$ *is p-independent.*

The proof is provided in Section 5.6.1. Note that $\mathrm{DFD}^2(p\|q)$ can be computed without the normalising constant of p by virtue of $\nabla^- p(\boldsymbol{x})/p(\boldsymbol{x})$, analogously to the continuous Fisher divergence $\mathrm{FD}(p\|q)$ in $\mathbb{R}^d$. All models p_θ used in this thesis are positive on $\mathcal{X}$, for which the assumption $p(\boldsymbol{x}^+) > 0$ in Proposition 1 is automatically satisfied. From Proposition 6, DFD between a model p_θ and an empirical distribution $p_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$ corresponding to data $\{\boldsymbol{x}_i\}_{i=1}^n$, is computed as

$$\mathrm{DFD}^2(p_\theta\|p_n) \stackrel{\theta}{=} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{p_\theta(\boldsymbol{x}_i^{j-})}{p_\theta(\boldsymbol{x}_i)}\right)^2 - 2\left(\frac{p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i^{j+})}\right) \tag{5.4}$$

where $\stackrel{\theta}{=}$ indicates equality up to an additive, θ-independent constant. In contrast to the continuous Fisher divergence, the θ-independent constant $C(p_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} 1 + (1 - p_n(\boldsymbol{x}_i^{j-})/p_n(\boldsymbol{x}_i))^2$ is well-defined for an empirical density p_n in DFD.

We next establish that DFD is in fact a Stein discrepancy. Define the *forward divergence* operator $\nabla^+\cdot$ for a $\mathbb{R}^d$-valued function $H : \mathcal{X} \to \mathbb{R}^d$ by $\nabla^+ \cdot H(\boldsymbol{x}) = \sum_{j=1}^{d} H_j(\boldsymbol{x}^{j+}) - H_j(\boldsymbol{x})$, where $H_j(\boldsymbol{x})$ denotes the j-th output coordinate of $H(\boldsymbol{x})$. Built on the forward divergence operator, we use the following Stein operator

$$\mathcal{S}_p[h](\boldsymbol{x}) = \frac{\nabla^- p(\boldsymbol{x})}{p(\boldsymbol{x})} \cdot H(\boldsymbol{x}) + \nabla^+ \cdot H(\boldsymbol{x}) \tag{5.5}$$

for a function $H : \mathcal{X} \to \mathbb{R}^d$, and set a Stein set equal to the unit ball of $L^2(q, \mathbb{R}^d)$.

Proposition 7. *Let* p *and* q *be probability distributions on* $\mathcal{X}$*, such that* $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$*. Consider a Stein discrepancy whose Stein operator is* (5.5) *and Stein set is* $\mathcal{U} = \{H : \mathcal{X} \to \mathbb{R}^d \mid \sum_{i=1}^{d} \mathbb{E}_{X\sim q}[H_i(X)^2] \leq 1\}$*. Then* $\mathrm{SD}^2(p\|q) = \mathrm{DFD}^2(p\|q)$*.*

The proof is contained in Section 5.6.2. A similar result immediately holds for the continuous Fisher divergence by replacing $\mathcal{X}$ with $\mathbb{R}^d$, ∇^- with ∇, and the Stein discrepancy (5.5) with the Langevin Stein discrepancy (2.5). This observation will allow us to conclude, later in Section 5.3.2, that DFD is a topologically stronger divergence than essentially any KSD. While a natural and computationally appealing choice of kernel K for KSD is not clear in discrete case, it does not prevent KSD from being applicable in discrete case. KSD is well-defined as long as a choice of kernel K and Stein operator $\mathcal{S}_p$ are specified because the abstract construction in Section 4.1 admits any domain $\mathcal{X}$ and Stein operator $\mathcal{S}_p$. For example, KSD under a specific choice of Stein operator $\mathcal{S}_p$ was considered in the discrete context in Yang et al. (2018). Using the Stein operator (5.5) and any kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$, KSD for a model p_θ and data p_n in discrete case is given by (4.2) whose expanded form is (4.3) replacing $\nabla \log p_\theta$ with $(\nabla^- p_\theta)/p_\theta$.

5.2. DFD-Bayes Methodology

We are now in a position to present DFD-Bayes. We select DFD as the Stein discrepancy in the SD-Bayes framework, where the resulting posterior is called the DFD-Bayes posterior.

Definition 10 (DFD-Bayes)**.** *Given a prior distribution π on Θ, a statistical model $p_\theta : \mathcal{X} \to (0, \infty)$ parametrised by $\theta \in \Theta$, and data $\{\boldsymbol{x}_i\}_{i=1}^n$, the* DFD-Bayes posterior *is*

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta n \, \mathrm{DFD}^2(p_\theta \| p_n)\right), \tag{5.6}$$

where $\beta \in (0, \infty)$ is a constant to be specified.

This is clearly a special case of SD-Bayes given $\mathrm{SD}^\gamma(p_\theta \| p_n) = \mathrm{DFD}^2(p_\theta \| p_n)$. The θ-independent constant $C(p_n)$ of $\mathrm{DFD}^2(p_\theta \| p_n)$ will be cancelled out by normalisation of the DFD-Bayes posterior. It is thus sufficient to use (5.4) in place of $\mathrm{DFD}^2(p_\theta \| p_n)$ for computation. The role of n in (5.6) is to ensure correct scaling of the generalised posterior as $n \to \infty$ limit, while the appropriate choice of β is crucial in calibrating the coverage of the generalised posterior at finite n. Section B.1 contains a detailed worked example of the DFD-Bayes posterior and a comparison with other posteriors using simple tractable models. In the same manner as KSD-Bayes, DFD-Bayes achieves tractable computation of the generalised posterior even for intractable models.

We next discuss computational appeals of DFD-Bayes. KSD-Bayes is applicable in discrete case using KSD with the Stein operator (5.5) plugged-in, but DFD-Bayes has several appealing advantages over KSD-Bayes in case of discrete intractable models.

5.2.1. Non-Conjugate Inference and Computation

The DFD-Bayes posterior is directly amenable to standard Markov chain Monte Carlo, in contrast to standard Bayes posteriors in the presence of intractable likelihood. This is because DFD in (5.6) does not depend on the intractable constant, with the cost of evaluating (5.6) as low as $O(d)$. Excluding the θ-independent constant cancelled out by normalisation of the DFD-Bayes posterior, the original formula (5.6) is rearranged as:

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta \sum_{i=1}^n \sum_{j=1}^d \left(\frac{p_\theta(\boldsymbol{x}_i^{j-})}{p_\theta(\boldsymbol{x}_i)}\right)^2 - 2\left(\frac{p_\theta(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i^{j+})}\right)\right), \tag{5.7}$$

where there is no interference by the intractable normalising constant of the model.

Remark 2. *The computational cost associated with evaluation of* (5.7) *is $O(nd)$, which improves on the $O(n^2 d)$ cost of KSD. Furthermore, if $\mathcal{X}$ is a finite set and count data are provided, indicating the number of times each of the elements of $\mathcal{X}$ occurred, then the complexity of* (5.7) *reduces to $O(d)$, independent of the size of the dataset. This is because* $\mathrm{DFD}^2(p_\theta \| p_n)$ *can be obtained by computing the intermediate quantity $(p_\theta(\boldsymbol{x}^{j-})/p_\theta(\boldsymbol{x}))^2 - 2p_\theta(\boldsymbol{x})/p_\theta(\boldsymbol{x}^{j+})$ for each state $\boldsymbol{x}$ of $\mathcal{X}$ once and taking the summation of the memorised quantities weighted by the empirical frequency of $\boldsymbol{x}$.*

The improved computational cost is appealing to discrete intractable models whose applications often involve data spaces of relatively large dimension, in which a large volume of data may be used to improve accuracy of inference. The linear computational cost $O(nd)$ facilitates inference in such situations that a large amount of data ought to be handled. Furthermore, gradient-based efficient samplers, such as the Langevin Monte Carlo algorithm, is effortlessly available up to differentiability of a prior. Decomposing a model into the normalising constant and non-normalised function as $p_\theta(\boldsymbol{x}) = q_\theta(\boldsymbol{x})/Z(\theta)$, the gradient of DFD is given as

$$\nabla_\theta \mathrm{DFD}^2(p_\theta \| p_n) = \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{q_\theta(\boldsymbol{x}_i^{j-})}{q_\theta(\boldsymbol{x}_i)}\right)\left(\frac{\nabla_\theta q_\theta(\boldsymbol{x}_i^{j-})q_\theta(\boldsymbol{x}_i) - q_\theta(\boldsymbol{x}_i^{j-})\nabla_\theta q_\theta(\boldsymbol{x}_i)}{q_\theta(\boldsymbol{x}_i)^2}\right) - \left(\frac{\nabla_\theta q_\theta(\boldsymbol{x}_i)q_\theta(\boldsymbol{x}_i^{j+}) - q_\theta(\boldsymbol{x}_i)\nabla_\theta q_\theta(\boldsymbol{x}_i^{j+})}{q_\theta(\boldsymbol{x}_i^{j+})^2}\right).$$

Under the assumption of Proposition 6, it is clear from the expression above that the gradient of DFD is available whenever $\nabla_\theta q_\theta(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \mathcal{X}$, that is, $q_\theta(\boldsymbol{x})$ is differentiable with respect to θ at each $\boldsymbol{x} \in \mathcal{X}$. In practice, the expression above is amenable to the use of automatic differentiation tools (Baydin et al., 2018).

It also makes DFD-Bayes attractive to discrete models that no choice of kernel is involved. In discrete domains, there are often no natural choices of kernel for KSD well-motivated by theoretical diagnosis, or natural choices, such as the heat kernel, are often impractical due to the high computational cost. This is in intriguing contrast to continuous case $\mathcal{X} = \mathbb{R}^d$ where multiple studies on KSD motivate the use of the IMQ kernel whose computational cost is no different from other common kernels in $\mathbb{R}^d$. Moreover, no dependency of kernel renders DFD-Bayes invariant to data transformation that preserves the order of $\mathcal{X}$. For example, defining alternative input states for implementational convenience within the same ordering—e.g., transforming $\{-1, 1\}$ to $\{0, 1\}$ by some bijective map—does not affect inference outcome of DFD-Bayes.

Remark 3. *DFD-Bayes is invariant to order-preserving transformations of the dataset. This is in contrast to KSD-Bayes, which is not invariant to how the data are represented.*

5.2.2. Limitations

There are three potential limitations of DFD-Bayes that will be discussed now, the first of which is specific to discrete case and the rest of which are similar to ones for KSD-Bayes in continuous case. First, as opposed to continuous case $\mathcal{X} = \mathbb{R}^d$ in Chapter 4, DFD-Bayes (and even KSD-Bayes in discrete case) does no longer satisfy conjugacy for exponential family models. This is because the Stein operator (5.5) used in discrete case relies on the difference operator rather than the log-derivative in continuous case. The lack of conjugate inference emphasises the appeal of the reduced computational cost of DFD-Bayes.

Second, DFD-Bayes was not derived as an approximation to standard Bayesian inference, and thus the semantics associated with the generalised posterior should not be confused

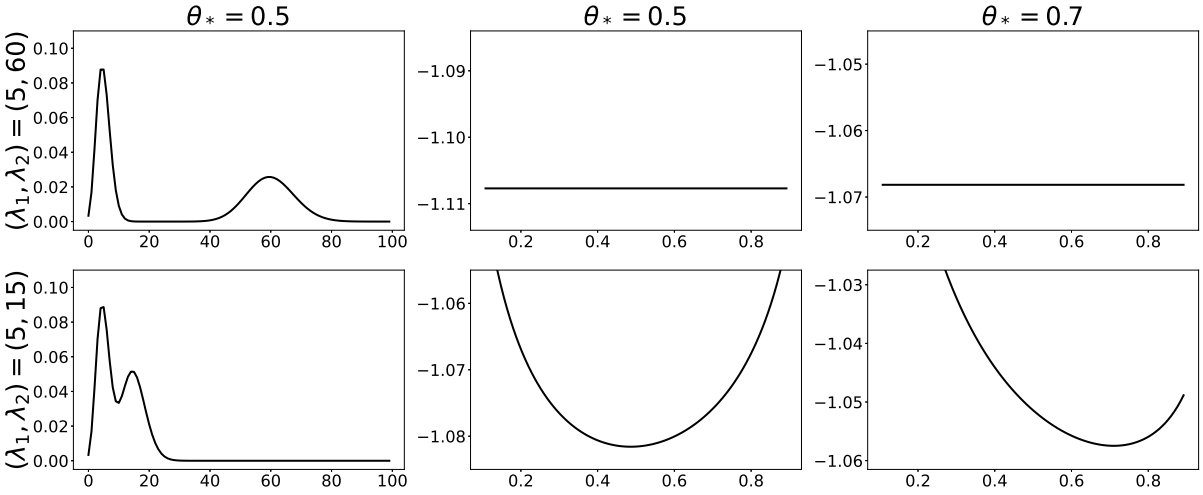


Figure 5.1 The form of the Poisson mixture model p_{θ_*} when $\theta_* = 0.5$ (left), DFD computed for data generated from the model p_{θ_*} with $\theta_* = 0.5$ (middle), and DFD computed for data generated from the model p_{θ_*} with $\theta_* = 0.7$ (right), for two cases where $\lambda_1 = 5, \lambda_2 = 60$ (top) and $\lambda_1 = 5, \lambda_2 = 15$ (bottom).

with the semantics of standard Bayesian inference; see Bissiri et al. (2016); Knoblauch et al. (2022) for a detailed discussion of this point. In particular, we need to calibrate DFD-Bayes through the selection of β, which is not a feature of standard Bayesian inference under well-specified models. Although we expect our bootstrap approach to outperform existing alternative approaches for small sample size n, it is possible that in those cases the bootstrap criterion for selecting β in Section 3.4 will fail, and in these circumstances the generalised posterior will fail to be calibrated.

Third, the generalised posterior may suffer from well-known drawbacks of score-based methods, including insensitivity to mixing proportions (Wenliang and Kanagawa, 2021). Indeed, for a two-component mixture model $p_\theta(\boldsymbol{x}) = (1-\theta)p_1(\boldsymbol{x}) + \theta p_2(\boldsymbol{x})$, we can compute the ratios

$$\rho_i := \left[\frac{\nabla^- p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x})}\right]_i = 1 - \frac{(1-\theta)p_1(\boldsymbol{x}^{i-}) + \theta p_2(\boldsymbol{x}^{i-})}{(1-\theta)p_1(\boldsymbol{x}) + \theta p_2(\boldsymbol{x})}$$

on which DFD is based. Suppose, informally, that the high probability regions R_1 of p_1 and R_2 of p_2 are separated, meaning $p_2 \approx 0$ on R_1 and $p_1 \approx 0$ on R_2. Then these ratios are approximately independent of θ on $R_1 \cup R_2$, since $\rho_i \approx 1 - p_1(\boldsymbol{x}^{i-})/p_1(\boldsymbol{x})$ for $\boldsymbol{x} \in R_1$ and $\rho_i \approx 1 - p_2(\boldsymbol{x}^{i-})/p_2(\boldsymbol{x})$ for $\boldsymbol{x} \in R_2$. It follows that $\mathrm{DFD}(p_\theta \| p_n)$ is approximately independent of θ whenever the data $\{\boldsymbol{x}\}_{i=1}^n \subseteq R_1 \cup R_2$. We illustrate this limitation for DFD using a mixture model of two Poisson distributions $p_\theta(x) = (1-\theta) \times q_{\lambda_1}(x) + \theta \times q_{\lambda_2}(x)$, where q_{λ_1} and q_{λ_2} are the Poisson distributions with rate parameters $\lambda_1 > 0$ and $\lambda_2 > 0$. Figure B.5 shows the surface geometry of DFD between the mixture model p_θ and data generated from the mixture model p_{θ_*} with the true mixture proportion θ_*, for two cases when the supports of the two Poisson distributions are highly isolated and when they

are not isolated. The correct mixture proportion θ_* was identified only in the latter case, while DFD was reduced to a constant in the former case. Thus, although DFD-Bayes may be applied to mixture models, supported by the theoretical guarantees of Theorem 5, the inferences for mixing proportions so-obtained can be data-inefficient. See Zhang et al. (2022) for a potential approach to remedy this general limitation of score-based methods.

5.3. Theoretical Assessment

The asymptotic behaviour of the standard Bayesian posterior is well-understood, with sufficient conditions for posterior consistency and asymptotic normality providing frequentist justification for Bayesian inference in the large data limit. As previous Chapter 4 established analogous conditions for KSD-Bayes, our attention now turns to establishing those conditions for DFD-Bayes. Analysis of DFD-Bayes is relatively more straightforward than KSD-Bayes because DFD falls into a case of additive loss whose convergence results are well-established in classical studies. Section 5.3.2 discusses a theoretical connection of DFD-Bayes and KSD-Bayes, illuminating that DFD dominates KSDs of essentially any kernels through a lens of the Stein discrepancy. Without loss of generality, we will give the proof for $\beta = 1$ for notational convenience. To extend the proof to arbitrary $\beta > 0$, simply replace $\mathrm{DFD}(p_\theta \| p_n)$ in all arguments by $\beta\,\mathrm{DFD}(p_\theta \| p_n)$; all of them hold immediately since β is a constant. The basic setting for which we derive our theory is the following:

Standing Assumption 2. *The data $\{\boldsymbol{x}_i\}_{i=1}^n$ consist of independent samples from a probability distribution p on $\mathcal{X}$. The distribution p and the statistical model p_θ for these data satisfy $(\nabla^- p)/p, (\nabla^- p_\theta)/p_\theta \in L^2(p, \mathbb{R}^d)$, for all $\theta \in \Theta$.*

The setting of independent data is broad enough to contain important examples of discrete intractable likelihood, including the models studied in Section 5.4. The other assumption simply ensures that $\mathrm{DFD}^2(p_\theta \| p_n)$ is well-defined for each $\theta \in \Theta$, due to Proposition 6.

5.3.1. Posterior Consistency and Bernstein–von Mises Theorem

We begin with posterior consistency of DFD-Bayes, which implies that the generalised posterior concentrates around the population minimiser θ_* of $\mathrm{DFD}^2(p_\theta \| p)$ with probability 1 as $n \to \infty$. Recall that Theorem 1 established posterior consistency of abstract generalised posteriors under Assumption 1 and Assumption 2. Because DFD is additive loss, Assumption 2 can be immediately verified by applying Proposition 1 for DFD. This leads to intended posterior consistency as follows:

Proposition 8. *Let $\sigma(\theta) := (\mathbb{V}_{X\sim\mathbb{P}}[\sum_{j=1}^d (p(X^{j-})/p(X))^2 - 2p(X)/p(X^{j+})])^{\frac{1}{2}}$. Suppose Assumption 1 for $D(\theta) = \mathrm{DFD}^2(p_\theta \| p)$, and $\sup_{\theta\in\Theta} \sigma(\theta) < \infty$. Then, for all $\delta \in (0, 1]$,*

$$\mathbb{P}\left(\int_\Theta \left|\mathrm{DFD}^2(p_\theta \| p) - \mathrm{DFD}^2(p_{\theta_*} \| p)\right| \pi_n^D(\theta)\mathrm{d}\theta < \frac{\alpha_1 + \alpha_2 + 8\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}\right) \geq 1 - \delta$$

where the probability is with respect to realisations of the dataset $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

Proof. Posterior consistency of DFD-Bayes follows from Theorem 1 that holds under Assumption 1 and 2. Assumption 1 is assumed. Assumption 2 is implied by Proposition 1 which holds under the assumption that $\sigma(\theta) < \infty$ at each $\theta \in \Theta$. This assumption trivially holds under the condition $\sup_{\theta\in\Theta} \sigma(\theta) < \infty$. □

We move on to a BvM result of DFD-Bayes, which implies the DFD-Bayes posterior is asymptotically normal around the population minimiser θ_*. In this setting, a natural first requirement is that the statistical model is identifiable in the large data limit:

Assumption 7. *There exists a unique minimiser* θ_* *of* $\mathrm{DFD}^2(p_\theta \| p)$ *and there exists a sequence* $\{\theta_n\}_{n=1}^\infty$ *s.t.* θ_n *minimises* $\theta \mapsto \mathrm{DFD}^2(p_\theta \| p_n)$ *a.s. for all* n *sufficiently large. Further, there exists a bounded convex open set* $U \subseteq \Theta$ *s.t.* $\theta_* \in U$ *and* $\theta_n \in U$ *a.s. for all* n *sufficiently large.*

This requirement corresponds to the precondition C3 of Assumption 3. It is worth highlighting that Assumption 7 does not require the model family $\{p_\theta \mid \theta \in \Theta\}$ to contain p—i.e., p_θ can be misspecified—which is in contrast to the assumptions needed for the classical Bernstein–von Mises theorem (van der Vaart, 1998, Theorem 10.1). On the other hand, if the model family $\{p_\theta \mid \theta \in \Theta\}$ contains p uniquely, existence of θ_* is immediate since DFD is a divergence and hence $\mathrm{DFD}^2(p_\theta \| p) = 0$ if and only if $p_\theta = p$. Our second requirement is a technical condition on the derivatives of the model, to ensure that the asymptotic limit is well-defined. It is helpful to introduce the shorthand $r_{j-}(\boldsymbol{x}, \theta) := p_\theta(\boldsymbol{x}^{j-})/p_\theta(\boldsymbol{x})$, and to let ∇_θ^s denote s-times differentiation with respect to θ. For a function $g : \Theta \to \mathbb{R}$, let $\nabla_\theta^2 g(\theta) \in \mathbb{R}^{p\times p}$ with entries $\partial_i \partial_j g(\theta)$, and let $\nabla_\theta^3 g(\theta) \in \mathbb{R}^{p\times p\times p}$ with entries $\partial_i \partial_j \partial_k g(\theta)$.

Assumption 8. *Assume that* $p_\theta(\boldsymbol{x})$ *is three times continuously differentiable in* U *for each fixed* $\boldsymbol{x} \in \mathcal{X}$*, and*

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U} \|\nabla_\theta^s r_{j-}(X^{j+}, \theta)\|\right] < \infty \qquad \text{and} \qquad \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U} \|\nabla_\theta^s (r_{j-}(X, \theta)^2)\|\right] < \infty$$

for all $j = 1, \ldots, d$ *and* $s = 1, 2, 3$.

It is straightforward to verify Assumption 8 as opposed to Assumption 7, as illustrated in the following example. It considers the exponential family, a large model class that encompasses models used in Section 5.4. For example, any model on a space $\mathcal{X}$ of finite cardinality has a representation as an exponential family model (Amari, 2016, Ch. 2.2.2).

Example 1 (Exponential Family)**.** *Consider an exponential family model* $p_\theta(\boldsymbol{x}) \propto \exp(\eta(\theta)\cdot T(\boldsymbol{x}) + b(\boldsymbol{x}))$*, where* $\eta : \Theta \to \mathbb{R}^k$*,* $T : \mathcal{X} \to \mathbb{R}^k$ *and* $b : \mathcal{X} \to \mathbb{R}$ *for some* $k \in \mathbb{N}$*. For this model, we have* $r_{j-}(\boldsymbol{x}, \theta) = \exp(\eta(\theta)\cdot(T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})) + b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x}))$*. Assumption 8 is satisfied if, for* $j = 1, \ldots, d$*, (i)* $\|\eta(\theta)\|$ *and* $\|\nabla_\theta^s \eta(\theta)\|$ *for* $s = 1, 2, 3$ *are bounded over* $\theta \in U$*,*

(ii) $\|T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})\|$ *is bounded over* $\boldsymbol{x} \in \mathcal{X}$*, and (iii)* $\mathbb{E}_{X\sim p}[\exp(b(X^{j-}) - b(X))^2] < \infty$*. The requirements (ii) and (iii) are immediate if* $\mathcal{X}$ *is a finite set.*

The calculations that accompany this example are provided in Section 5.6.3. The following theorem establishes that the BvM theorem holds for DFD-Bayes.

Theorem 5. *Suppose that Assumptions 7 and 8 hold. Suppose that a prior* π *is positive and continuous at* θ_**. Denote by* $\widetilde{\pi}_n^D$ *a density on* $\mathbb{R}^p$ *of a random variable* $\sqrt{n}(\theta - \theta_n)$ *for* $\theta \sim \pi_n^D$*. If* $H_* := \nabla_\theta^2 \mathrm{DFD}^2(p_\theta \| p)|_{\theta=\theta_*}$ *is nonsingular, then*

$$\int_{\mathbb{R}^p} \left| \widetilde{\pi}_n^D(\theta) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}} \exp\left(-\frac{1}{2}\theta \cdot H_* \theta\right) \right| \mathrm{d}\theta \xrightarrow{\text{a.s.}} 0 \tag{5.8}$$

where the a.s. convergence is with respect to realisations of the dataset $\{x_i\}_{i=1}^n$*.*

The proof is contained in Section 5.6.5.

5.3.2. Connection to KSD-Bayes

Finally, this section elaborates a certain theoretical connection of DFD-Bayes and KSD-Bayes. To be precise, we deduce that DFD is topologically stronger than KSDs of essentially any kernels. Informally this means that the statistical efficiency of DFD-Bayes outperforms one of KSD-Bayes in case a model is well-specified, in a sense that the loss surface of DFD distinguishes an optimal parameter from other non-optimal parameters more distinctively than one of KSD.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a scalar-valued kernel with the associated reproducing kernel Hilbert space $\mathcal{H}_k$. Let $\mathcal{H}_k^d := \mathcal{H}_k \times \cdots \times \mathcal{H}_k$, that is, a space of functions $H : \mathcal{X} \to \mathbb{R}^d$ whose i-th output coordinate $H_i : \mathcal{X} \to \mathbb{R}$ belongs to $\mathcal{H}_k$ for each $i = 1, \ldots, d$. This space $\mathcal{H}_k^d$ is a simple case of vector-valued RKHS defined by the Cartesian product of the same scalar-valued RKHS $\mathcal{H}_k$. For simplicity, we use the unit ball of such vector-valued RKHS $\mathcal{H}_k^d$ as the Stein set to construct KSD. Recall from Proposition 7 that DFD is a Stein discrepancy built on the unit ball of $L^2(q, \mathbb{R}^d)$ as the Stein set. The following proposition shows that DFD built on the unit ball of $L^2(q, \mathbb{R}^d)$ dominates KSD built on the unit ball of $\mathcal{H}_k^d$ for essentially any kernel k.

Proposition 9. *Let* p *and* q *be probability distributions on* $\mathcal{X}$ *s.t.* $(\nabla^- p)/p, (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$*. Let* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a kernel such that* $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$*. Let* $\mathcal{S}_p$ *be a Stein operator in* (5.5)*. Then the corresponding KSD satisfies that* $\mathrm{KSD}(p\|q) \leq \mathrm{DFD}(p\|q)$*.*

The proof is contained in Section 5.6.6. Informally, there are two important implications of this result for each case when a model is well-specified and misspecified. If a model is well-specified, DFD-Bayes concentrates around an optimal parameter θ_* faster than KSD-Bayes, since $\mathrm{KSD}(p\|q) \leq \mathrm{DFD}(p\|q)$ and the larger loss induces a lighter-tailed generalised posterior. However, on the flip side, this simultaneously manifests stronger robustness of KSD, since KSD induces a heavier-tail generalised posterior than DFD and that limits

the sensitivity of KSD-Bayes to perturbation in data, such as a contaminating outlier. Indeed, our analysis in Section 4.3.3 illustrated compelling robustness of KSD-Bayes albeit in continuous case. This is also supported in Section B.2 by empirical comparison of DFD-Bayes and KSD-Bayes under severe model misspecification.

This argument is not restricted to discrete case and immediately applicable for continuous case. Recall the Langevin Stein operator (2.5) for continuous domain $\mathbb{R}^d$ and the continuous Fisher divergence $\mathrm{FD}(p\|q) = \mathbb{E}_{X\sim q}[\|\nabla \log p(X) - \nabla \log q(X)\|^2]$.

Proposition 10. *Let p and q be positive densities on $\mathbb{R}^d$ s.t. $\nabla \log p, \nabla \log q \in L^2(q, \mathbb{R}^d)$. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d} k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$. Let $\mathcal{S}_p$ be a Stein operator in (2.5). Then the corresponding KSD satisfies that $\mathrm{KSD}(p\|q) \leq \mathrm{FD}(p\|q)$.*

Proof. The proof immediately follows from the same argument as Section 5.6.6 using $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator (2.5) instead. □

This completes theoretical assessment of DFD-Bayes. We next provide detailed empirical assessment of DFD-Bayes.

5.4. Empirical Assessment

We perform an empirical assessment of DFD-Bayes, focusing on three important instances of discrete intractable likelihood. First, in Section 5.4.1 we consider a relatively simple model for over- and under-dispersed count data, called the Conway–Maxwell–Poisson model. Section 5.4.2 concerns an application to Ising-type models for discrete spatial data. Finally, we apply DFD-Bayes to perform inference for the parameters of flexible multivariate models for count data in Section 5.4.3. Source code to reproduce these experiments can be downloaded from https://github.com/takuomatsubara/Discrete-Fisher-Bayes.

5.4.1. Conway–Maxwell–Poisson Model

The first model we consider is a generalisation of the Poisson model for over- and under-dispersed count data, due to Conway and Maxwell (1962). This model is on $\mathcal{X} = \mathbb{N} \cup \{0\}$ (hence $d = 1$ and $\mathrm{card}(\mathcal{X}) = \infty$) and generalises the Poisson distribution through the inclusion of an additional parameter controlling how the data are dispersed. Since the work of Shmueli et al. (2005), this model has been used in a wide range of fields including transport, finance and retail. The model has two parameters $\theta \in \Theta = (0, \infty)^2 \cup ([0, 1] \times \{0\})$ (and hence $p = 2$) and its probability mass function is given by $p_\theta(x) = \tilde{p}_\theta(x) Z_\theta^{-1}$ where $\tilde{p}_\theta(x) = (\theta_1)^x (x!)^{-\theta_2}$. The normalising constant is given by $Z_\theta = \sum_{y=0}^{\infty} \tilde{p}_\theta(y)$, which has no analytical form except for certain special cases of $\theta \in \Theta$, including the case $\theta_2 = 1$ for which the standard Poisson model is recovered.

This model is an ideal test-bed for DFD-Bayes: although the likelihood is formally intractable, it is relatively straightforward to directly approximate the normalising con-

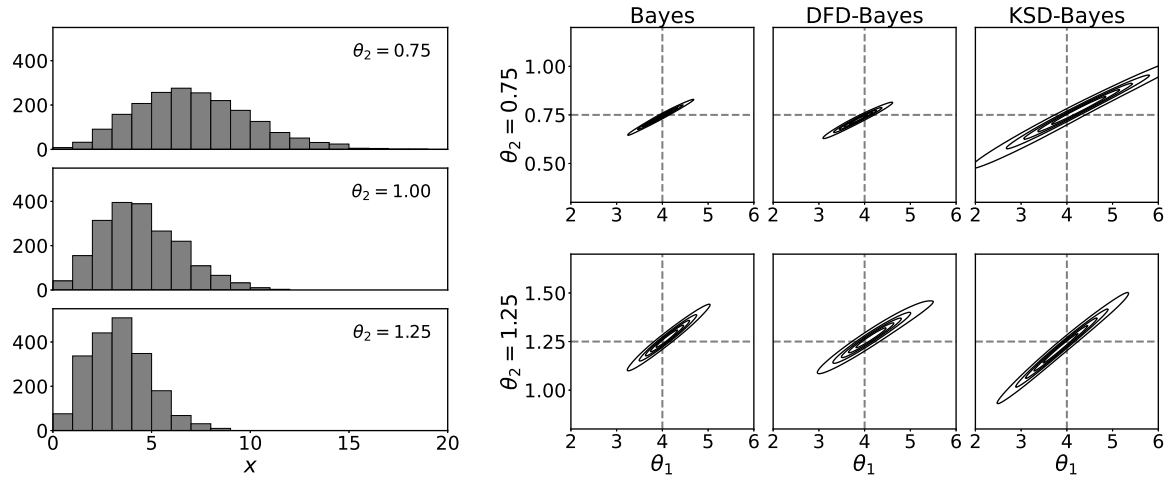


Figure 5.2 Comparison of standard Bayesian inference with the generalised posteriors from DFD-Bayes and KSD-Bayes on the Conway–Maxwell–Poisson model in the over-dispersed case $\theta_2 = 0.75$ and the under-dispersed case $\theta_2 = 1.25$ for $n = 2,000$.

stant[2]. This enables a direct comparison with standard Bayesian inference in the case where the model is well-specified. To this end, we simulated two datasets from the model: (i) an under-dispersed case where $\theta^* = (4, 1.25)$, and (ii) an over-dispersed case where $\theta^* = (4, 0.75)$, shown in Figure 5.2 (left). Three inference methods were compared: standard Bayesian inference, KSD-Bayes, and DFD-Bayes we have proposed. The settings of KSD-Bayes are described in Section B.3.1. In each case, the prior π was taken to be the chi-squared distribution with 3 degrees of freedom for each of θ_1 and θ_2 independently. A Metropolis–Hastings algorithm was used to sample from all the posteriors; and details can be found in Section B.3.2. The weight β in DFD-Bayes and KSD-Bayes was calibrated by our approach described in Section 3.4.

Figure 5.2 (right) illustrates the posteriors, based on typical datasets of size $n = 2,000$. The estimated value of β_* was 1.91 for DFD-Bayes and 5.04 for KSD-Bayes in the over-dispersed case $\theta_2 = 0.75$, and 0.46 for DFD-Bayes and 2.51 for KSD-Bayes in the under-dispersed case $\theta_2 = 1.25$. The left panel of Figure 5.3 displays the distribution of calibrated weight β_* as in Section 3.4 over multiple instances of the dataset, along with the values advocated in Lyddon et al. (2019). For both methods, the calibrated weight is stably estimated.

The inferences obtained using DFD-Bayes resembled those obtained using standard Bayesian inference, irrespective of whether the data were over- or under-dispersed. Those obtained using KSD-Bayes were more conservative than standard Bayes and DFD-Bayes, although the maximum a posteriori estimator approximated the true parameter well. Note that the credible regions of the generalised posteriors can substantially differ from those of standard Bayesian inference; in our approach a credible region of a generalised posterior is calibrated with reference to the distribution of a corresponding frequentist estimator

[2]The standard Bayesian inferences reported in this section used the approximation $Z_\theta \approx \sum_{y=0}^{99} \tilde{p}_\theta(y)$ and the associated approximate likelihood. Alternative estimators are available; see Benson and Friel (2021).

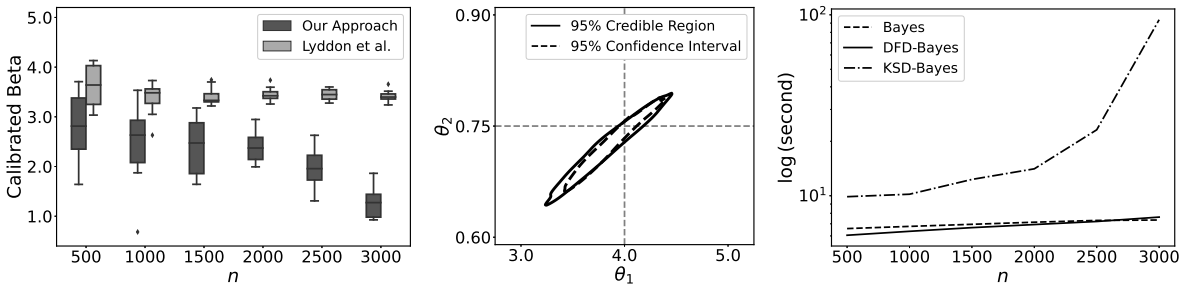


Figure 5.3 Distribution of β_* across different realisations of the dataset at each data number n for $\theta_2 = 0.75$ (left), comparison of a 95% credible region of the DFD-Bayes posterior and a 95% confidence interval of the frequentist counterpart for $n = 2000$ (centre), and comparison of computational times of each Metropolis–Hastings algorithm (right). The confidence interval was estimated by a 95% highest probability density region of a kernel density estimator applied to the 100 bootstrap minimisers used in calibration of β.

estimated by bootstrapping, leading to approximately correct frequentist coverage as shown in Figure 5.3 (middle). Calibration led to improved inference outcomes for both DFD-Bayes and KSD-Bayes. In KSD-Bayes case for example, the value of $\beta_* \geq 1$ intensified the concentration around the true parameter by placing more importance on the loss than the prior. In addition, our approach to calibration is relatively more conservative than Lyddon et al. (2019) because the prior is taken into account.

There is a stark difference in computational cost between DFD-Bayes and KSD-Bayes[3], as demonstrated in the right panel of Figure 5.3. Indeed, the computational cost of DFD-Bayes is seen to increase linearly with n, while the cost of KSD-Bayes increases quadratically.

Finally, to assess performance in a real-world data setting, we apply DFD-Bayes to infer the parameters of a Conway–Maxwell–Poisson model using the sales dataset of Shmueli et al. (2005). All relevant details are contained in Section B.3.3. Figure 5.4 compares our fitted model to a standard Bayesian analysis using the Poisson distribution, which is the closest analysis one can perform without confronting an intractable likelihood. As observed in the central panel of Figure 5.4, the Poisson model is not able to capture over-dispersion of the data, whereas the Conway–Maxwell–Poisson model fitted using DFD-Bayes, shown in the right panel, provides a reasonable fit. The DFD-Bayes posterior (left) appears approximately normal, in line with Theorem 5.

5.4.2. Ising Model

The aim of this section is to consider a more challenging instance of discrete intractable likelihood, where the data are high-dimensional (i.e. d is large) and the cardinality of each coordinate domain S_i is small. A small cardinality of S_i is particularly interesting, because the intuition that our difference operators arise from discretisation of continuous differential

[3]The cost of standard Bayesian inference in this experiment is entirely determined by the accuracy with which the normalisation constant is approximated; since direct approximation of the normalisation constant is infeasible in general, we do not report this cost.

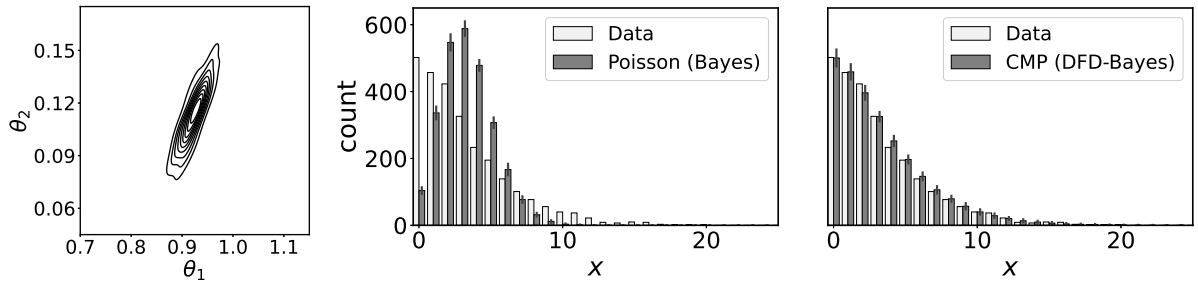


Figure 5.4 Comparison of DFD-Bayes for the Conway–Maxwell–Poisson model and standard Bayes for the Poisson distribution on the sales data of Shmueli et al. (2005). Left: The generalised posterior distribution produced using DFD-Bayes. Centre: Posterior predictive distribution, at the level of the data, for a Poisson model with standard Bayesian inference performed. Right: Posterior predictive distribution, at the level of the data, for a Conway–Maxwell–Poisson model with DFD-Bayes inference performed. In both cases, error bars indicate one standard deviation of the posterior predictive distribution.

operators fails to hold. This setting is typified by the Ising model (which has $S_i = \{0, 1\}$), variants of which are used to model diverse phenomena, such as the network structure of the amino-acid sequences (Xue et al., 2012). The computational challenge of performing Bayesian inference for Ising-type models has, to-date, principally been addressed using techniques such as pseudo-likelihood (see the recent survey in Bhattacharyya and Atchade, 2019). Unfortunately, these do not necessarily lead to asymptotically exact inference since the correct likelihood is replaced by an approximation.

Let G be an undirected graph on a d-dimensional vertex set and let $\mathcal{N}_i$ denote the neighbours of a node i, with self-edges excluded. An Ising model describes a discrete process that assigns each vertex of G either the value 0 or 1, and thus the data domain is $\mathcal{X} = \{0, 1\}^d$. The probability mass function has the exponential family form

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\frac{1}{\theta}\sum_{i=1}^{d}\sum_{j\in\mathcal{N}_i} x_i x_j\right) \tag{5.9}$$

where θ is a temperature parameter, controlling the propensity for neighbouring vertices to share a common value. Here we consider the ferromagnetic Ising model, which has $\theta \in (0, \infty)$. To conduct a simulation study, we consider the case where G is a $m \times m$ grid. Simulating from Ising models is challenging due to the high-dimensional discrete domain, so here we restrict attention to $m \in \{5, \ldots, 10\}$ to ensure that data were accurately simulated[4]. A total of $n = 1,000$ data points were generated from an Ising model with $\theta = 5$, using an extended run of a Metropolis—Hastings algorithm, the details of which are contained in Section B.3.4. A chi-squared prior with degree of freedom 3 was used. Three inference methods were compared: the KSD-Bayes method, the proposed DFD-Bayes

[4]The value of m used in these experiments was not constrained by the computational demand of DFD-Bayes, which scales as $O(m^2)$.

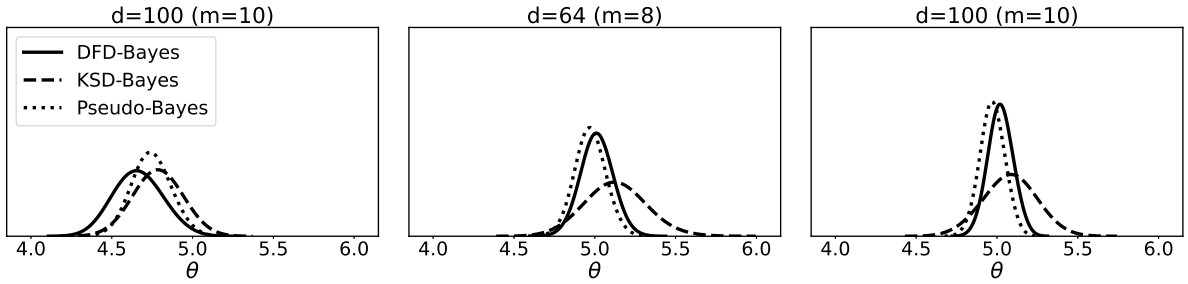


Figure 5.5 Comparison of approximate Bayesian inference based on pseudo-likelihood, DFD-Bayes and KSD-Bayes, applied to the Ising model with $\theta = 5$ for $n = 1,000$ and $d = 10 \times 10$. For all methods, the value β_* from Section 3.4 was used.

method, and standard Bayesian inference based on the pseudo-likelihood

$$\tilde{p}_\theta(\boldsymbol{x}) = \prod_{i=1}^{d} p_\theta(x_i | \{x_j : j \in \mathcal{N}_i\}),$$

where $p_\theta(x_i|\{x_j : j \in \mathcal{N}_i\})$ is a restriction of the original model (5.9) to the i-th coordinate x_i under the condition $\{x_j : j \in \mathcal{N}_i\}$ that results in a Bernoulli distribution of x_i for each $i = 1, \cdots, d$ (Besag, 1974). The latter Pseudo-Bayes approach can be viewed as a special case of generalised Bayes inference, since it replaces the original likelihood loss of the model (5.9) with the pseudo-likelihood loss, and therefore we also applied the proposed calibration procedure to this method. The settings of KSD-Bayes are described in Section B.3.5. A Metropolis–Hastings algorithm was also used to sample from all generalised posteriors, the details for which are contained in Section B.3.6.

Results are presented for three different datasets of size $n = 1,000$ and dimension $d = 36$ ($m = 6$), $d = 64$ ($m = 8$), and $d = 100$ ($m = 10$) in Figure 5.5. For the lowest dimension $d = 36$, all the approaches produced similar posteriors. For the higher-dimensional cases, it can be seen that the DFD-Bayes and Pseudo-Bayes posteriors concentrate around the true parameter $\theta = 5$. The KSD-Bayes posterior is more conservative, whilst DFD-Bayes gives a comparable result to Pseudo-Bayes. For $d = 100$, the total computational time required to perform this analysis (including calibration) was 540 seconds for DFD-Bayes, 2,353 seconds for KSD-Bayes, and 1,053 seconds for Pseudo-Bayes each in average over 10 independent experiments, confirming that DFD-Bayes incurs a significantly lower computational cost than both alternatives. The value of the weight obtained through our calibration method for $d = 100$ in Figure 5.5 was 0.013 for DFD-Bayes, 0.157 for KSD-Bayes, and 0.579 for Pseudo-Bayes. These small values of weight indicated that the calibration worked effectively, preventing the over-concentration of each posterior.

5.4.3. Multivariate Count Data

Finally, we consider a problem involving multivariate count data. Count data occur in diverse application areas, and variables in such data are rarely independent, yet the literature on statistical modelling of such data is limited. Poisson graphical models and

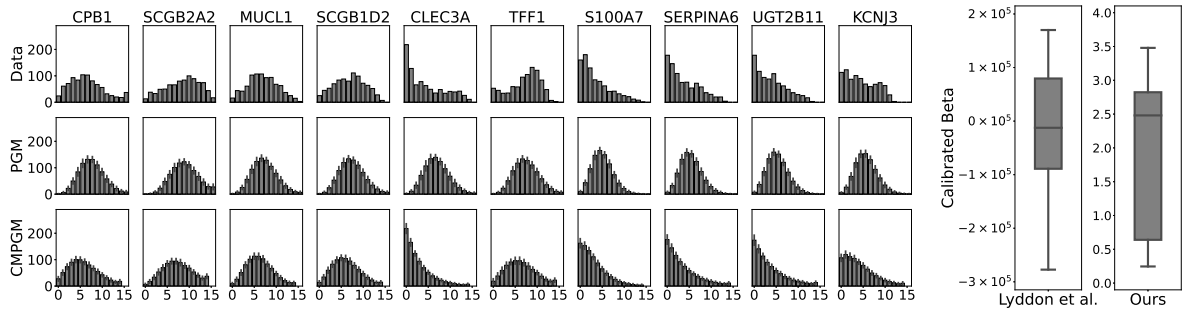


Figure 5.6 Left: Posterior predictive distributions from the Poisson graphical model and the Conway–Maxwell–Poisson graphical model. Right: Sampling distributions of β_* for the Conway–Maxwell–Poisson graphical model by Lyddon et al. (2019) and by the proposed approach, computed using 10 independent realisations of the dataset.

their extensions have emerged as a powerful tool for modelling such data; see the recent review of Inouye et al. (2017). To the best of our knowledge a complete Bayesian treatment of Poisson graphical models has yet to be attempted[5], and we speculate that this is due to the computational challenges of the associated intractable likelihood. Our aim here is to assess the suitability of DFD-Bayes for learning the parameters of a Poisson graphical model.

Let G be an undirected graph on a vertex set $\{1, \ldots, d\}$ and let $\mathcal{M}_i$ denote the neighbours of a node i that are contained in the set $\{i+1, \ldots, d\}$. A Poisson graphical model has a probability mass function

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d} \theta_i x_i - \sum_{i=1}^{d} \sum_{j \in \mathcal{M}_i} \theta_{i,j} x_i x_j - \sum_{i=1}^{d} \log(x_i!)\right)$$

where the parameters θ consist of both the linear coefficients $\theta_i \in (-\infty, \infty)$ and the interaction coefficients $\theta_{i,j} \in [0, \infty)$. Our aim is to reproduce an analysis of a breast cancer gene expression dataset described in Inouye et al. (2017), but in a generalised Bayesian framework. For this problem, $n = 878$, $d = 10$, and $p = 64$ which renders the computational cost of $O(n^2 d)$ at every MCMC step and of $O(p^2 n^2 d)$ at calibration associated with KSD-Bayes inefficient. Full details of the dataset are contained in Section B.3.7. Independent standard normal priors were employed for each θ_i, and half-normal distributions with scale $(d(d-1)/2)^{-1}$ were employed for each $\theta_{i,j}$. A No-U-Turn Sampler was used to sample from the DFD-Bayes posterior, as described in Section B.3.8. The total computational time required to perform this analysis, including calibration, was $1,896$ seconds. Results, in Figure 5.6, demonstrate that the Poisson graphical model is in fact a poor fit for these data, since the data show signs of being under-dispersed relative to the standard Poisson model. However, in terms of identifying the best parameter values for this model, DFD-Bayes appears to have performed well.

[5]Though we note that a pairwise Markov random field whose marginals are close to being Poisson was considered in Roy and Dunson (2020), and a specific generalisation of the Conway-Maxwell-Poisson was used in Piancastelli et al. (2021).

As a possible improvement, and to further stress-test the DFD-Bayes method, we considered a generalisation of the Poisson graphical model that allows for over- and under-dispersion, analogous to Conway and Maxwell (1962). This model takes the form

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d}\theta_i x_i - \sum_{i=1}^{d}\sum_{j\in\mathcal{M}_i}\theta_{i,j}x_i x_j - \sum_{i=1}^{d}\theta_{0,i}\log(x_i!)\right)$$

where the additional parameters $\theta_{0,i} \in [0,\infty)$ control the dispersion, with $\theta_{0,i} = 1$ recovering the standard Poisson marginal. This time, $p = 74$ as opposed to $p = 64$ for the Poisson-based model. For this Conway–Maxwell–Poisson graphical model, the same priors as the Poisson graphical model were used for θ_i and $\theta_{i,j}$, and half-normal priors with scale $1/\sqrt{2}$ were used for each $\theta_{0,i}$. Results in Figure 5.6 demonstrate an improved fit to the dataset. Indeed, the optimal β for the Poisson graphical model was $\beta_* = 0.2150$, which is smaller than the corresponding value $\beta_* = 0.9971$ for the Conway–Maxwell–Poisson graphical model, resulting in a conservative inference outcome when the statistical model is most misspecified and supporting the effectiveness of the proposed approach to calibration.

The right panel of Figure 5.6 shows the sampling distributions of estimators for the weight β in the context of the Conway–Maxwell–Poisson graphical model, computed using bootstrap resampling of the gene expression dataset. It can be seen that the asymptotic approach proposed in Lyddon et al. (2019) is severely numerically unstable and can even lead to a negative weight, while the approach proposed in Section 3.4 remains stable within a reasonable range between 0 and 3.5. The lack of stability of the approach by Lyddon et al. (2019) arises from the need to invert a covariance matrix of derivatives of the loss, which can become numerically singular if the parameter dimension is high. In contrast, our approach involves no matrix inversion. This real-data analysis using flexible parametric models highlights the value in being able to perform rapid and automatic (i.e. free from user-specified degrees of freedom) generalised Bayesian inference for discrete intractable likelihood.

5.5. Concluding Remark

In this chapter, we proposed DFD-Bayes, that is, the SD-Bayes methodology resulting from the use of DFD. Similarly to KSD-Bayes in Chapter 4, DFD-Bayes is computable by any standard MCMC algorithms even if intractable models are used. While both KSD-Bayes and DFD-Bayes does not admit the conjugate inference in the discrete case, DFD-Bayes enjoys the linear computational cost $\mathcal{O}(n)$ to the data size n in contrast to the quadratic cost $\mathcal{O}(n^2)$ of KSD-Bayes. DFD-Bayes also benefits from independence of user-specified hyperparameter, such as a kernel in KSD-Bayes whose natural choice in discrete domains is often impractical to compute. It achieves highly efficient inference with few hyperparameters for discrete intractable models. There was little concern about robustness for real-world datasets in Chapter 4, where model misspecification associated

with the use of tractable models in standard Bayesian inference was resolved by using more complex intractable models with DFD-Bayes. In circumstances where robustness is concerned, KSD-Bayes can produce enhanced robustness as discussed in Section B.2.

5.6. Proofs of Chapter 5

This section contains all the deferred proof of theoretical results in Chapter 5. The proof of posterior consistency of DFD-Bayes is placed in the main text since it is immediate from Theorem 1 and proposition 1 in Chapter 3. The proofs of other results and useful lemmas are contained in this section.

5.6.1. Proof of Proposition 6

First, we introduce three technical lemmas that will be useful:

Lemma 14. *For any $\boldsymbol{x} \in \mathcal{X}$ and $i = 1, \ldots, d$, it holds that $(\boldsymbol{x}^{i-})^{i+} = \boldsymbol{x}$ and $(\boldsymbol{x}^{i+})^{i-} = \boldsymbol{x}$.*

Proof. Since $\mathcal{X} = S_1 \times \cdots \times S_d$ from the Standing Assumption,

$$\boldsymbol{x}^{i-} = (x_1, \cdots, x_i^-, \cdots, x_d), \qquad \boldsymbol{x}^{i+} = (x_1, \cdots, x_i^+, \cdots, x_d). \tag{5.10}$$

It is thus sufficient to show that $(x_i^-)^+ = x_i$ and $(x_i^+)^- = x_i$ for any $i = 1, \ldots, d$. Consider, therefore, a set $S \cong I \subseteq \mathbb{Z}$ with more than one element. Our aim is to establish the identity $(s^-)^+ = s$ and $(s^+)^- = s$ for all $s \in S$. Existence of the least and greatest element $s_{\min}$ and $s_{\max}$ of S determines four qualitatively distinct cases to be checked: (i) neither of them exist; (ii) both of them exists; (iii) only $s_{\min}$ exists; (iv) only $s_{\max}$ exists. Recall that we identify the case (iv) with (iii) without loss of generality by reversing the ordering of S. The identity for (i) & (ii) is trivial since the maps $s \mapsto s^-$ is bijective from S to itself with inverse $s \mapsto s^+$. For case (iii), we have $(s^-)^+ = s$ for $s \neq s_{\min}$ and $(s^+)^- = s$ for all $s \in S$. Recalling the definition $s_{\min}^- = \star$ and $\star^+ = s_{\min}$ completes the argument. □

Lemma 15. *For any $f, g : \mathcal{X} \to \mathbb{R}$ and any $i = 1, \ldots, d$, suppose $\sum_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x}) g(\boldsymbol{x}^{i-})| < \infty$, that is, the series is absolutely convergent. Then we have*

$$\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}^{i-}) = \sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}^{i+}) g(\boldsymbol{x}). \tag{5.11}$$

Proof. Since $\mathcal{X} = S_1 \times \cdots \times S_d$ from the Standing Assumption 1, the series can be expressed as

$$\begin{aligned}
\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}^{i-}) &= \sum_{x_1 \in S_1} \cdots \sum_{x_i \in S_i} \cdots \sum_{x_d \in S_d} f(x_1, \cdots, x_i, \cdots, x_d) g(x_1, \cdots, x_i^-, \cdots, x_d), \\
\sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}^{i+}) g(\boldsymbol{x}) &= \sum_{x_1 \in S_1} \cdots \sum_{x_i \in S_i} \cdots \sum_{x_d \in S_d} f(x_1, \cdots, x_i^+, \cdots, x_d) g(x_1, \cdots, x_i, \cdots, x_d).
\end{aligned}$$

Holding the coordinates $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d$ fixed, and exploiting absolute convergence to justify the interchange of summations, the claimed result follows if

$$\sum_{x_i \in S_i} \tilde{f}(x_i)\tilde{g}(x_i^-) = \sum_{x_i \in S_i} \tilde{f}(x_i^+)\tilde{g}(x_i) \tag{5.12}$$

where $\tilde{f}(x_i) := f(x_1, \cdots, x_i, \cdots, x_d)$ and $\tilde{g}(x_i) := g(x_1, \cdots, x_i, \cdots, x_d)$ are viewed as functions on S_i.

Consider, therefore, an arbitrary set $S \cong I \subseteq \mathbb{Z}$, for which we aim to establish the identity $\sum_{s\in S} h(s)k(s^-) = \sum_{s\in S} h(s^+)k(s)$ for any functions $h, k : S \to \mathbb{R}$ s.t. $\sum_{s\in S} |h(s)k(s^-)| < \infty$. From the definition of an order isomorphism, the elements of S can be indexed as $S = \{s_i : i \in I\}$, where $s_i < s_j$ if and only if $i < j$. The identity therefore can be written as $\sum_{i\in I} h(s_i)k(s_i^-) = \sum_{i\in I} h(s_i^+)k(s_i)$, and will be verified for the three qualitatively distinct cases of index set I described in the proof of Lemma 14:

(i) $I = \mathbb{Z}$. The result is immediate, since $(s_i, s_i^-) = (s_i, s_{i-1})$ and $(s_i^+, s_i) = (s_{i+1}, s_i)$ range over the same set for $i \in I$. The series $\sum_{i\in I} h(s_i^+)k(s_i)$ is absolutely convergent since the sets $\{h(s_i)k(s_i^-)\}_{i\in I}$ and $\{h(s_i^+)k(s_i)\}_{i\in I}$ in the two series are equal.

(ii) $I = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$. In this case $s_{\min} = s_1$ and $s_{\max} = s_n$, and it follows from the definition of decrements and increments that

$$\begin{aligned}\sum_{i\in I} h(s_i)k(s_i^-) &= h(s_1)k(s_1^-) + h(s_2)k(s_1) + \cdots + h(s_n)k(s_{n-1}) \\ &= h(s_n^+)k(s_n) + h(s_2)k(s_1) + \cdots + h(s_n)k(s_{n-1}) = \sum_{i\in I} h(s_i^+)k(s_i),\end{aligned}$$

where the sets $\{h(s_i)k(s_i^-)\}_{i\in I}$ and $\{h(s_i^+)k(s_i)\}_{i\in I}$ are again equal.

(iii) $I = \{1, 2, \ldots\}$. In this case $s_{\min} = s_1$, and it follows from $s_1^- = \star$ and $k(\star) = 0$ that

$$\begin{aligned}\sum_{i\in I} h(s_i)k(s_i^-) &= \underbrace{h(s_1)k(\star)}_{=0} + h(s_2)k(s_1) + h(s_3)k(s_2) + \cdots \\ &= h(s_2)k(s_1) + h(s_3)k(s_2) + \cdots = \sum_{i\in I} h(s_i^+)k(s_i).\end{aligned}$$

The series $\sum_{i\in I} h(s_i^+)k(s_i)$ is absolutely convergent since the set $\{h(s_i^+)k(s_i)\}_{i\in I}$ is a subset of the absolutely summable set $\{h(s_i)k(s_i^-)\}_{i\in I}$.

This completes the proof. □

Lemma 16. *Let $\mathcal{P}_q(\mathcal{X})$ be a set of probability mass functions absolutely continuous to q on $\mathcal{X}$. The map $\mu_p := (\nabla^- p)/p$ for $p \in \mathcal{P}_q(\mathcal{X})$ is an injection $\mu : \mathcal{P}_q(\mathcal{X}) \to L^2(q, \mathbb{R}^d)$.*

Proof. It suffices to show that each value $p(\boldsymbol{x})$, for $\boldsymbol{x}$ in the support of q, can be explicitly recovered from μ_p. Note that, since p takes values in $(0, \infty)$ in the support of q, the embedding μ_p is well-defined in the support of q, which is sufficient to be an element of

$L^2(q, \mathbb{R}^d)$. From the Standing Assumption 1, we have that $\mathcal{X} = S_1 \times \cdots \times S_d$, where each $S_i \cong I_i \subseteq \mathbb{Z}$ is a set with more than one element. Since S_i serve only as index sets, we can without loss of generality assume that S_i is a consecutive subset of $\mathbb{Z}$ and that $0 \in S_i$, for each $i = 1, \ldots, d$. The idea of the proof is first to demonstrate that each of the quantities $p(\boldsymbol{x})$ can be explicitly expressed in terms of μ_p, $p(\mathbf{0})$ and $\{p(\boldsymbol{y}) : \|\boldsymbol{y}\|_1 < \|\boldsymbol{x}\|_1\}$, where $\|\boldsymbol{x}\|_1 := |x_1| + \cdots + |x_d|$. It would then follow from a simple inductive argument that $p(\boldsymbol{x})$ can be expressed in terms of μ_p and $p(\mathbf{0})$. Finally, the constraint that $\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) = 1$ uniquely determines $p(\mathbf{0})$, demonstrating that $p(\boldsymbol{x})$ can be explicitly recovered.

Given $\boldsymbol{x} \in \mathcal{X}$, assume $\boldsymbol{x} \neq \mathbf{0}$, for otherwise the claim will trivially hold. Then let $i \in \{1, \ldots, d\}$ be such that $x_i \neq 0$. If $x_i > 0$, then from the definition of $\mu_p(\boldsymbol{x})_i = 1 - p(\boldsymbol{x}^{i-})/p(\boldsymbol{x})$ we have the relation

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}^{i-})}{1 - \mu_p(\boldsymbol{x})_i}.$$

Conversely, if $x_i < 0$, then using Lemma 14 we have $\mu_p(\boldsymbol{x}^{i+})_i = 1 - p(\boldsymbol{x})/p(\boldsymbol{x}^{i+})$ and we have the relation

$$p(\boldsymbol{x}) = [1 - \mu_p(\boldsymbol{x}^{i+})_i] p(\boldsymbol{x}^{i+}).$$

The previously described inductive argument completes the proof. □

Now we prove the main result:

Proof of Proposition 6. Expanding the square gives that

$$\mathrm{DFD}(p\|q) = \mathbb{E}_{X \sim q}\Bigg[\sum_{i=1}^{d} \Bigg(\frac{p(X) - p(X^{j-})}{p(X)}\Bigg)^2 - 2\underbrace{\frac{p(X) - p(X^{j-})}{p(X)} \frac{q(X) - q(X^{j-})}{q(X)}}_{=:(*)} + \Bigg(\frac{q(X) - q(X^{j-})}{q(X)}\Bigg)^2\Bigg].$$

Denote by $\mathrm{supp}(q)$ the support of q i.e. $\{\boldsymbol{x} \in \mathcal{X} \mid q(\boldsymbol{x}) \neq 0\}$. For the term $\mathbb{E}_{X \sim q}[(*)]$, it follows from the definition $\mathbb{E}_{X \sim q}[f(X)] = \sum_{\boldsymbol{x} \in \mathrm{supp}(q)} f(\boldsymbol{x}) q(\boldsymbol{x})$ that

$$\begin{aligned}
\mathbb{E}_{X \sim q}[(*)] &= \sum_{j=1}^{d} \mathbb{E}_{X \sim q}\left[\frac{p(X) - p(X^{j-})}{p(X)} - \frac{p(X) - p(X^{j-})}{p(X)} \frac{q(X^{j-})}{q(X)}\right] \\
&= \sum_{j=1}^{d} \Bigg\{ \sum_{\boldsymbol{x} \in \mathrm{supp}(q)} \frac{p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})} q(\boldsymbol{x}) - \underbrace{\sum_{\boldsymbol{x} \in \mathrm{supp}(q)} \frac{p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})} q(\boldsymbol{x}^{j-})}_{(**)} \Bigg\},
\end{aligned}$$

We apply Lemma 15 to the term $(**)$ with $f(\boldsymbol{x}) = (p(\boldsymbol{x}) - p(\boldsymbol{x}^{j-}))/p(\boldsymbol{x})$ and $g(\boldsymbol{x}) = q(\boldsymbol{x})$, where $f(\boldsymbol{x}^{j+})$ is well-defined for all $\boldsymbol{x} \in \mathrm{supp}(q)$ due to the assumption that $p(\boldsymbol{x}^{j+}) > 0$

for $\boldsymbol{x} \in \operatorname{supp}(q)$ and Lemma 15 is thus applicable. This reveals that

$$(**) = \sum_{\boldsymbol{x} \in \operatorname{supp}(q)} \frac{p(\boldsymbol{x}^{j+}) - p(\boldsymbol{x})}{p(\boldsymbol{x}^{j+})} q(\boldsymbol{x})$$

for each $j = 1, \ldots, d$, where Lemma 14 is used to deduce that $(\boldsymbol{x}^{j-})^{j+} = \boldsymbol{x}$. Hence,

$$\mathbb{E}_{X \sim q}[(*)] = \mathbb{E}_{X \sim q}\left[\sum_{i=1}^{d} \frac{p(X) - p(X^{j-})}{p(X)} - \frac{p(X^{j+}) - p(X)}{p(X^{j+})}\right] = \mathbb{E}_{X \sim q}\left[\sum_{i=1}^{d} -\frac{p(X^{j-})}{p(X)} + \frac{p(X)}{p(X^{j+})}\right].$$

Plugging this equality in DFD at the top and completing the expansion establish that

$$\begin{aligned}
\operatorname{DFD}(p\|q) &= \mathbb{E}_{X \sim q}\left[\sum_{i=1}^{d} \left(1 - \frac{p(X^{j-})}{p(X)}\right)^2 + 2\frac{p(X^{j-})}{p(X)} - 2\frac{p(X)}{p(X^{j+})} + \left(1\frac{q(X^{j-})}{q(X)}\right)^2\right] \\
&= \mathbb{E}_{X \sim q}\left[\sum_{i=1}^{d} \left(\frac{p(X^{j-})}{p(X)}\right)^2 - 2\frac{p(X)}{p(X^{j+})}\right] + \underbrace{\mathbb{E}_{X \sim q}\left[\sum_{i=1}^{d} 1 + \left(1 - \frac{q(X^{j-})}{q(X)}\right)^2\right]}_{=:C(q)}.
\end{aligned}$$

Finally we verify that $\operatorname{DFD}(p\|q) = 0$ if and only if $p = q$. From Lemma 16 we have the injective embedding $p \mapsto \mu_p := (\nabla^- p)/p$. Clearly, the map $p \mapsto \mu_p$ is also an injection into $L^2(q, \mathbb{R}^d)$, equipped with the canonical norm $\|\nu\|_{L^2(q,\mathbb{R}^d)} := \mathbb{E}_{X \sim q}[\|\nu(X)\|^2]$ for $\nu \in L^2(q, \mathbb{R}^d)$. From (5.2) we recognise that $\operatorname{DFD}(p\|q) = \|\mu_p - \mu_q\|^2_{L^2(q,\mathbb{R}^d)}$ is the squared distance between μ_p and μ_q according to the canonical norm of $L^2(q, \mathbb{R}^d)$. Since $\|\mu_p - \mu_q\|_{L^2(q,\mathbb{R}^d)} = 0$ if and only if $\mu_p = \mu_q$ in $L^2(q, \mathbb{R}^d)$, it follows from injectivity of $p \mapsto \mu_p$ that $\operatorname{DFD}(p\|q) = 0$ if and only if $p = q$, as required. □

5.6.2. Proof of Proposition 7

Proof. From (2.6) and (5.5), we have that

$$\operatorname{SD}(p\|q) = \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{X \sim q}\left[\frac{\nabla^- p(X)}{p(X)} \cdot h(X) - \frac{\nabla^- q(X)}{q(X)} \cdot h(X)\right]\right|.$$

Note that $L^2(q, \mathbb{R}^d)$ is a Hilbert space when equipped with the inner product $\langle f, g\rangle_{L^2(q,\mathbb{R}^d)} := \mathbb{E}_{X \sim q}[f(X) \cdot g(X)]$. Thus, we can view $\operatorname{SD}(p\|q)$ as the maximum of the inner product

$$\operatorname{SD}(p\|q) = \sup_{h \in \mathcal{H}} \left| \left\langle \frac{\nabla^- p}{p} - \frac{\nabla^- q}{q}, h \right\rangle_{L^2(q,\mathbb{R}^d)} \right|, \tag{5.13}$$

which is well-defined since $u := (\nabla^- p)/p - (\nabla^- q)/q \in L^2(q, \mathbb{R}^d)$. Let $\|\cdot\|_{L^2(q,\mathbb{R}^d)}$ denote the norm of $L^2(q, \mathbb{R}^d)$, so that $\mathcal{H}$ is the set of $f \in L^2(q, \mathbb{R}^d)$ for which $\|f\|_{L^2(q,\mathbb{R}^d)} \leq 1$. By the Cauchy–Schwarz inequality, the inner product in (5.13) attains its supremum at

$h = u/\|u\|_{L^2(q,\mathbb{R}^d)} \in \mathcal{H}$. Therefore

$$\mathrm{SD}(p\|q) = \sup_{h\in\mathcal{H}} |\langle u, h\rangle_{L^2(q,\mathbb{R}^d)}| = \|u\|_{L^2(q,\mathbb{R}^d)} = \sqrt{\mathbb{E}_{X\sim q}\left[\left\|\frac{\nabla^- p(X)}{p(X)} - \frac{\nabla^- q(X)}{q(X)}\right\|^2\right]},$$

which concludes the proof. □

5.6.3. Assumption 8 for Exponential Family

The aim of this section is to establish when Assumption 8 is satisfied for the exponential family model. For better presentation, let $T_{j-}(\boldsymbol{x}) := T(\boldsymbol{x}^{j-}) - T(\boldsymbol{x})$ and $b_{j-}(\boldsymbol{x}) := b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})$ to see that $r_{j-}(\boldsymbol{x},\theta) = \exp(\eta(\theta)\cdot T_{j-}(\boldsymbol{x}) + b_{j-}(\boldsymbol{x}))$. In addition, let $T_{j+}(\boldsymbol{x}) := T(\boldsymbol{x}) - T(\boldsymbol{x}^{j+})$ and $b_{j+}(\boldsymbol{x}) := b(\boldsymbol{x}) - b(\boldsymbol{x}^{j+})$ to see that $r_{j-}(\boldsymbol{x}^{j+},\theta) = \exp(\eta(\theta)\cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x}))$. It is straightforward to see that, for any $\boldsymbol{x}\in\mathcal{X}$,

$$\begin{aligned}
\nabla_\theta r_{j-}(\boldsymbol{x}^{j+},\theta) &= \nabla_\theta\eta(\theta)\cdot T_{j+}(\boldsymbol{x})\,\exp(\eta(\theta)\cdot T_{j+}(\boldsymbol{x}) + b_{j+}(\boldsymbol{x})) \\
&= \nabla_\theta\eta(\theta)\cdot T_{j+}(\boldsymbol{x})\,\exp(\eta(\theta)\cdot T_{j+}(\boldsymbol{x}))\,\exp(b_{j+}(\boldsymbol{x})) \\
\nabla_\theta (r_{j-}(\boldsymbol{x},\theta)^2) &= 2 r_{j-}(\boldsymbol{x},\theta)\,\nabla_\theta r_{j-}(\boldsymbol{x},\theta) \\
&= 2\nabla_\theta\eta(\theta)\cdot T_{j-}(\boldsymbol{x})\,\exp(2\eta(\theta)\cdot T_{j-}(\boldsymbol{x}))\,\exp(2b_{j-}(\boldsymbol{x}))
\end{aligned}$$

By assumption, $T_{j-}(\boldsymbol{x})$ is bounded over all $\boldsymbol{x}\in\mathcal{X}$, which in turn shows that $T_{j+}(\boldsymbol{x}) = T_{j-}(\boldsymbol{x}^{j+})$ is bounded over all $\boldsymbol{x}\in\mathcal{X}$ since $\boldsymbol{x}^{j+}\in\mathcal{X}$. Further, by assumption, $\sup_{\theta\in U}\|\nabla_\theta\eta(\theta)\| < \infty$ and $\sup_{\theta\in U}\|\eta(\theta)\| < \infty$. Let M be a constant that upper bounds all the terms $\sup_{\boldsymbol{x}\in\mathcal{X}}\|T_{j-}(\boldsymbol{x})\|$, $\sup_{\boldsymbol{x}\in\mathcal{X}}\|T_{j+}(\boldsymbol{x})\|$, $\sup_{\theta\in U}\|\nabla_\theta\eta(\theta)\|$ and $\sup_{\theta\in U}\|\eta(\theta)\|$. Then we have

$$\begin{aligned}
\sup_{\theta\in U}\|\nabla_\theta r_{j-}(\boldsymbol{x}^{j+},\theta)\| &\le M^2\exp\left(M^2\right)\exp(b_{j+}(\boldsymbol{x})), \\
\sup_{\theta\in U}\|\nabla_\theta (r_{j-}(\boldsymbol{x},\theta)^2)\| &\le 2M^2\exp\left(2M^2\right)\exp(2b_{j-}(\boldsymbol{x})).
\end{aligned}$$

Taking the expectations,

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta r_{j-}(X^{j+},\theta)\|\right] \le M^2\exp(M^2)\mathbb{E}_{X\sim p}\left[\exp(b_{j+}(X))\right], \tag{5.14}$$

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta (r_{j-}(X,\theta)^2)\|\right] \le 2M^2\exp(2M^2)\mathbb{E}_{X\sim p}\left[\exp(2b_{j-}(X))\right]. \tag{5.15}$$

By assumption $\mathbb{E}_{X\sim p}\left[\exp(2b_{j-}(X))\right] = \mathbb{E}_{X\sim p}\left[\exp(b_{j-}(X))^2\right] < \infty$, and we now argue that this also implies $\mathbb{E}_{X\sim p}\left[\exp(b_{j+}(X))\right] < \infty$. Indeed, from Lemma 15,

$$\begin{aligned}
\mathbb{E}_{X\sim p}\left[\exp(b_{j+}(X))\right] &= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\exp(b(\boldsymbol{x}) - b(\boldsymbol{x}^{j+})) = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}^{j-})\exp(b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\frac{p(\boldsymbol{x}^{j-})}{p(\boldsymbol{x})}\exp(b(\boldsymbol{x}^{j-}) - b(\boldsymbol{x})) = \mathbb{E}_{X\sim p}\left[\frac{p(X^{j-})}{p(X)}\exp(b_{j-}(X))\right].
\end{aligned}$$

Now, using the Cauchy–Schwartz inequality,

$$\mathbb{E}_{X\sim p}\left[\exp(b_{j+}(X))\right] \leq \mathbb{E}_{X\sim p}\left[\frac{p(X^{j-})^2}{p(X)^2}\right]\mathbb{E}_{X\sim p}\left[\exp(2b_{j-}(X))\right]. \tag{5.16}$$

Existence of the first term in (5.16) is implied by the Standing Assumption $(\nabla^- p)/p \in L^2(p, \mathbb{R}^d)$, while existence of the second term in (5.16) was assumed. Therefore, we have shown that (5.14) and (5.15) exist. Repeating an essentially identical argument, it is straightforward to see also that

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta^s r_{j-}(X^{j+},\theta)\|\right] < \infty \quad \text{and} \quad \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta^s (r_{j-}(X,\theta)^2)\|\right] < \infty \tag{5.17}$$

for $s = 2, 3$ as claimed.

5.6.4. Assumption 8 for Poisson, Ising, and Conway-Maxwell-Poisson Models

Assumption 8 for the Poisson and Ising models used in the experiments can be verified as a special case of exponential family. Any Poisson model can be written in the form

$$p_\theta(x) \propto \exp\left(\log(\theta_1)\, x - \sum_{k=1}^{x}\log(k)\right).$$

This falls into a class of exponential family by setting $\eta(\theta) = \log(\theta)$, $T(x) = x$, and $b(x) = -\sum_{k=1}^{x}\log(k)$. This gives that $T(x-1) - T(x) = -1$ and $b(x-1) - b(x) = \log(x)$. The condition derived in the preceding section is then satisfied provided that $\mathbb{E}_{X\sim p}[\exp(\log(X))^2] = \mathbb{E}_{X\sim p}[X^2] < \infty$, i.e. p has a second moment. Similarly, any Ising model can be written in the form

$$p_\theta(x) \propto \exp(\theta \cdot T(\boldsymbol{x}))$$

where $T : \mathcal{X} \to \mathbb{R}^k$ is a vector of summary statistics that define the model. For Ising models, $\mathcal{X}$ is of finite cardinality and $T(\boldsymbol{x})$ is hence bounded for any $\boldsymbol{x} \in \mathcal{X}$. The conditions are then automatically satisfied.

The Conway-Maxwell-Poisson model falls into a class of exponential family, but it is beyond the simplified case. Nonetheless, Assumption 8 is still verifiable. Recall that the Conway-Maxwell-Poisson model has the form $p_\theta(x) \propto (\theta_1)^x (x!)^{-\theta_2}$ whose ratio function is given by $r_{j-}(x,\theta) = p_\theta(x-1)/p_\theta(x) = x^{\theta_2}/\theta_1$ where $\theta_1, \theta_2 \in (0,\infty)$. The derivative of the ratio with respect to $\theta = (\theta_1, \theta_2)$ is then given by

$$\nabla_\theta r_{j-}(x+1,\theta) = \left(-\frac{(x+1)^{\theta_2}}{\theta_1^2}, \frac{(x+1)^{\theta_2}\log(x+1)}{\theta_1}\right), \quad \nabla_\theta (r_{j-}(x,\theta))^2 = \left(-\frac{x^{2\theta_2}}{\theta_1^3}, \frac{x^{2\theta_2}\log x}{\theta_1^2}\right).$$

Note that the term $x^{2\theta_2}\log x$ in $\nabla_\theta(r_{j-}(x,\theta))^2$ is well-defined even at $x=0$ since it converges to 0 as $x\to 0$ if $\theta_2>0$ despite the individual term $\log x$ alone is not well-defined for $x=0$. Let M_1 and M_2 be the infimum value of θ_1 and the supremum value of θ_2 for (θ_1,θ_2) in the bounded set U to see that

$$\sup_{\theta\in U}\|\nabla_\theta r_{j-}(x+1,\theta)\| = \left|\frac{(x+1)^{M_2}}{M_1^2}\right| + \left|\frac{(x+1)^{M_2}\log(x+1)}{M_1}\right|,$$
$$\sup_{\theta\in U}\|\nabla_\theta(r_{j-}(x,\theta))^2\| = \left|\frac{x^{2M_2}}{M_1^3}\right| + \left|\frac{x^{2M_2}\log x}{M_1^2}\right|.$$

We can derive the same quantity up to constants in the power exponent of each term for the second and third derivative. Then Assumption 8 imposes that expectations of these quantities with respect to the data generating distribution $x\sim p$ are finite. For example, the expectations for the first derivatives are

$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta r_{j-}(X+1,\theta)\|\right] = \frac{1}{M_1^2}\mathbb{E}_{X\sim p}\left[\left|(x+1)^{M_2}\right|\right] + \frac{1}{M_1}\mathbb{E}_{X\sim p}\left[\left|(x+1)^{M_2}\log(x+1)\right|\right],$$
$$\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta(r_{j-}(x,\theta))^2\|\right] = \frac{1}{M_1^3}\mathbb{E}_{X\sim p}\left[\left|x^{2M_2}\right|\right] + \frac{1}{M_1^2}\mathbb{E}_{X\sim p}\left[\left|x^{2M_2}\log x\right|\right],$$

where the boundedness is translated into the moment condition of p as above.

5.6.5. *Proof of Theorem 5*

Proof. Let $r_{j-}(\boldsymbol{x},\theta) := p_\theta(\boldsymbol{x}^{j-})/p_\theta(\boldsymbol{x})$ and $r_{j+}(\boldsymbol{x},\theta) := p_\theta(\boldsymbol{x})/p_\theta(\boldsymbol{x}^{j+})$ to set

$$D_n(\theta) := \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d\Big(r_{j-}(\boldsymbol{x}_i,\theta)\Big)^2 - 2r_{j+}(\boldsymbol{x}_i,\theta).$$

Let $R(\boldsymbol{x}_i,\theta) := \sum_{j=1}^d(r_{j-}(\boldsymbol{x}_i,\theta))^2 - 2r_{j+}(\boldsymbol{x}_i,\theta)$. In what follows, we set $D(\theta) := \mathbb{E}_{X\sim p}[R(X,\theta)]$ and verify that preconditions C1-C5 of Theorem 5 are satisfied. Note that C3 holds directly by Assumption 7 and C5 is also assumed directly in Theorem 5.

C1: By the strong law of large numbers (Durrett, 2010, Theorem 2.5.10),

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^n R(\boldsymbol{x}_i,\theta) \quad\xrightarrow{\text{a.s.}}\quad \mathbb{E}_{X\sim p}[R(X,\theta)] = D(\theta), \tag{5.18}$$

provided that $\mathbb{E}_{X\sim p}[|R(X,\theta)|]<\infty$ for each $\theta\in\Theta$. By the triangle inequality,

$$\begin{aligned}\mathbb{E}_{X\sim p}[|R(X,\theta)|] &= \mathbb{E}_{X\sim p}\left[|R(X,\theta)|\right] + C(p) - C(p)\\ &= \mathbb{E}_{X\sim p}\left[\left|R(X,\theta)+1+\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right|\right] + \mathbb{E}_{X\sim p}\left[1+\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right]\\ &= \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p_\theta(X)}{p_\theta(X)} - \frac{\nabla^- p(X)}{p(X)}\right\|^2\right] + 1 + \mathbb{E}_{X\sim p}\left[\left\|\frac{\nabla^- p(X)}{p(X)}\right\|^2\right]\end{aligned}$$

where the last equality holds from Proposition 6 and both the quantities are finite by Standing Assumption 1. Hence $\mathbb{E}_{X\sim p}[|R(X,\theta)|] < \infty$ and (5.18) holds for each $\theta \in \Theta$.

C2: From Assumption 8, we have that $r_{j+}(\boldsymbol{x},\theta)$ and $r_{j-}(\boldsymbol{x},\theta)$ are three times continuously differentiable with respect to $\theta \in U$ for all $\boldsymbol{x} \in \mathcal{X}$, and thus $D_n(\theta)$ is three times continuously differentiable with respect to $\theta \in U$. For any $s \in \{1,2,3\}$,

$$\nabla_\theta^s D_n(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^s R(\boldsymbol{x}_i,\theta) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^d \nabla_\theta^s (r_{j-}(\boldsymbol{x}_i,\theta)^2) - 2\nabla_\theta^s r_{j+}(\boldsymbol{x}_i,\theta). \tag{5.19}$$

By the triangle inequality, we have an upper bound

$$\sup_{\theta\in U} \|\nabla_\theta^s D_n(\theta)\| \le \frac{1}{n}\sum_{i=1}^n \underbrace{\sum_{j=1}^d \sup_{\theta\in U} \left\|\nabla_\theta^s (r_{j-}(\boldsymbol{x}_i,\theta)^2)\right\| + 2\sup_{\theta\in U}\|\nabla_\theta^s r_{j+}(\boldsymbol{x}_i,\theta)\|}_{=:G(\boldsymbol{x}_i)}.$$

The quantity $\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i)$ is a random variable dependent on $\{\boldsymbol{x}_i\}_{i=1}^n$. By the strong law of large numbers (Durrett, 2010, Theorem 2.5.10),

$$\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) \xrightarrow{\text{a.s.}} \mathbb{E}_{X\sim p}[G(X)] < \infty$$

provided that $\mathbb{E}_{X\sim p}[|G(X)|] < \infty$. Indeed, this condition holds since from positivity of G

$$\mathbb{E}_{X\sim p}[|G(X)|] = \sum_{j=1}^d \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\left\|\nabla_\theta^s (r_{j-}(X,\theta)^2)\right\|\right] + 2\mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|\nabla_\theta^s r_{j+}(X,\theta)\|\right],$$

where the right-hand side is finite by Assumption 8. Then

$$\limsup_{n\to\infty}\sup_{\theta\in U}\|\nabla_\theta^s D_n(\theta)\| \le \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n G(\boldsymbol{x}_i) \overset{\text{a.s.}}{=} \mathbb{E}_{X\sim p}[G(X)] < \infty$$

for any $s \in \{1,2,3\}$, which establishes C2.

C4: Let $h(\boldsymbol{x},\theta) := \nabla_\theta^2 R(\boldsymbol{x},\theta)$. From (5.19), $H_n(\theta) = \frac{1}{n}\sum_{i=1}^n h(\boldsymbol{x}_i,\theta)$. By the strong law of large numbers (Durrett, 2010, Theorem 2.5.10), we have $H_n(\theta) \xrightarrow{\text{a.s.}} \mathbb{E}_{X\sim p}[h(X,\theta)]$ provided that $\mathbb{E}_{X\sim p}[\|h(X,\theta)\|] < \infty$. Indeed, this condition holds for all $\theta \in U$, since we have the upper bound

$$\mathbb{E}_{X\sim p}[\|h(X,\theta_*)\|] \le \mathbb{E}_{X\sim p}\left[\sup_{\theta\in U}\|h(X,\theta)\|\right] \le \mathbb{E}_{X\sim p}\left[|G(X)|\right] < \infty$$

where the right-hand side is bounded by the preceding argument. It remains to verify that $H_* := \lim_{n\to\infty} H_n(\theta_*)$ is equal to $\nabla_\theta^2 \operatorname{DFD}(p_\theta\|p)|_{\theta=\theta_*}$, from which C4 follows since H_* was assumed to be nonsingular in the statement of Theorem 5. By the Lebesgue's dominated

convergence theorem, for each $\theta \in U$,

$$\lim_{n\to\infty} H_n(\theta) = \mathbb{E}_{X\sim p}[\nabla_\theta^2 R(\boldsymbol{x}, \theta)] = \nabla_\theta^2 \mathbb{E}_{X\sim p}[R(\boldsymbol{x}, \theta)] = \nabla_\theta^2 D(\theta).$$

provided that $\mathbb{E}_{X\sim p}[\sup_{\theta\in U} \|\nabla_\theta^2 R(\boldsymbol{x}, \theta)\|] < \infty$. This condition holds for all $\theta \in U$ since $\mathbb{E}_{X\sim p}[\sup_{\theta\in U} \|\nabla_\theta^2 R(\boldsymbol{x}, \theta)\|] \leq \mathbb{E}_{X\sim p}[|G(X)|] < \infty$. Since $\theta_* \in U$ in particular, $H_* = \nabla_\theta^2 D(\theta) \mid_{\theta=\theta_*} = \nabla_\theta^2 \operatorname{DFD}(p_\theta \| p)|_{\theta=\theta_*}$, as claimed.

Thus, preconditions C1-C5 are satisfied and the result follows from Assumption 3. □

5.6.6. Proof of Proposition 9

We first establish that the Stein set $\{h \in \mathcal{H}_k^d : \sum_{i=1}^d \|h_i\|_{\mathcal{H}_k}^2 \leq 1\}$ constructed from $\mathcal{H}_k^d$ is contained in another Stein set $\{h \in L^2(q, \mathbb{R}^d) : \|h\|_{L^2(q,\mathbb{R}^d)}^2 \leq 1\}$ constructed from $L^2(q, \mathbb{R}^d)$ for any domain $\mathcal{X}$, under a standard condition on the reproducing kernel. This in turn shows that DFD dominates KSD.

Proposition 11. *Let q be a probability distribution on $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel such that $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$. Then the unit ball of $\mathcal{H}_k^d$ is contained in the unit ball of $L^2(q, \mathbb{R}^d)$.*

Proof. First, let $f : \mathcal{X} \to \mathbb{R}^d$ be any element of $\mathcal{H}_k^d$, where its i-th output-coordinate $f_i : \mathcal{X} \to \mathbb{R}$ belongs to $\mathcal{H}_k$ each. From the reproducing property of $\mathcal{H}_k$, followed by the Cauchy–Schwartz inequality, the norm of f in $L^2(q, \mathbb{R}^d)$ is upper bounded as follows:

$$\begin{aligned}
\|f\|_{L^2(q,\mathbb{R}^d)}^2 = \sum_{i=1}^d \mathbb{E}_{X\sim q}[f_i(X)^2] &= \sum_{i=1}^d \mathbb{E}_{X\sim q}[\langle f_i(\cdot), k(X, \cdot)\rangle_{\mathcal{H}_k}^2] \\
&\leq \sum_{i=1}^d \mathbb{E}_{X\sim q}\left[\|f_i\|_{\mathcal{H}_k}^2 \|k(X, \cdot)\|_{\mathcal{H}_k}^2\right] = \sum_{i=1}^d \mathbb{E}_{X\sim q}\left[\|f_i\|_{\mathcal{H}_k}^2 k(X, X)\right] \\
&= \left(\sum_{i=1}^d \|f_i\|_{\mathcal{H}_k}^2\right) \mathbb{E}_{X\sim q}\left[k(X, X)\right] = \|f\|_{\mathcal{H}_k^d}^2 \, \mathbb{E}_{X\sim q}\left[k(X, X)\right].
\end{aligned}$$

The continuous embedding of $\mathcal{H}_k^d$ in $L^2(q, \mathbb{R}^d)$ therefore holds, and moreover the embedding constant is at most one, since $\mathbb{E}_{X\sim q}[k(X, X)] \leq 1$ due to the assumption that $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \leq 1$. Therefore, it follows that the unit ball of $\mathcal{H}_k^d$ is contained in the unit ball of $L^2(q, \mathbb{R}^d)$. □

Now we move on to the main proof.

Proof. From the construction of KSD and DFD based on (2.6), it is clear to see that

$$\operatorname{KSD}(p\|q) = \sup_{\|h\|_{\mathcal{H}_k^d}\leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| \leq \sup_{\|h\|_{L^2(q,\mathbb{R}^d)}\leq 1} |\mathbb{E}_{X\sim q}\left[S_p[h](X)\right]| = \operatorname{DFD}(p\|q),$$

where the inequality follows from Proposition 11 showing that $\{\|h\|_{\mathcal{H}_k^d} \leq 1\}$ is a smaller set than $\{\|h\|_{L^2(q,\mathbb{R}^d)} \leq 1\}$, and the final equality follows from Proposition 7. □

Chapter 6. Conclusion

Prior to this thesis, there was little or no existing literature concerning generalised Bayesian inference in the setting of intractable likelihood. Existing approaches to Bayesian inference for intractable likelihood fell into three broad categories: (1) simulation-based methods such as *approximate Bayesian computation* (Beaumont et al., 2002; Frazier, 2020; Marin et al., 2012; Price et al., 2018; Tavaré et al., 1997) and *exchange algorithm* (Murray et al., 2006; Møller et al., 2006), including MCMC-based simulation methods (Park and Haran, 2018b); (2) plugin-based methods such as *pseudo-marginal* MCMC (Andrieu and Roberts, 2009) and *Russian roulette*-based MCMC (Lyne et al., 2015); (3) approximate likelihood methods such as *pseudo-likelihood* (Besag, 1974; Dryden et al., 2002) and *composite likelihood* (Eidsvik et al., 2014), which are of course also applicable beyond the Bayesian context. Both (1) and (2) rely on either the ability to exactly or approximately simulate from the generative model, or the ability to unbiasedly estimate the likelihood, whilst (3) represents a collection of approaches that are tailored to particular statistical models. These algorithms aim to approximate the standard Bayesian posterior, which often do not confer robustness in situations where the model is misspecified.

This thesis established SD-Bayes, a new Bayesian methodology for intractable models based on generalised Bayesian inference and Stein discrepancy. Unlike the existing approaches, the SD-Bayes methodology does not rely on any approximation, estimation, or sampling of intractable models. Instead, it leverages a Stein discrepancy—a statistical divergence for probability distributions that is computable without knowing their normalising constants—as a loss in generalised Bayesian inference. The resulting generalised posterior is independent of normalising constants of intractable models and therefore accessible by any standard MCMC algorithms. In Chapter 3, we established several theoretical underpinnings of generalised Bayesian inference. We derived posterior consistency and the Bernstein–von Mises theorem of generalised posteriors, which guarantees their appealing regularity in the limiting regime of the number of data. Furthermore, we formulated a qualitative criterion, termed global bias-robustness, that implies a strong insensitivity of generalised posteriors to an outlier mixed in data. This criterion rigorously emphasises an advantage of generalised Bayesian inference over standard Bayesian inference in the setting of data containing an outlier. Finally, we discussed a novel calibration algorithm of generalised posteriors through minimisation of the objective based on a Stein discrepancy, which is computationally efficient and numerically stable even when a datum or a parameter is high dimensional.

In Chapter 4, we provided a general formulation of KSD as a particularly useful Stein discrepancy to intractable models in continuous domains. We then proposed KSD-Bayes, the SD-Bayes methodology resulting from the use of KSD. We demonstrated that KSD-Bayes is not only computable by any standard MCMC algorithms, but also admits fully conjugate analysis for exponential family models. KSD-Bayes provides robust generalised Bayesian inference in this context, including a theoretical guarantee of global bias-robustness over Θ. From a theoretical perspective, the soundness of KSD-Bayes, in terms of posterior consistency and asymptotic normality of the generalised posterior, was further established. In Chapter 5, we established DFD, a novel extension of the Fisher divergence to a wide class of discrete domains, and concretise SD-Bayes with DFD for intractable models in discrete domains. The approach, called DFD-Bayes, is distinguished by its suitability for standard MCMC algorithms as with KSD-Bayes, its lack of user-specified hyperparameters such as a choice, and its linear computational cost in the dataset size n per-iteration of MCMC. Furthermore, posterior consistency and asymptotic normality of the generalised posterior was established. In discrete case, DFD-Bayes outperformed KSD-Bayes in our experiments both in terms of inferential performance and computational cost. However, one of the significant advantages of KSD-Bayes is robustness in the presence of outliers contained in a dataset. In discrete case, this is confirmed through an additional experiment on the Ising model in Section B.2. Thus, in settings where robust inference is required, the KSD-Bayes approach may be preferred.

Although SD-Bayes, whose concrete special cases include KSD-Bayes and DFD-Bayes, has a number of appealing features, it is not a panacea for intractable likelihood. As discussed in Sections 4.2.3 and 5.2.2, score-based methodologies in general can suffer from insensitivity to mixture proportions, which limits its applicability to models and datasets that are not "too multi-modal". In addition, KSD-Bayes is not invariant to transformations of data, while DFD-Bayes achieves the invariance in discrete case. DFD-Bayes is not equipped with such strong robustness to model misspecification as KSD-Bayes demonstrated by the virtue of an appropriate choice of kernel. Some of future works could focus on generalising our DFD construction to allow for further robustness as per the diffusion score-matching framework of Barp et al. (2019). This thesis focused on independent and identically distributed data, meaning that, for example, regression models were not considered. Relaxing the independence and identical distribution assumptions represents a natural direction for future work, and a road map is provided by recent research in the score-matching literature (Xu et al., 2022). There also exists various open avenues for future work in more broad contexts of generalised Bayesian inference. Calibration of generalised posteriors in non-asymptotic regime remains an open problem in generalised Bayesian inference, where that in asymptotic regime has been solved by Frazier et al. (2023) selecting a class of loss whose generalised posterior and frequentist counterpart admit the same asymptotic covariance. In non-asymptotic regime, a certain stable algorithm may be required especially when the parameter θ is high-dimensional

relative to the size n of the dataset. Cases where model misspecification can occurs are clearly not limited to the case of outliers considered in the global bias-robustness. There appears to be an urgent need for extensive theory on in which circumstances generalised Bayesian inference should be or should not be used, so that generalised Bayesian inference can be safely applied in practice. While the focus of this thesis was entirely on posteriors of Bayesian inference for complex models, priors and predictive distributions are also fundamental components of Bayesian inference. It is further vital for Bayesian inference for complex models to tackle an effective choice of priors and guarantees appropriate predictive distributions, in which two other works of the author partially addressed in the context of *Bayesian neural networks* and *calibration error metrics* before. All of these are challenging yet exciting directions of future work.

References

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41.

Amari, S. (1997). Information geometry. *Contemporary Mathematics*, 203:81–96.

Amari, S. (2016). *Information Geometry and Its Applications*. Springer.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., and Swan, Y. (2021). Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*.

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

Baraud, Y. and Birgé, L. (2020). Robust Bayes-like estimation: Rho-Bayes estimation. *The Annals of Statistics*, 48(6):3699 – 3720.

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

Basu, A., Shioya, H., and Park, C. (2019). *Statistical Inference: The Minimum Distance Approach*. Chapman and Hall/CRC.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–43.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Benson, A. and Friel, N. (2021). Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, 16(3):905–931.

Berger, J., Moreno, E., Pericchi, L., Bayarri, M., Bernardo, J., Cano, J., Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., Dasgupta, A., Gustafson, P., and and, L. W. (1994). An overview of robust Bayesian analysis. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 3(1):5–124.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian Theory*. John Wiley & Sons.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302.

Bhattacharyya, A. and Atchade, Y. (2019). A Bayesian analysis of large discrete graphical models. *arXiv:1907.01170*.

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

Bochkina, N. A. and Green, P. J. (2014). The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850 – 1878.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Canu, S. and Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720.

Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 10:377–408.

Carmeli, C., De Vito, E., Toigo, A., and Umanità, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61.

Carter, L. L. and Cashwell, E. D. (1975). *Particle-transport simulation with the Monte Carlo method*. Technical Information Center Energy Research and Development Administration.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1011–1021.

Cherief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, pages 1–21.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Chung, F. R. and Graham, F. C. (1997). *Spectral graph theory*. American Mathematical Soc.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A Kernel Test of Goodness of Fit. In *Proceedings of The 33rd International Conference on Machine Learning*.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

Dawid, A. P., Lauritzen, S., and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608.

Dawid, A. P. and Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, 10(2):479–499.

de Heide, R., Kirichenko, A., Grunwald, P., and Mehta, N. (2020). Safe-bayesian generalized linear regression. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108:2623–2633.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 943–970.

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

Dryden, I., Ippoliti, L., and Romagnoli, L. (2002). Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 11(2):370–388.

Duncan, A., NÃ¼sken, N., and Szpruch, L. (2023). On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39.

Durrett, R. (2010). *Probability: Theory and Examples (4th Edition)*. Cambridge University Press.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.

Elçi, E. M., Grimm, J., Ding, L., Nasrawi, A., Garoni, T. M., and Deng, Y. (2018). Lifted worm algorithm for the Ising model. *Physical Review E*, 97(4):042126.

Everitt, R. G. (2012). Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960.

Frazier, D. T. (2020). Robust and efficient approximate Bayesian computation: A minimum distance approach. *arXiv:2006.14126*.

Frazier, D. T., Kohn, R., Drovandi, C., and Gunawan, D. (2023). Reliable bayesian inference in misspecified models.

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607.

Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. (2022). Bayesian inference using synthetic likelihood: Asymptotics and adjustments. *Journal of the American Statistical Association*, 0(0):1–12.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):421–444.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Ghosh, A. and Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68:413–437.

Girolami, M. (2020). Introducing data-centric engineering: An open access journal dedicated to the transformation of engineering design and practice. *Data-Centric Engineering*, 1:e1.

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

Gong, W., Li, Y., and Hernández-Lobato, J. M. (2021). Sliced kernelized Stein discrepancy. In *Proceedings of the 9th International Conference on Learning Representations*.

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

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 226–234.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. *Proceedings of the 34th International Conference on Machine Learning*, 70:1292–1301.

Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic Stein discrepancies. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Grünwald, P. (2011). Safe learning: Bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420.

Grünwald, P. (2012). The safe Bayesian. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pages 169–183.

Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., et al. (2016). Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318.

Hoeffding, W. (1961). The strong law of large numbers for U-statistics. *Institute of Statistics Mimeo Series*, 302.

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, 23(3):556–584.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley.

Huggins, J. H. and Mackey, L. (2018). Random feature Stein discrepancies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1903–1913.

Huggins, J. H. and Miller, J. W. (2020). Robust inference and model criticism using bagged posteriors. *arXiv: 1912.07104*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.

Insua, D. R. and Ruggeri, F. (2000). *Robust Bayesian Analysis*. Springer.

Jewson, J., Smith, J. Q., and Holmes, C. (2018). Principled Bayesian minimum divergence inference. *Entropy*, 20(6):442.

Jiang, X., Li, Q., and Xiao, G. (2021). Bayesian modeling of spatial transcriptomics data via a modified Ising model. *arXiv:2104.13957*.

Kleijn, B. and van der Vaart, A. (2012a). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354 – 381.

Kleijn, B. J. and van der Vaart, A. W. (2012b). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Knoblauch, J., Jewson, J., and Damoulas, T. (2022). An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109.

Lecun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). *A tutorial on energy-based learning*. MIT Press.

Liang, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22(2):1081–1114.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, pages 71–105.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 276–284.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*.

Liu, S., Kanamori, T., Jitkrittum, W., and Chen, Y. (2019). Fisher efficient inference of intractable models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467.

Lyu, S. (2009). Interpretation and generalization of score matching. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, page 359–366.

Ma, Y.-A., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient MCMC. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 2917–2925.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22:1167–1180.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2023). Approximating Bayes in the 21st Century. *Statistical Science*, pages 1 – 26.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Miller, J. W. (2021). Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53.

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125.

Moores, M., Nicholls, G., Pettitt, A., and Mengersen, K. (2020). Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *Bayesian Analysis*, 15(1):1–27.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9:541–548.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Nakagawa, T. and Hashimoto, S. (2020). Robust Bayesian inference via γ-divergence. *Communications in Statistics - Theory and Methods*, 49(2):343–360.

Oates, C. J. (2013). *Bayesian inference for protein signalling networks*. PhD thesis, University of Warwick.

Ollila, E. and Raninen, E. (2019). Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions. *IEEE Transactions on Signal Processing*, 67(10):2707–2719.

Pacchiardi, L. and Dutta, R. (2021). Generalized Bayesian likelihood-free inference using scoring rules estimators. *arXiv:2104.03889*.

Park, J. and Haran, M. (2018a). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.

Park, J. and Haran, M. (2018b). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.

Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press.

Piancastelli, L. S. C., Friel, N., Barreto-Souza, W., and Ombao, H. (2021). Multivariate Conway-Maxwell-Poisson distribution: Sarmanov method and doubly-intractable Bayesian inference. *arXiv:2107.07561*.

Postman, M., Huchra, J., and Geller, M. (1986). Probes of large-scale structure in the corona borealis region. *The Astronomical Journal*, 92:1238–1247.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11.

Propp, J. and Wilson, D. (1998). Coupling from the past: a user's guide. *Microsurveys in Discrete Probability*, 41:181–192.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2021). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. To appear.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2022). Optimal thinning of mcmc output. *Journal of the Royal Statistical Society: Series B (to appear)*.

Robert, C. P. (2007). *The Bayesian Choice*. Springer.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624.

Rousseau, J. (2016). On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3(1):211–231.

Roy, A. and Dunson, D. B. (2020). Nonparametric graphical model for counts. *Journal of Machine Learning Research*, 21(229):1–21.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A. (2015). Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*.

Sutherland, D. J., Strathmann, H., Arbel, M., and Gretton, A. (2018). Efficient and principled score estimation with Nyström kernel exponential families. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 652–660.

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Takenouchi, T. and Kanamori, T. (2017). Statistical inference with unnormalized discrete models and localized homogeneous divergences. *Journal of Machine Learning Research*, 18(1):1804–1829.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wan, Y.-W., Allen, G. I., and Liu, Z. (2015). TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, 32(6):952–954.

Wang, D., Tang, Z., Bajaj, C., and Liu, Q. (2019). Stein variational gradient descent with matrix-valued kernels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Wei, C. and Murray, I. (2017). Markov Chain Truncation for Doubly-Intractable Inference. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:776–784.

Wenliang, L., Sutherland, D. J., Strathmann, H., and Gretton, A. (2019). Learning deep kernels for exponential family densities. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6737–6746.

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

Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144.

Wu, P.-S. and Martin, R. (2020). A comparison of learning rate selection methods in generalized Bayesian inference. *arXiv:2012.11349*.

Xu, J., Scealy, J. L., Wood, A. T. A., and Zou, T. (2022). Generalized score matching for regression. *arXiv:2203.09864*.

Xue, L., Zou, H., and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(115):3813–3847.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. *Proceedings of the 35th International Conference on Machine Learning*, pages 5561–5570.

Yu, M., Kolar, M., and Gupta, V. (2016). Statistical inference for pairwise graphical models using score matching. *Advances in Neural Information Processing Systems*, 29.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., and Briol, F.-X. (2022). Towards healing the blindness of score matching. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Appendix A. Supplementary Material for Chapter 4

This appendix contains additional empirical results and auxiliary theoretical results referred in Chapter 4. Section A.1 investigates the sensitivity of the generalised posterior to the choice of parameters employed in the kernel K. Section A.2 investigates the sampling distribution of β, controlling the scale of the generalised posterior, when estimated using the approach proposed by Lyddon et al. (2019). An extended discussion of the choice of weighting function, M, and the associated trade-off between statistical efficiency and robustness, is contained in Section A.3. A comparison of KSD-Bayes with other generalised Bayesian procedures developed for a *tractable* likelihood is presented in Section A.4. Section A.5 contains the detail of the default selection of β adopted in Section 4.4. Finally, technical lemmas on the derivatives of KSD that were used as intermediate results in the proofs are established in Section A.6.

A.1. Sensitivity to Kernel Parameters

The kernel K that we recommend as a default in Section 4.1.2 has no degrees of freedom to be specified (with the exception of the weighting function M, whose choice is further explored in Section A.3). Nevertheless, it is interesting to ask whether the generalised posterior is sensitive to our recommended choice of kernel. To this end, we considered the family of kernels of the form

$$K(x, x') = \left(1 + \sigma^{-2}\|x - x'\|_2^2\right)^{-\gamma} \times I_d \tag{A.1}$$

where $\sigma > 0$ and $\gamma \in (0, 1)$. Our recommended kernel sets σ equal to a regularised version of the sample standard deviation of the dataset and $\gamma = 1/2$. To investigate how the generalised KSD-Bayes posterior depends on the choice of σ and γ, we re-ran the normal location model experiment from Section 4.4.1 using values $\sigma \in \{0.5, 1, 2\}$ and $\gamma \in \{0.1, 0.5, 0.9\}$. To limit scope, we consider the performance of the robust version of KSD-Bayes from Section 4.4.1, with weight function $M(x) = (1 + x^2)^{-1/2}$, in the case where the contaminant is fixed to $y = 10$ and the proportion of contamination is varied in $\epsilon \in \{0, 0.1, 0.2\}$. Results in Figure A.1 indicate that the generalised posterior is insensitive to σ, with almost identical output for each value of σ considered. The results for $\gamma \in \{0.5, 0.9\}$ were almost identical, but the generalised posterior appeared to be less robust to contamination when $\gamma = 0.1$. These results support the default

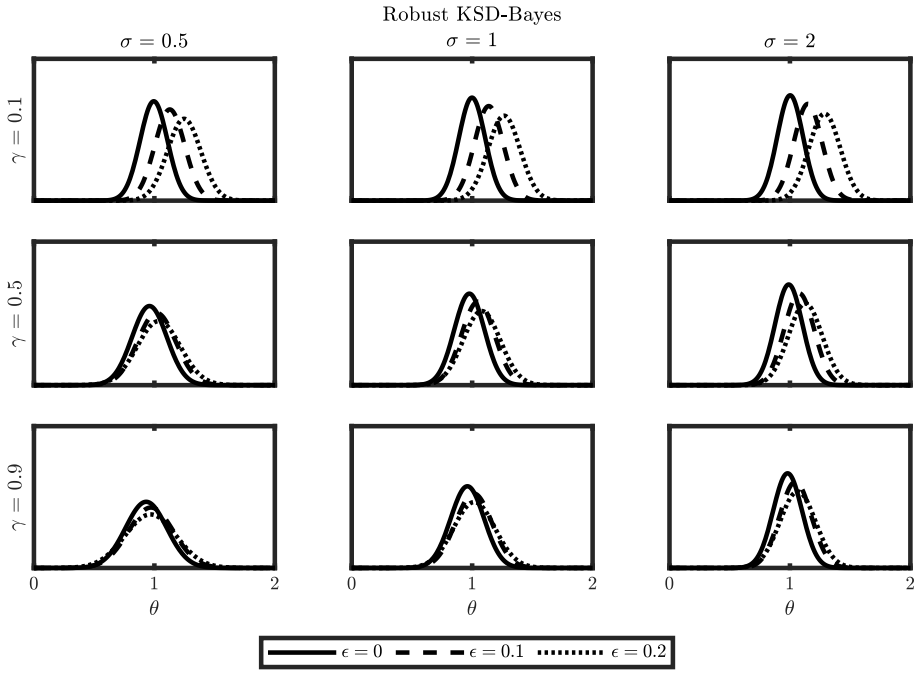


Figure A.1 Sensitivity to kernel parameters: Kernels of the form (A.1), with length-scale parameter σ and exponent γ, are considered in the context of the normal location model in Section 4.4.1. The settings $\sigma \approx 1$, $\gamma = 0.5$ (central panel) were used in the main text. The true parameter value is $\theta = 1$, while a proportion ϵ of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. Here $y = 10$ is fixed and $\epsilon \in \{0, 0.1, 0.2\}$ are considered.

choices recommended in the main text ($\sigma \approx 1$, $\gamma = 0.5$) and provide reassurance that the generalised posterior is not overly sensitive to how these values are specified.

A.2. Sampling Distribution of β

An important component of the KSD-Bayes method is the use of a data-adaptive β. In this appendix the sampling distribution of this data-adaptive β is investigated. Of particular interest are (1) the extent to which β varies at small sample sizes, and (2) how the behaviour of β changes when the data-generating model is misspecified. To investigate, we considered multiple independent realisations of the dataset in the context of the normal location model from Section 4.4.1, collecting the corresponding estimates of β together into box plots, so that the sampling distribution of β can be visualised. To limit scope, we consider the performance of the standard version of KSD-Bayes from Section 4.4.1 (i.e. with weight function $M(x) = 1$), in the case where the contaminant is fixed to $y = 10$ and the proportion of contamination is varied in $\epsilon \in \{0, 0.1, 0.2\}$. The dataset sizes $n \in \{10, 50, 100\}$ were considered. Results in Figure A.2 show that, in the case $\epsilon = 0$ where the model is well-specified, the value $\beta = 1$ is typically selected. This value ensures that the scale of the KSD-Bayes posterior matches that of the standard posterior in this example, so that the approach used to select β can be considered successful. In the misspecified regimes $\epsilon \in \{0.1, 0.2\}$, with small n the estimation of an appropriate weight β is expected to be difficult and indeed the default choice of $\beta = 1$ in (A.7) is automatically

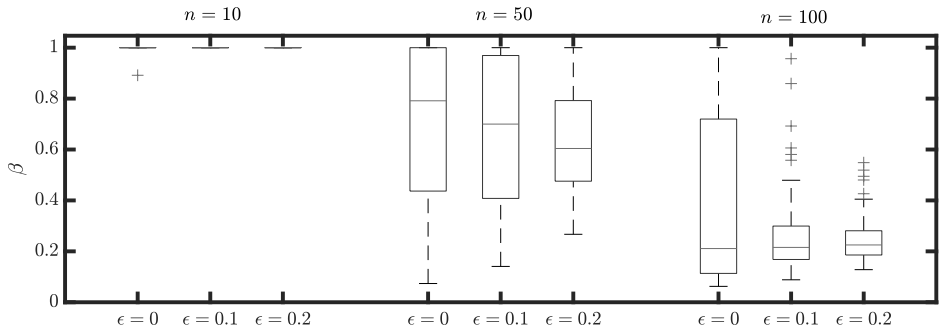


Figure A.2 Sampling distribution of β: Box plots are used to summarise the sampling distribution of β in the context of the normal location model in Section 4.4.1. The sample size n and the contamination proportion ϵ were each varied.

adopted. At larger values of n it is possible to reliably estimate a weight $\beta < 1$ and this weight is seen to be smaller on average when data are more contaminated. These results support our recommended approach to selecting β in (A.7).

A.3. Efficiency/Robustness Trade-Off

There is a well-known trade-off between statistical efficiency and robustness to model misspecification, as exemplified by the data-agnostic statistician who is robust by not learning from data. Minimum distance estimation, which can be considered the frequentist analogue of generalised Bayesian inference, can strike an attractive balance between these competing goals (see e.g. Basu et al., 2019; Lindsay, 1994). In Section 4.3.3 it was demonstrated that global bias-robustness can be achieved using KSD-Bayes through the inclusion of an appropriate weighting function M in the kernel, and in Section 4.4 it was demonstrated that KSD-Bayes can learn from data whilst being bias-robust. However, it remains to investigate the extent to which statistical efficiency is lost in KSD-Bayes, compared to standard Bayesian inference, in the case where the data-generating model is correctly specified. In this appendix, we return to the normal location model of Section 4.4.1 and explore the effect of the choice of weighting function M on the efficiency of the inferences that are produced.

Recall from Theorem 4 that KSD-Bayes is globally bias-robust if there is a function $\gamma : \Theta \to \mathbb{R}$ such that

$$\sup_{y \in \mathbb{R}^d} \Big(\nabla_y \log p_\theta(y) \cdot K(y,y) \nabla_y \log p_\theta(y) \Big) \leq \gamma(\theta) \tag{A.2}$$

where $\sup_{\theta \in \Theta} |\pi(\theta)\gamma(\theta)| < \infty$ and $\int_\Theta \pi(\theta)\gamma(\theta) \mathrm{d}\theta < \infty$. For our recommended kernel K in (4.4), the expression on the left-hand side of (A.2) reduces to

$$\sup_{y \in \mathbb{R}^d} \| M(y)^\top \nabla_y \log p_\theta(y) \|_2^2.$$

For the normal location model in Section 4.4.1 we have $\nabla_y \log p_\theta(y) = \theta - y$ and thus, with our recommended kernel from Equation (4.4), we have

$$\|M(y)^\top \nabla_y \log p_\theta(y)\|_2^2 = (y-\theta)^2 M(y)^2. \tag{A.3}$$

In order that (A.3) is bounded over $y \in \mathbb{R}$ we require $M(y)$ to decay at the rate $\mathcal{O}(|y|^{-1})$ as $|y| \to \infty$. This decay is achieved, for example, by functions of the form

$$M(y) = \left(\frac{a^2}{a^2 + (y-b)^2}\right)^{c/2} \tag{A.4}$$

for any $a \neq 0$, $b \in \mathbb{R}$ and any $c \geq 1$, although of course there are infinitely many other such functions that could be considered. The particular value $c = 1$, which we considered in Section 4.4.1 of the main text and consider here in the sequel, represents the smallest value of c for which (A.3) is bounded over $y \in \mathbb{R}$. For this choice we have that (A.3) is maximised by $y = \theta \pm \sqrt{a^2 + (\theta - b)^2}$ and

$$\sup_{y \in \mathbb{R}^d} (y-\theta)^2 M(y)^2 = \frac{[a^2 + (\theta-b)^2]a^2}{a^2 + [\theta - b \pm \sqrt{a^2 + (\theta-b)^2}]^2} \leq a^2 + (\theta - b)^2 =: \gamma(\theta).$$

For this bound $\gamma(\theta)$, all conditions of Theorem 4 are satisfied. The aim in what follows is to investigate how the performance of KSD-Bayes depends on the specific choices of a and b and in (A.4).

To limit scope, we consider performance in the case where the contaminant is fixed to $y = 10$ and the proportion of contamination is varied in $\epsilon \in \{0, 0.1, 0.2\}$. The dataset sizes were fixed at $n = 100$ as per the main text. Recall from Section 4.4.1 of the main text that the choices $a = 1$, $b = 0$ lead to statistical efficiency comparable to that of standard Bayesian inference. Results in Figure A.3 show that $a = 0.1$ led to almost total robustness to contamination at the expense of inefficient estimation, with the spread of the generalised posterior approximately twice as large as the case where $a = 1$. The setting $a = 10$ causes the generalised posterior to approximate the non-robust KSD-Bayes approach with $M \equiv 1$, as would be expected from inspection of (A.4). The generalised posterior was somewhat insensitive to b, though we note that the choice $b = -5$ conferred additional robustness at the expense of efficiency, while the choice $b = 5$ sacrificed both robustness and efficiency, in both cases relative to $b = 0$. These results broadly support the choices of $a = 1$ and $b = 0$ for this inference problem, as we considered in the main text.

A.4. Comparison with Robust Generalised Bayesian Procedures

This paper presented a generalised Bayesian approach to inference for models that involve an intractable likelihood. However, several generalised Bayesian approaches exist for *tractable* likelihood, and it is interesting to ask how the performance of KSD-Bayes

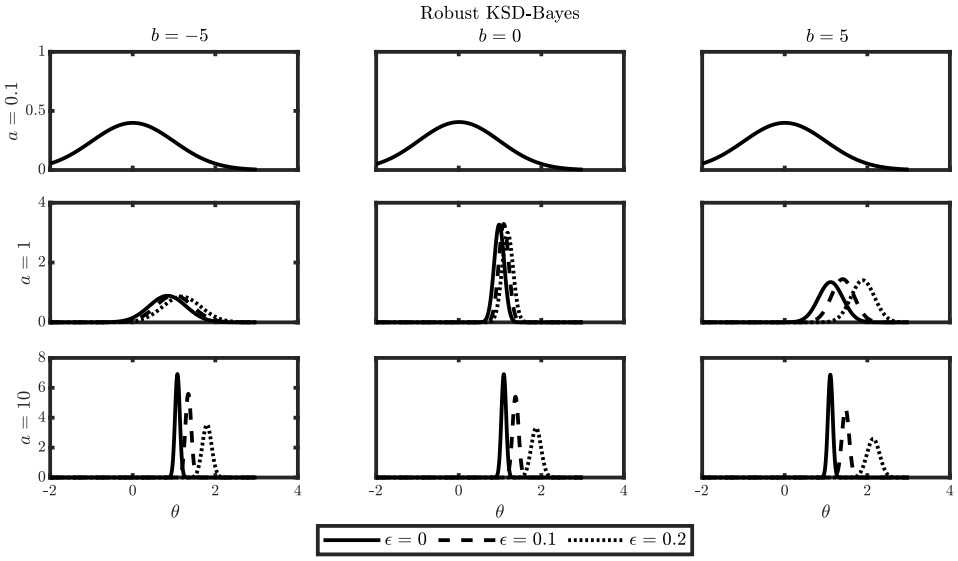


Figure A.3 Efficiency/robustness trade-off: Weight functions of the form (A.4), with length-scale parameter a and location parameter b, are considered in the context of the normal location model in Section 4.4.1. The settings $a = 1$, $b = 0$ (central panel) were used in the main text. The true parameter value is $\theta = 1$, while a proportion ϵ of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. Here $y = 10$ is fixed and $\epsilon \in \{0, 0.1, 0.2\}$ are considered.

compares to these existing approaches in the case of a tractable likelihood. To this end, we return to the normal location model of Section 4.4.1, which has a tractable likelihood, and consider two distinct generalised Bayesian procedures that have been developed in this context; the *power posterior* approach of Holmes and Walker (2017) and the *MMD-Bayes* approach of Cherief-Abdellatif and Alquier (2020). These approaches are representative of two of the main classes of robust statistical methodology; data-adaptive scaling parameters β and minimum discrepancy methods. Both approaches are briefly recalled:

Power Posteriors Motivated by the *coherence* argument of Bissiri et al. (2016), the authors Holmes and Walker (2017) consider a generalised posterior of the form, for some $\beta > 0$,

$$\pi_n(\theta) \propto \pi(\theta) \exp\left\{\beta \sum_{i=1}^{n} \log p_\theta(x_i)\right\},$$

which we call a *power posterior* (e.g. following Friel and Pettitt, 2008). To select an appropriate value for β, with the intention to "allow for Bayesian learning under model misspecification", the authors first introduce the function

$$\Delta(x) = \int_\Theta \pi(\theta) \, \|\partial^1 \log p_\theta(x)\|_2^2 \, \mathrm{d}\theta,$$

where we recall that, in our notation, $\partial^1 = (\partial_{\theta_1}, \ldots, \partial_{\theta_p})$. Then the authors set

$$\beta = \left\{ \frac{\int_{\mathcal{X}} p_{\hat{\theta}_n}(x)\Delta(x)\mathrm{d}x}{\frac{1}{n}\sum_{i=1}^n \Delta(x_i)} \right\}^{\frac{1}{2}}, \tag{A.5}$$

where $\hat{\theta}_n$ is a maximiser of the likelihood. The motivation for (A.5) is quite involved, so we refer the reader to Holmes and Walker (2017) for further background. The authors prove that $\beta \to 1$ in probability when the model is well-specified (Holmes and Walker, 2017, Lemma 2.1), and present empirical evidence of robustness when the model is misspecified.

For the normal location model of Section 4.4.1 we can compute $\partial^1 \log p_\theta(x) = x - \theta$, $\Delta(x) = 1 + x^2$, $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i$, and $\int_{\mathcal{X}} p_{\hat{\theta}_n}(x)\Delta(x)\mathrm{d}x = 2 + (\hat{\theta}_n)^2$, leading to the recommended weight

$$\beta = \left\{ \frac{2 + \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2}{1 + \frac{1}{n}\sum_{i=1}^n x_i^2} \right\}^{\frac{1}{2}}$$

and an associated generalised posterior that is again Gaussian with mean $(\frac{\beta n}{1+\beta n})(\frac{1}{n}\sum_{i=1}^n x_i)$ and variance $\frac{1}{1+\beta n}$.

MMD-Bayes An analogue of KSD-Bayes for tractable likelihood is provided by the MMD-Bayes approach of Cherief-Abdellatif and Alquier (2020), where a *maximum mean discrepancy* (MMD) is employed in place of KSD. In identical notation to that used in **??**, the MMD-Bayes generalised posterior is defined, for some $\beta > 0$, as

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left\{-\beta n \,\mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{P}_n)\right\} \tag{A.6}$$

where, for a given reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the MMD between distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$ is defined as

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}},$$

where the Bochner intergals $\mu_{\mathbb{P}}(\cdot) = \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}\mathbb{P}(x)$ and $\mu_{\mathbb{Q}}(\cdot) = \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}\mathbb{Q}(x)$ are the *kernel mean embeddings* of $\mathbb{P}$ and $\mathbb{Q}$ in $\mathcal{H}$. The authors prove a generalisation bound for MMD-Bayes (Cherief-Abdellatif and Alquier, 2020, Theorem 1), which they interpret as showing "the MMD-Bayes posterior distribution is robust to misspecification". The authors do not recommend a default choice of β in the main text[1], but in private correspondence they recommend $\beta = O(1)$, and we use $\beta = 1$ as a default. The kernel $k(x, y) = \exp(-\|x - y\|_2^2/d)$ was used in our experiment, following Appendix F in Cherief-Abdellatif and Alquier (2020).

[1]Cherief-Abdellatif and Alquier (2020) absorbed the n factor in (A.6) into their definition of β, but for convenience of the reader we have adjusted the presentation of MMD-Bayes to match that used for KSD-Bayes in the main text.

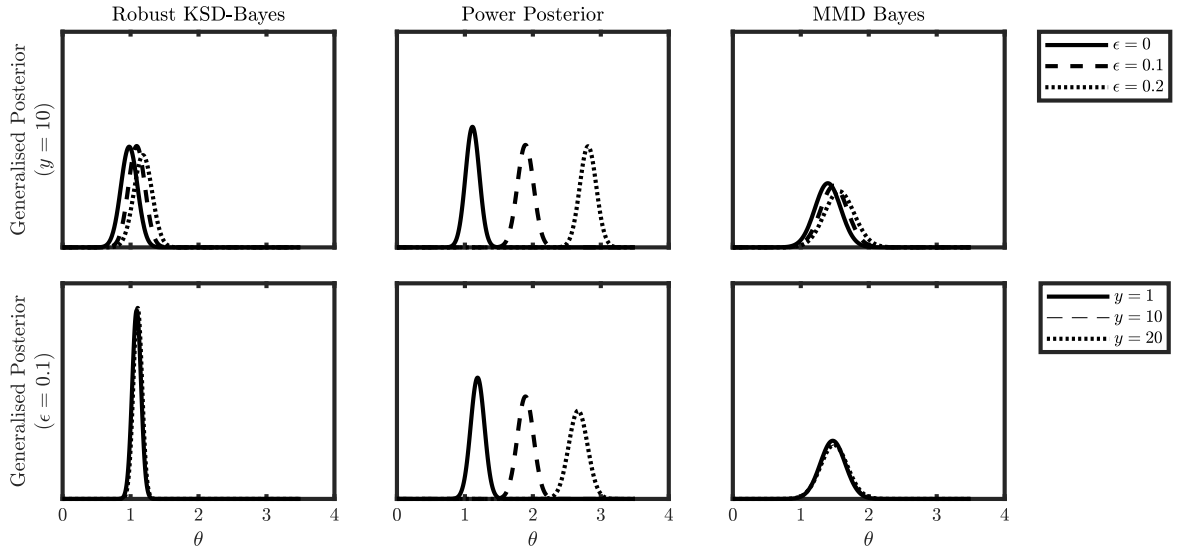


Figure A.4 Comparison with robust generalised Bayesian procedures: Robust KSD-Bayes (this paper), *power posterior* (Holmes and Walker, 2017) and *MMD-Bayes* (Cherief-Abdellatif and Alquier, 2020) approaches are considered in the context of the normal location model in Section 4.4.1. The true parameter value is $\theta = 1$, while a proportion ϵ of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. In the top row $y = 10$ is fixed and $\epsilon \in \{0, 0.1, 0.2\}$ are considered, while in the bottom row $\epsilon = 0.1$ is fixed and $y \in \{1, 10, 20\}$ are considered.

For the normal location model of Section 4.4.1 we can compute the kernel mean embeddings $\mu_{\mathbb{P}_\theta}(x) = \sqrt{\frac{1}{3}}\exp(-\frac{1}{3}(x-\theta)^2)$, $\mu_{\mathbb{P}_n}(x) = \frac{1}{n}\exp(-(x-x_i)^2)$, obtaining an overall expression for the MMD:

$$\mathrm{MMD}(\mathbb{P}_\theta, \mathbb{P}_n)^2 = \frac{1}{3}\exp\left(-\frac{\theta^2}{6}\right) - \frac{2}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{3}}\exp\left(-\frac{(\theta - x_i)^2}{3}\right) + \frac{1}{n^2}\sum_{i,j=1}^{n}\exp\left(-(x_i - x_j)^2\right)$$

The un-normalised density associated with this generalised posterior can be pointwise evaluated; we do this over a fine grid to approximate the normalisation constant in the experiments that we report.

Results The experiment of Section 4.4.1 was conducted using the power posterior and MMD-Bayes methods just described, with results shown in Figure A.4. Power posteriors exhibited similar performance to (non-robust) KSD-Bayes (i.e. with $M \equiv 1$; see Figure 4.2 in the main text), and was therefore less robust to contamination compared with robust KSD-Bayes (i.e. with $M(x) = (1 + x^2)^{-1/2}$). MMD-Bayes generalised posteriors provided similar performance to robust KSD-Bayes in this experiment, albeit exhibiting greater spread. The spread of the MMD-Bayes generalised posterior might be improved if a data-adaptive learning rate β is used, but such an approach was not proposed in Cherief-Abdellatif and Alquier (2020).

A.5. Default Setting for β in Section 4.4

For a simple normal location model, as described in Section 4.4.1, and in a well-specified setting, the asymptotic variance of the KSD-Bayes posterior with $\beta = 1$ is never smaller than that of the standard posterior. This provides a heuristic motivation for the default $\beta = 1$. However, in a misspecified setting smaller values of β are needed to avoid over-confidence in the generalised posterior, taking misspecification into account; see the recent review of Wu and Martin (2020). Here we aim to pick β such that the scale of the asymptotic precision matrix of the generalised posterior (H_*; Proposition 5) matches that of the minimum KSD point estimator ($H_* J_*^{-1} H_*$; Lemma 7), the approach proposed in Lyddon et al. (2019). This ensures the scale of the generalised posterior matches the scale of the sampling distribution of a closely related estimator whose frequentist properties can be analysed when the statistical model is misspecified. Since $\mathbb{P}$ is unknown, estimators of H_* and J_* are required. We propose the following default for β:

$$\beta = \min(1, \beta_n) \qquad \text{where} \qquad \beta_n = \frac{\operatorname{tr}(H_n J_n^{-1} H_n)}{\operatorname{tr}(H_n)}, \tag{A.7}$$

where the matrix H_* is approximated using $H_n := \nabla_\theta^2 \operatorname{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)\big|_{\theta=\theta_n}$, and the matrix J_* is approximated using

$$J_n := \frac{1}{n}\sum_{i=1}^{n} S_n(x_i, \theta_n) S_n(x_i, \theta_n)^\top, \quad S_n(x, \theta) := \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta (\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x_i)).$$

The minimum of $\beta = 1$ and $\beta = \beta_n$ taken in (A.7) provides a safeguard against selecting a value of β that over-shrinks the posterior covariance matrix—a phenomenon that we observed for the experiments reported in Sections 4.4.2 to 4.4.4, due to poor quality of the approximations H_n and J_n when n is small.

A.6. Derivative Bounds

Our auxiliary results here mainly concern moments of derivative quantities, and the aim of this section is to establish the main bounds that will be used. Recall that ∂^1, ∂^2 and ∂^3 denote the partial derivatives $(\partial/\partial\theta_h)$, $(\partial^2/\partial\theta_h\partial\theta_k)$ and $(\partial^3/\partial\theta_h\partial\theta_k\partial\theta_l)$ respectively. For the proofs in this section, we make the index explicit by re-writing them as $\partial^1_{(h)}$, $\partial^2_{(h,k)}$ and

$\partial^3_{(h,k,l)}$. For $x \in \mathcal{X}$ and $(h,k,l) \in \{1,\dots,p\}^3$, we define

$$m^0(x) := \sup_{\theta\in\Theta} \sqrt{\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x)}, \qquad m^1(x) := \sup_{\theta\in\Theta} \sqrt{\sum_{h=1}^{p} (\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})K(x,x)},$$
$$m^2(x) := \sup_{\theta\in\Theta} \sqrt{\sum_{h,k=1}^{p} (\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_\theta})K(x,x)},$$
$$m^3(x) := \sup_{\theta\in\Theta} \sqrt{\sum_{h,k,l=1}^{p} (\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})K(x,x)}.$$

where we continue to use the convention that the first and second operator in expressions such as $(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})K(x,x')$ are respectively applied to the first and second argument of K. Further define

$$M^1(x,x') := m^1(x)m^0(x') + m^0(x)m^1(x'),$$
$$M^2(x,x') := m^2(x)m^0(x') + 2m^1(x)m^1(x') + m^0(x)m^2(x'),$$
$$M^3(x,x') := m^3(x)m^0(x') + 3m^2(x)m^1(x') + 3m^1(x)m^2(x') + m^0(x)m^3(x').$$

Based on these quantities, we now provide three technical results, Lemma 17, Lemma 19 and Lemma 18.

Lemma 17. *Suppose Assumption 5 ($r_{max} = 3$) holds. For each $r = 1,2,3$, and for any $x, x' \in \mathcal{X}$,*

$$\sup_{\theta\in\Theta} \left\| \nabla^r_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \Big) \right\|_2 \le M^r(x,x'). \tag{A.8}$$

If instead Assumption 5 ($r_{max} = 1$) holds, then (A.8) *holds for $r = 1$.*

Proof. We first derive the upper bound for $r = 1$ and then apply the same argument for the remaining upper bound for $r = 2$ and $r = 3$. By the definition of ∇_θ,

$$\sup_{\theta\in\Theta} \left\| \nabla_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \Big) \right\|_2 = \sup_{\theta\in\Theta} \sqrt{\sum_{h=1}^{p} \Big(\partial^1_{(h)} \Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \Big) \Big)^2}. \tag{A.9}$$

By Lemma 9 and Standing Assumption 2, we have $\mathcal{S}_{\mathbb{P}_\theta}K(x,\cdot) \in \mathcal{H}$ for any $x \in \mathcal{X}$ and

$$(*_1) := \partial^1_{(h)} \Big(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \Big) = \partial^1_{(h)} \Big(\langle \mathcal{S}_{\mathbb{P}_\theta}K(x,\cdot), \mathcal{S}_{\mathbb{P}_\theta}K(x',\cdot) \rangle_{\mathcal{H}} \Big). \tag{A.10}$$

From Assumption 5 ($r_{max} = 1$), the operator $(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})$ exists over Θ and satisfies the preconditions of Lemma 9. Hence, by setting $\mathcal{S}_{\mathbb{Q}} = (\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})$ in Lemma 9, we have that $(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_\theta})K(x,\cdot) \in \mathcal{H}$ for each $x \in \mathcal{X}$. Let $f_\theta(\cdot) = \mathcal{S}_{\mathbb{P}_\theta}K(x,\cdot)$ and $g_\theta(\cdot) = \mathcal{S}_{\mathbb{P}_\theta}K(x',\cdot)$. Then the following product rule holds:

$$\partial^1_{(h)}\langle f_\theta, g_\theta \rangle_{\mathcal{H}} = \langle \partial^1_{(h)} f_\theta, g_\theta \rangle_{\mathcal{H}} + \langle f_\theta, \partial^1_{(h)} g_\theta \rangle_{\mathcal{H}}, \tag{A.11}$$

which is verified from definition of differentiation as a limit and continuity of the inner product. Note that $\partial_{(h)} f_\theta(\cdot) = (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot) \in \mathcal{H}$ and $\partial_{(h)} g_\theta(\cdot) = (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x', \cdot) \in \mathcal{H}$. Therefore, by (A.11) and the Cauchy–Schwarz inequality,

$$\begin{aligned}(*_1) &= \left\langle \partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta} K(x, \cdot), \mathcal{S}_{\mathbb{P}_\theta} K(x', \cdot) \right\rangle_{\mathcal{H}} + \left\langle \mathcal{S}_{\mathbb{P}_\theta} K(x, \cdot), \partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta} K(x', \cdot) \right\rangle_{\mathcal{H}} \\ &\leq \underbrace{\left\| (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot) \right\|_{\mathcal{H}}}_{(*_a)} \underbrace{\| \mathcal{S}_{\mathbb{P}_\theta} K(x', \cdot) \|_{\mathcal{H}}}_{(*_b)} + \underbrace{\| \mathcal{S}_{\mathbb{P}_\theta} K(x, \cdot) \|_{\mathcal{H}}}_{(*_c)} \underbrace{\left\| (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x', \cdot) \right\|_{\mathcal{H}}}_{(*_d)}.\end{aligned}$$

For the original term (A.9), by the triangle inequality,

$$\sup_{\theta \in \Theta} \sqrt{\sum_{h=1}^p (*_1)^2} \leq \sup_{\theta \in \Theta} \sqrt{\sum_{h=1}^p \Big((*_a)(*_b) + (*_c)(*_d) \Big)^2} \leq \sup_{\theta \in \Theta} \sqrt{\sum_{h=1}^p (*_a)^2 (*_b)^2} + \sup_{\theta \in \Theta} \sqrt{\sum_{h=1}^p (*_c)^2 (*_d)^2}.$$

For the term $(*_a)$, expanding the norm yields that

$$(*_a)^2 = \left\langle (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot), (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot) \right\rangle_{\mathcal{H}} = (\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta})(\partial^1_{(h)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, x).$$

A similar argument applied to $(*_b)^2$, $(*_c)^2$ and $(*_d)^2$ leads to the overall bound

$$\sup_{\theta \in \Theta} \left\| \nabla_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big) \right\|_2 \leq m^1(x) m^0(x') + m^0(x) m^1(x') = M^1(x, x').$$

The upper bounds for $r = 2$ and $r = 3$ are obtained by an analogous argument. Indeed, from the definition of ∇^2_θ and ∇^3_θ,

$$\begin{aligned}\sup_{\theta \in \Theta} \left\| \nabla^2_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big) \right\|_2 &= \sup_{\theta \in \Theta} \sqrt{\sum_{h,k=1}^p \Big(\partial^2_{(h,k)} \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big) \Big)^2} =: (*''), \\ \sup_{\theta \in \Theta} \left\| \nabla^3_\theta \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big) \right\|_2 &= \sup_{\theta \in \Theta} \sqrt{\sum_{h,k,l=1}^p \Big(\partial^3_{(h,k,l)} \Big(\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') \Big) \Big)^2} =: (*'''),\end{aligned}$$

From Assumption 5 ($r_{max} = 3$), the operators $(\partial^2_{(h,k)} \mathcal{S}_{\mathbb{P}_\theta})$ and $(\partial^3_{(h,k,l)} \mathcal{S}_{\mathbb{P}_\theta})$ exist over Θ and satisfy the preconditions of Lemma 9. Hence, from Lemma 9, $\partial^2_{(h,k)} f_\theta(\cdot) = (\partial^2_{(h,k)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot) \in \mathcal{H}$ and $\partial^3_{(h,k,l)} f_\theta(\cdot) = (\partial^3_{(h,k,l)} \mathcal{S}_{\mathbb{P}_\theta}) K(x, \cdot) \in \mathcal{H}$ for any $x \in \mathcal{X}$, and in turn $\partial^2_{(h,k)} g_\theta(\cdot) \in \mathcal{H}$ and $\partial^3_{(h,k,l)} g_\theta(\cdot) \in \mathcal{H}$. Repeated application of the product rule (A.11) gives that

$$\begin{aligned}\partial^2_{(h,k)} \langle f_\theta, g_\theta \rangle_{\mathcal{H}} =& \langle \partial^2_{(h,k)} f_\theta, g_\theta \rangle_{\mathcal{H}} + \langle \partial^1_{(h)} f_\theta, \partial^1_{(k)} g_\theta \rangle_{\mathcal{H}} + \langle \partial^1_{(k)} f_\theta, \partial^1_{(h)} g_\theta \rangle_{\mathcal{H}} + \langle f_\theta, \partial^2_{(h,k)} g_\theta \rangle_{\mathcal{H}}, \\ \partial^3_{(h,k,l)} \langle f_\theta, g_\theta \rangle_{\mathcal{H}} =& \langle \partial^3_{(h,k,l)} f_\theta, g_\theta \rangle_{\mathcal{H}} + \langle \partial^2_{(h,k)} f_\theta, \partial^1_{(l)} g_\theta \rangle_{\mathcal{H}} + \langle \partial^2_{(h,l)} f_\theta, \partial^1_{(k)} g_\theta \rangle_{\mathcal{H}} + \langle \partial^2_{(k,l)} f_\theta, \partial_{(h)} g_\theta \rangle_{\mathcal{H}} \\ &+ \langle \partial^1_{(h)} f_\theta, \partial^2_{(k,l)} g_\theta \rangle_{\mathcal{H}} + \langle \partial^1_{(k)} f_\theta, \partial^2_{(h,l)} g_\theta \rangle_{\mathcal{H}} + \langle \partial^1_{(l)} f_\theta, \partial^2_{(h,k)} g_\theta \rangle_{\mathcal{H}} + \langle f_\theta, \partial^3_{(h,k,l)} g_\theta \rangle_{\mathcal{H}}.\end{aligned}$$

Following the same argument as the preceding upper bound for $r = 1$, the triangle inequality and Cauchy–Schwarz imply that

$$\begin{aligned}
(*'') &\leq m^2(x)m^0(x') + m^1(x)m^1(x') + m^1(x)m^1(x') + m^0(x)m^2(x') \\
&= m^2(x)m^0(x') + 2m^1(x)m^1(x') + m^0(x)m^2(x') = M^2(x, x'), \\
(*''') &\leq m^3(x)m^0(x') + m^2(x)m^1(x') + m^2(x)m^1(x') + m^2(x)m^1(x') \\
&\quad + m^1(x)m^2(x') + m^1(x)m^2(x') + m^1(x)m^2(x') + m^0(x)m^3(x') \\
&= m^3(x)m^0(x') + 3m^2(x)m^1(x') + 3m^1(x)m^2(x') + m^0(x)m^3(x') = M^3(x, x'),
\end{aligned}$$

which are the claimed upper bounds for the cases $r = 2$ and $r = 3$. □

Lemma 18. *Suppose Assumption 5 ($r_{max} = 3$) holds. For $r = 0, 1, 2, 3$, $\mathbb{E}_{X\sim\mathbb{P}}[|m^r(X)|] < \infty$ and $\mathbb{E}_{X\sim\mathbb{P}}[|m^r(X)|^2] < \infty$. For $r = 1, 2, 3$, $\mathbb{E}_{X,X'\sim\mathbb{P}}[|M^r(X, X')|] < \infty$ and $\mathbb{E}_{X\sim\mathbb{P}}[|M^r(X, X)|] < \infty$. If instead Assumption 5 ($r_{max} = 1$) holds, these results hold for $0 \leq r \leq 1$.*

Proof. First, note that positivity of $m^r(\cdot)$ and $M^r(\cdot)$ implies that the absolute value signs can be neglected. Moreover, from Jensen's inequality $(\mathbb{E}_{X\sim\mathbb{P}}[m^r(X)])^2 \leq \mathbb{E}_{X\sim\mathbb{P}}[m^r(X)^2]$. Thus, it is sufficient to show that (a) $\mathbb{E}_{X\sim\mathbb{P}}[m^r(X)^2] < \infty$, (b) $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')] < \infty$ and (c) $\mathbb{E}_{X\sim\mathbb{P}}[M^r(X, X)] < \infty$.

Part (a): The argument is analogous for each $r = 0, 1, 2, 3$ and we present it with $r = 3$. The bounded follows from Jensen's inequality and the triangle inequality:

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}}\left[m^3(X)^2\right] &\leq \mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\theta\in\Theta}\sum_{h,k,l=1}^{p}(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})K(X, X)\right] \\
&\leq \sum_{h,k,l=1}^{p}\mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\theta\in\Theta}\left((\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_\theta})K(X, X)\right)\right]
\end{aligned}$$

where the terms in the sum are finite by Assumption 5 ($r_{max} = 3$).

Part (b): Since X, X' are independent in the expectation $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')]$, it is clear from the definition of M^r that $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')]$ exists if the expectation of each term $m^s(X)$, $s \leq r$, exists. Thus, by part (a), $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')] < \infty$ for $r = 1, 2, 3$.

Part (c): From the definition of $M^r(x, x)$ for $r = 1, 2, 3$,

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}}[M^1(X, X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)m^0(X)], \\
\mathbb{E}_{X\sim\mathbb{P}}[M^2(X, X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)m^0(X)] + 2\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)m^1(X)], \\
\mathbb{E}_{X\sim\mathbb{P}}[M^3(X, X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^3(X)m^0(X)] + 6\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)m^1(X)].
\end{aligned}$$

Applying the Cauchy Schwartz inequality for each term

$$\mathbb{E}_{X\sim\mathbb{P}}[M^1(X,X)] \le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]},$$
$$\mathbb{E}_{X\sim\mathbb{P}}[M^2(X,X)] \le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]} + 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]},$$
$$\mathbb{E}_{X\sim\mathbb{P}}[M^3(X,X)] \le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^3(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]} + 6\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}.$$

Since each of the latter expectations is finite by part (a), $\mathbb{E}_{X\sim\mathbb{P}}[M^r(X,X)] < \infty$ for $r = 1, 2, 3$.

Inspection of the proof reveals that these results hold for $r = 0, 1$ if instead Assumption 5 ($r_{max} = 1$) holds. □

Lemma 19. *Suppose Assumption 5 ($r_{max} = 3$) holds. Then, for $r = 1, 2, 3$,*

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^r(x_i, x_j) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X,X')] < \infty. \tag{A.12}$$

If instead Assumption 5 ($r_{max} = 1$) holds, then (A.12) *holds for $r = 1$.*

Proof. The proof is based on the strong law of large numbers, the sufficient conditions for which are provided by Lemma 18, which shows that $\mathbb{E}_{X\sim\mathbb{P}}[|m^r(X)|] < \infty$ for $r = 0, 1, 2, 3$ under Assumption 5 ($r_{max} = 3$). Then the strong law of large numbers (Durrett, 2010, Theorem 2.5.10) yields that $(1/n)\sum_{i=1}^{n} m^r(x_i) \xrightarrow{a.s.} \mathbb{E}_{X\sim\mathbb{P}}[m^r(X)] =: (*_r)$ for $r = 0, 1, 2, 3$. Then, from the definition of M^1,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^1(x_i,x_j) &= \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(m^1(x_i)m^0(x_j) + m^0(x_i)m^1(x_j)\Big)\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m^1(x_i) \times \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} m^0(x_j) + \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m^0(x_i) \times \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} m^1(x_j).\end{aligned}$$

Since each limit in the right-hand side converges a.s. to either $(*_0)$ or $(*_1)$, so that

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^1(x_i,x_j) \xrightarrow{a.s.} &\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)] \times \mathbb{E}_{X\sim\mathbb{P}}[m^0(X)] + \mathbb{E}_{X\sim\mathbb{P}}[m^0(X)] \times \mathbb{E}_{X\sim\mathbb{P}}[m^1(X)]\\ &= \mathbb{E}_{X,X'\sim\mathbb{P}}[m^1(X)m^0(X') + m^0(X)m^1(X')] = \mathbb{E}_{X,X'\sim\mathbb{P}}[M^1(X,X')],\end{aligned}$$

where X, X' are independent. An analogous argument holds for $M^2(x_i, x_j)$ and $M^3(x_i, x_j)$, giving that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^2(x_i,x_j) \xrightarrow{a.s.} (*_2)(*_0) + 2(*_1)(*_1) + (*_0)(*_2) = \mathbb{E}_{X,X'\sim\mathbb{P}}[M^2(X,X')],$$
$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^3(x_i,x_j) \xrightarrow{a.s.} (*_3)(*_0) + 3(*_2)(*_1) + 3(*_1)(*_2) + (*_0)(*_3) = \mathbb{E}_{X,X'\sim\mathbb{P}}[M^3(X,X')].$$

Inspection of the proof reveals that (A.12) still holds for $r = 1$ if Assumption 5 ($r_{max} = 1$) holds instead. □

Appendix B. Supplementary Material for Chapter 5

This supplementary material is structured as follows: Illustrative analysis of the DFD and the DFD-Bayes using simple tractable models is presented in Section B.1. Robustness of KSD in discrete case is explored in Section B.2. Full details on our numerical experiments are provided in Section B.3

B.1. Illustrative Analysis with Tractable Models

This section provides illustrative analysis of DFD-Bayes, including comparison with standard Bayesian inference and KSD-Bayes, using simple tractable models. We first demonstrate the calculation of the DFD using the Bernoulli model. We then compare the properties of DFD-Bayes with standard Bayesian inference and KSD-Bayes, using the same Bernoulli model. We next discuss the influence of model misspecification on each posterior using the Poisson model. Finally, we provide an empirical illustration of the limitations of the DFD discussed in Section 5.2.2. The Bernoulli and Poisson models are used for illustration and comparison in this section, since they are tractable and enable standard Bayesian inference to be performed.

B.1.1. The DFD for the Bernoulli Model

For $x \in \{0, 1\}$, the Bernoulli model can be expressed by

$$p_\theta(x) = \theta^x (1-\theta)^{1-x} \tag{B.1}$$

where θ is the probability of $x = 1$. Recall that $p_\theta(1^+) = p_\theta(0)$ and $p_\theta(0^-) = p_\theta(1)$ under our increment/decrement rule. Both the increment and decrement of $p_\theta(1)$ are simply equal to $p_\theta(0)$, and likewise both the increment and decrement of $p_\theta(0)$ are equal to $p_\theta(1)$. Hence, they can be expressed by

$$p_\theta(x^+) = p_\theta(x^-) = \theta^{1-x}(1-\theta)^x, \tag{B.2}$$

that is $p_\theta(x^+) = \theta$ if $x = 0$ and $p_\theta(x^+) = 1 - \theta$ if $x = 1$. Plugging these into equation (5) in the manuscript with $d = 1$ gives an explicit form of the DFD:

$$\mathrm{DFD}(p_\theta \| p_n) \overset{\theta}{=} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\theta^{1-x_i}(1-\theta)^{x_i}}{\theta^{x_i}(1-\theta)^{1-x_i}} \right)^2 - 2 \left(\frac{\theta^{x_i}(1-\theta)^{1-x_i}}{\theta^{1-x_i}(1-\theta)^{x_i}} \right) \tag{B.3}$$

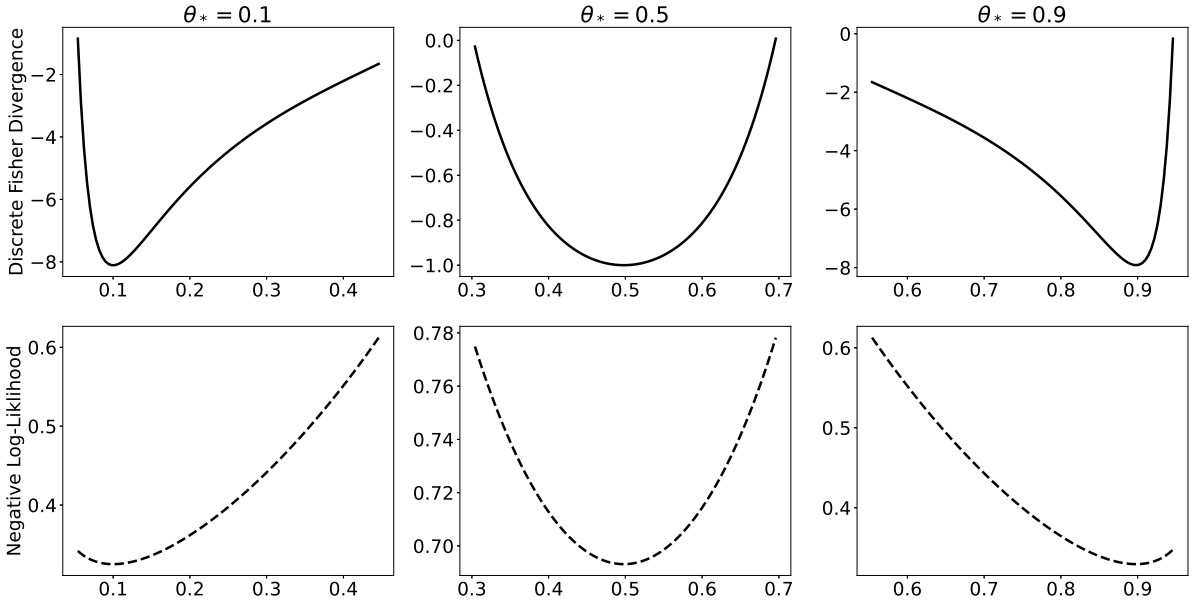


Figure B.1 The DFD (top, solid) and the negative log-likelihood (bottom, dash) between the Bernoulli model and data generated from the Bernoulli model of three different parameters $\theta_* = 0.1$ (left), $\theta_* = 0.5$ (centre), and $\theta_* = 0.9$ (right). They both identify the correct parameter θ_* in each case albeit the different loss surface geometries.

Figure B.1 shows the DFD in (B.3) computed in three cases where 500 random samples are generated from the Bernoulli model with $\theta = 0.1$, $\theta = 0.5$ and $\theta = 0.9$, comparing the loss surface geometry with that of the negative log-likelihood. Both of the losses identify the parameter correctly in each case.

Although the geometrical shape of (B.3) is different from the negative log-likelihood, we can observe in Figure B.1 that the DFD is symmetric under the relabelling $y_i = 1 - x_i$ similarly to the negative log-likelihood in this example. This can indeed be verified as follows. If all data are relabelled, the above formula corresponds to

$$\mathrm{DFD}(p_\theta \| p_n) \overset{\theta}{=} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\theta^{y_i}(1-\theta)^{1-y_i}}{\theta^{1-y_i}(1-\theta)^{y_i}} \right)^2 - 2 \left(\frac{\theta^{1-y_i}(1-\theta)^{y_i}}{\theta^{y_i}(1-\theta)^{1-y_i}} \right). \tag{B.4}$$

With a transform of the parameter $\rho = 1 - \theta$ applied, it further corresponds to

$$\mathrm{DFD}(p_\theta \| p_n) \overset{\theta}{=} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\rho^{1-y_i}(1-\rho)^{y_i}}{\rho^{y_i}(1-\rho)^{1-y_i}} \right)^2 - 2 \left(\frac{\rho^{y_i}(1-\rho)^{1-y_i}}{\rho^{1-y_i}(1-\rho)^{y_i}} \right). \tag{B.5}$$

It is clear from comparison of (B.3) and (B.5) here that the DFD of θ based on the original data x_i is equivalent to that of $\rho = 1 - \theta$ based on the relabelled data $y_i = 1 - x_i$.

B.1.2. Illustrative Comparison of DFD-Bayes with standard Bayes and KSD-Bayes

First, we derive the negative log-likelihood and the KSD for the Bernoulli model. The negative log-likelihood is

$$\mathrm{NLL}(p_\theta \| p_n) = -\frac{1}{n}\sum_{i=1}^{n} x_i \log(\theta) + (1-x_i)\log(1-\theta). \tag{B.6}$$

The KSD in the discrete context was considered in Yang et al. (2018). Letting $\rho_-(\theta, x) := p_\theta(x^-)/p_\theta(x) = \theta^{1-2x}(1-\theta)^{-1+2x}$, the KSD given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is derived as

$$\begin{aligned}\mathrm{KSD}(p_\theta \| p) \overset{\theta}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} &\left(1-\rho_-(\theta,x_i)\right) k(x_i,x_j) \left(1-\rho_-(\theta,x_j)\right) + \\ &\left(1-\rho_-(\theta,x_i)\right)\Big(k(x_i,x_j) - k(x_i,x_j^-)\Big) + \Big(k(x_i,x_j) - k(x_i^-,x_j)\Big)\left(1-\rho_-(\theta,x_j)\right)\Big]. \end{aligned} \tag{B.7}$$

The DFD-Bayes posterior, the standard posterior, and the KSD-Bayes posterior are recovered from generalised posterior (2.2) built upon losses (B.3), (B.6), and (B.7), where β is set to 1 for the standard posterior.

Next, we provide an analytical comparison of the credible regions of each posterior. As discussed in Section 5.2.2, a generalised posterior produces a credible region that differs from that of a standard posterior even in the asymptotic regime. For illustration, we derive the asymptotic variance of each posterior for the Bernoulli model. The asymptotic distribution of each posterior (appropriately centred) follows a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ whose variance σ^2 is the inverse loss-Hessian at the minimiser θ_*. To simplify the derivation, we use the Hamming distance kernel $k(x, x') = \mathbb{1}_{x=x'}$, that is 1 when $x = x'$ and otherwise 0, for the KSD. Let $\rho_+(\theta, x) := p_\theta(x)/p_\theta(x^+) = \theta^{-1+2x}(1-\theta)^{1-2x}$. By routine calculation, the second derivatives of each loss in the limit $n \to \infty$ are

$$\begin{aligned}
\frac{\partial^2}{\partial^2\theta}\mathrm{NLL}(p_\theta\|p) &= \mathbb{E}_{X\sim p}\left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\right], \\
\frac{\partial^2}{\partial^2\theta}\mathrm{DFD}(p_\theta\|p) &= \mathbb{E}_{X\sim p}\left[2\rho_-(\theta,X)\frac{\partial^2}{\partial^2\theta}\rho_-(\theta,X) + 2\left(\frac{\partial}{\partial\theta}\rho_-(\theta,X)\right)^2 - 2\frac{\partial^2}{\partial^2\theta}\rho_+(\theta,X)\right], \\
\frac{\partial^2}{\partial^2\theta}\mathrm{KSD}(p_\theta\|p) &= \mathbb{E}_{X\sim p}\left[2\rho_-(\theta,X)\frac{\partial^2}{\partial^2\theta}\rho_-(\theta,X) + 2\left(\frac{\partial}{\partial\theta}\rho_-(\theta,X)\right)^2 - 2\frac{\partial^2}{\partial^2\theta}\rho_-(\theta,X)\right].
\end{aligned}$$

For the KSD, given that $k(x_1, x_2) - k(x_1, x_2^-)$ and $k(x_1, x_2) - k(x_1^-, x_2)$ are 1 when $x = x'$ and otherwise -1, we simplify the expression as

$$\begin{aligned}\mathrm{KSD}(p_\theta\|p) &\overset{\theta}{=} \mathbb{E}_{X_1,X_2\sim p}\Big[\left(1-\rho_-(\theta,X_1)\right)k(X_1,X_2)\left(1-\rho_-(\theta,X_2)\right)\Big] \\ &\overset{\theta}{=} \mathbb{E}_{X\sim p}\Big[\left(1-\rho_-(\theta,X)\right)^2\Big] \overset{\theta}{=} \mathbb{E}_{X\sim p}\Big[\left(\rho_-(\theta,X)\right)^2 - 2\rho_-(\theta,X)\Big].\end{aligned}$$

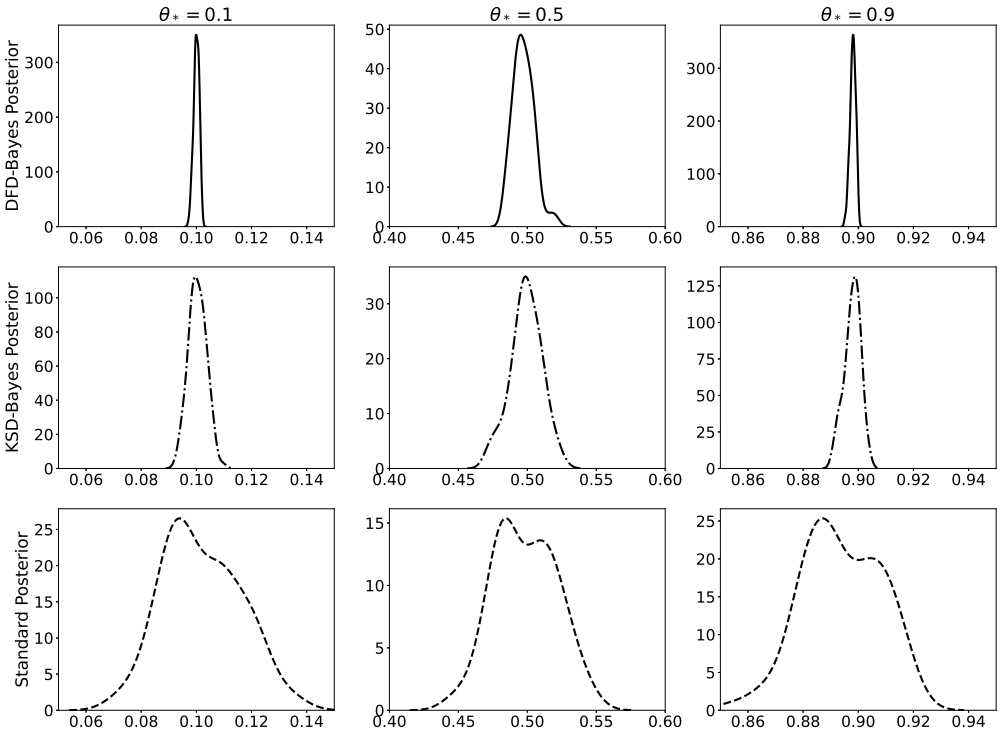


Figure B.2 The DFD-Bayes posterior (top, solid), the KSD-Bayes posterior (middle, dash-dot), and the standard posterior (bottom, dash) computed without β calibrated, for data generated from the Bernoulli model with three different parameters $\theta_* = 0.1$ (left), $\theta_* = 0.5$ (centre), and $\theta_* = 0.9$ (right). While their scales and geometries are different, all methods identify the correct parameter θ_*.

Suppose that the population loss minimiser is $\theta_* = 0.5$, meaning that the data-generating distribution p is the Bernoulli model with $\theta_* = 0.5$. We then have $\rho_-(\theta_*, x) = 1$, $\frac{\partial}{\partial\theta}\rho_-(\theta_*, x) = 2^2(1-2x)$, $\frac{\partial^2}{\partial^2\theta}\rho_-(\theta_*, x) = 2^4(1-2x)^2$, and $\frac{\partial^2}{\partial^2\theta}\rho_+(\theta_*, x) = -2^4(1-2x)^2$. These gives us that

$$
\begin{aligned}
(\partial^2/\partial^2\theta)\mathrm{NLL}(p_\theta \| p)|_{\theta=\theta_*} &= \mathbb{E}_{X\sim p}\left[2^2 \times (X + 1 - X)\right] = 4, \\
(\partial^2/\partial^2\theta)\mathrm{DFD}(p_\theta \| p)|_{\theta=\theta_*} &= \mathbb{E}_{X\sim p}\left[3 \times 2^5 \times (1-2X)^4\right] = 96, \\
(\partial^2/\partial^2\theta)\mathrm{KSD}(p_\theta \| p)|_{\theta=\theta_*} &= \mathbb{E}_{X\sim p}\left[2 \times 2^4 \times (1-2X)^2\right] = 32.
\end{aligned}
$$

By taking the inverse, the asymptotic variance σ^2 for the standard Bayes, the DFD-Bayes, and the KSD-Bayes is each given by 1/4, 1/96, and 1/32. In this example, the above calculation suggests that the DFD-Bayes has the narrowest credible region. The difference in these values emphasises the importance of calibrating β, which we do for all of our experiments in the manuscript.

Finally, we empirically demonstrate the difference between the posteriors and the influence of β. We computed each posterior in cases where (i) β is *not* calibrated i.e. $\beta = 1$

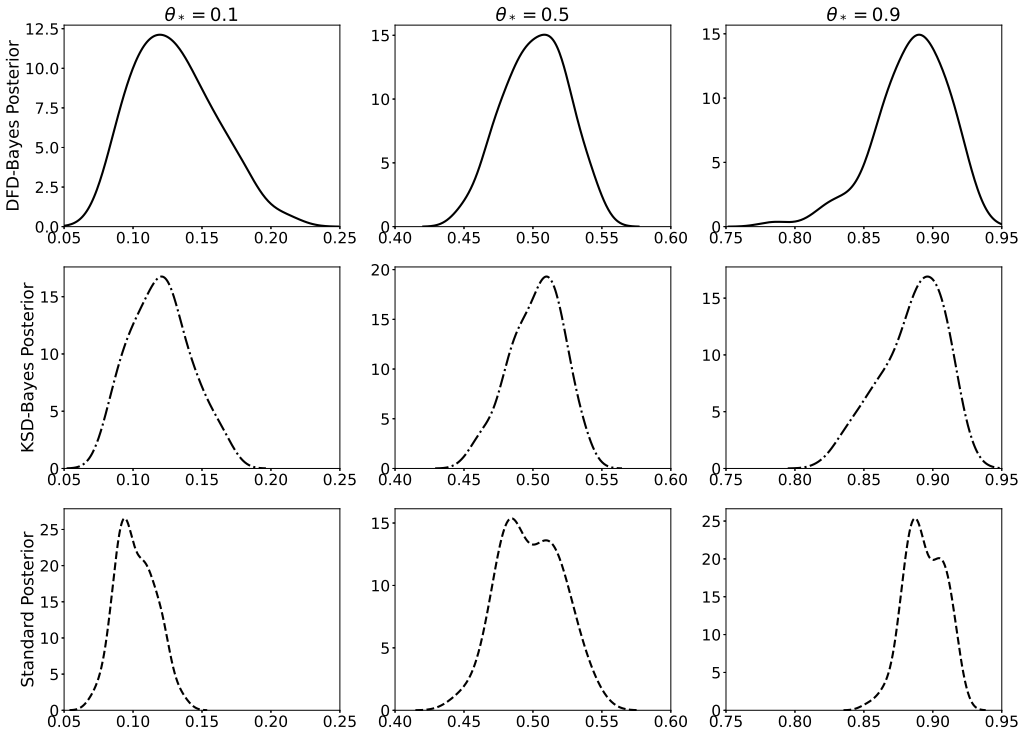


Figure B.3 The DFD-Bayes posterior (top, solid), the KSD-Bayes posterior (middle, dash-dot), and the standard posterior (bottom, dash) computed with β calibrated, for data generated from the Bernoulli model with three different parameters $\theta_* = 0.1$ (left), $\theta_* = 0.5$ (centre), and $\theta_* = 0.9$ (right). While their scales and geometries are different, all methods identify the correct parameter θ_*.

and (ii) β *is* calibrated (except for the standard posterior, which has $\beta = 1$). A Metropolis–Hastings algorithm was adopted to sample from all the posteriors. A Gaussian random walk proposal with covariance $\sigma^2 = 0.01$ was used. In total, 100 samples were obtained from each posterior by thinning 2,000 samples, after an initial burn-in of length 2,000. Figure B.2 shows each posterior computed without β calibrated. It confirms that, without calibration of β, the DFD-Bayes posterior has the narrowest credible region, which agrees with the analytical illustration provided above. Figure B.3 shows each posterior computed with β calibrated, where the result for the standard posterior is identical to Figure B.2 as $\beta = 1$. For the DFD-Bayes and the KSD-Bayes, calibration of β was performed by our proposal in Section 3.4, where we used 100 bootstrap minimisers to compute the analytical solution of β_* in (3.10). It demonstrates that calibration of β prevents over-concentration of the DFD-Bayes and the KSD-Bayes.

B.1.3. Influence of Model Misspecification

Next, we turn our attention to the influence of model misspecification on each method. It is convenient to consider the Poisson model to introduce a synthetic model misspecification.

For $x \in \mathbb{N}_0$, the Poisson model is

$$p_\theta(x) = \frac{\theta^x \exp(\theta)}{x!}. \tag{B.8}$$

Then, the negative log-likelihood and the DFD are

$$\mathrm{NLL}(p_\theta \| p_n) \overset{\theta}{=} \frac{1}{n}\sum_{i=1}^{n} -x_i \log(\theta) + \theta, \tag{B.9}$$

$$\mathrm{DFD}(p_\theta \| p_n) \overset{\theta}{=} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i}{\theta}\right)^2 - 2\frac{x_i + 1}{\theta}. \tag{B.10}$$

Letting $\rho_-(\theta, x) := p_\theta(x^-)/p_\theta(x) = x_i/\theta$, the KSD is

$$\begin{aligned}\mathrm{KSD}(p_\theta \| p) \overset{\theta}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} &(1-\rho_-(\theta,x_i))\,k(x_i,x_j)\,(1-\rho_-(\theta,x_j)) + \\ &(1-\rho_-(\theta,x_i))\left(k(x_i,x_j) - k(x_i,x_j^-)\right) + \left(k(x_i,x_j) - k(x_i^-,x_j)\right)(1-\rho_-(\theta,x_j)).\end{aligned} \tag{B.11}$$

For the KSD, we use a similar choice of kernel to Section 4.1.2, that induces a robustness suitable for this example: $k(x, x') = m(x)\exp(-\mathbb{1}_{x=x'})m(x')$ where $m(x) = \sigma(15 - x)$ based on a sigmoid function $\sigma(t) = (1 + \exp(-t))^{-1}$.

For illustration, we synthetically introduce model misspecification by mixing outliers into the data. We sampled 500 data points $\{x_i\}_{i=1}^n$ from the Poisson model with the parameter $\theta_* = 5$, and replaced the $100 \times \epsilon$ percent of data with an outlier $y = 20$ that is larger than the 99.9% percentile of the Poisson distribution of $\theta_* = 5$. This causes a synthetic model misspecification because the dataset is generated from a mixture of the Possion model and the Dirac distribution at $y = 20$, which cannot be adequately explained by only the Poisson model. The sensitivity of each posterior to the outlier can be analytically investigated. The standard Bayesian posterior is modestly impacted by the outlier y, given that the negative log-likelihood (B.9) is a linear function of each datum x_i. On the other hand, in this example, DFD-Bayes may be more severely impacted, given the DFD (B.10) is a quadratic function of each datum x_i. The growth rate of the KSD with respect to each datum x_i is determined by the choice of kernel k. We compute each posterior for two cases when $\epsilon = 0.0$ (no outlier contained) and $\epsilon = 0.1$ (10% outliers contained), to empirically demonstrate the impact of the model misspecification. The Metropolis–Hastings algorithm with the Gaussian random walk proposal of $\sigma^2 = 0.1$ is used to sample from each posterior with calibration applied. In total, 100 samples were obtained from each posterior by thinning 2,000 samples, after an initial burn-in of length 2,000.

Figure B.4 demonstrates the sensitivity of the standard Bayesian posterior and DFD-Bayes to the outliers, whlie KSD-Bayes shows insensitivity due to the careful choice of kernel. See also Section B.2 for more discussion on robustness of KSD-Bayes. In this example, the sensitivity of the DFD-Bayes to the outlier was higher than the standard

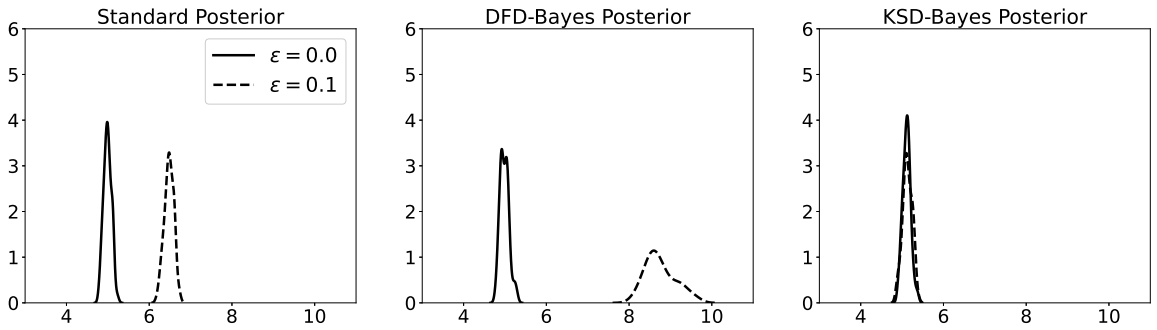


Figure B.4 The standard posterior (left), The DFD-Bayes posterior (centre), and the KSD-Bayes posterior (right) computed with β calibrated for data when $\epsilon = 0.0$ (solid line) and $\epsilon = 0.1$ (dash line), that is, the 10% of data is replaced with outlier y.

Bayesian posterior, as anticipated. Barp et al. (2019) proposed a robust analogue of the Fisher divergence in the continuous case. Although this is not a focus of this work, a similar approach may be applied to the discrete case when severe model misspecification is anticipated. This would be an interesting avenue for further work, but our present interest is in computation for discrete intractable likelihood.

B.1.4. Limitation of DFD-Bayes for Inference of Mixture Parameters

Finally, we provide an empirical illustration of the limitation of score-based methods in Section 5.2.2. It has been pointed out that score-based methods generally exhibit insensitivity to mixing proportions when mixture components have isolated high-probability regions (Wenliang and Kanagawa, 2021; Zhang et al., 2022). In the continuous case, this can be observed using a mixture model of two Gaussian distributions $\mathbb{P}_\theta(x) = (1-\theta) \times \mathcal{N}(-\mu, 1) + \theta \times \mathcal{N}(\mu, 1)$ whose parameter is the mixture ratio. Zhang et al. (2022) illustrated how the Fisher divergence is approximately constant over Θ if μ is large enough to isolate the components $\mathcal{N}(-\mu, 1)$ and $\mathcal{N}(\mu, 1)$. We illustrate the same limitation for the DFD using a mixture model of two Poisson distributions $p_\theta(x) = (1-\theta) \times q_{\lambda_1}(x) + \theta \times q_{\lambda_2}(x)$, where q_{λ_1} and q_{λ_2} are the Poisson distributions with rate parameters $\lambda_1 > 0$ and $\lambda_2 > 0$. Figure B.5 shows the geometry of the DFD between the mixture model p_θ and data generated from the mixture model p_{θ_*} with the true mixture proportion θ_*, for two cases when the supports of the two Poisson distributions are highly isolated and when they are not isolated. The correct mixture proportion θ_* was identified only in the latter case, while in the former case the DFD was approximately constant. See Zhang et al. (2022) for a potential approach to remedy this general limitation of score-based methods.

B.2. Robustness of the KSD in Discrete Case

Section 5.3.2 indicates statistical efficiency of the DFD over the KSD. If one's model is well-specified, minimising the DFD leads us to a correct model faster than the KSD. However, this does not mean that the use of the DFD is always better than the KSD. In

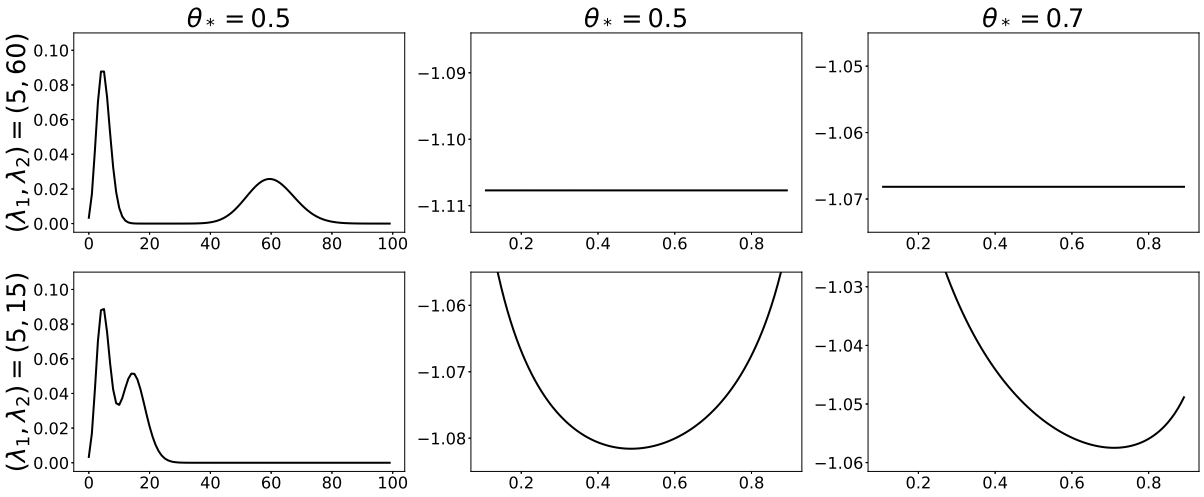


Figure B.5 The form of the Poisson mixture model p_{θ_*} when $\theta_* = 0.5$ (left), the DFD computed for data generated from the model p_{θ_*} with $\theta_* = 0.5$ (middle), and the DFD computed for data generated from the model p_{θ_*} with $\theta_* = 0.7$ (right), for two cases where $\lambda_1 = 5, \lambda_2 = 60$ (top) and $\lambda_1 = 5, \lambda_2 = 15$ (bottom).

particular, the KSD can be equipped with strong robustness by choosing an appropriate kernel. To demonstrate this, we compare three posteriors of Pseudo-Bayes, DFD-Bayes, and KSD-Bayes for the same Ising model as Section 5.4.2 with $d = 100$ ($m = 10$) in a setting where a dataset contains extreme outliers with a proportion ϵ.

We approximately draw 1000 samples $\{x_i\}_{i=1}^{1000}$ from the Ising model p_θ with $\theta = 5$ by the same Metropolis–Hastings algorithm as Section 5.4.2. To study the robustness of the posteriors, we replaced a proportion $\epsilon = 0.1$ of the data with the vector $(1, 1, \cdots, 1)$ corresponding to the extreme value in $\mathcal{X}$ that is rarely drawn from the model. Section 4.3.3 showed that KSD-Bayes can satisfy strong qualitative robustness called "global bias-robustness" by choosing a kernel appropriately. For this example, we use the same choice of kernel as Section 4.1.2 below:

$$k(\boldsymbol{x}, \boldsymbol{x}') = m(\boldsymbol{x}) \exp\left(-\frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(x_i - x_i')\right) m(\boldsymbol{x})$$

where $m(x) = \sigma(90 - |\sum_i x_i|)$ based on a sigmoid function $\sigma(t) = (1 + \exp(-t))^{-1}$. This is indeed a proper choice of kernel, and the function $m(\boldsymbol{x})$ in the definition of kernel is designed to restrict the influence of extreme data whose norm is closer to or larger than 90.

In Figure B.6 demonstrated that KSD-Bayes offered a correct inference outcome even when the dataset contains outliers, being less affected by the outliers. On the other hand, the Pseudo-Bayes and DFD-Bayes posteriors placed the majority of the probability mass on smaller θ than the correct value $\theta = 5$. The extreme value $(1, 1, \cdots, 1)$ of the outliers is more likely to be drawn from the model of $\theta \ll 1$; the posteriors of Pseudo-Bayes and DFD-Bayes were thus pulled in the direction of smaller θ.

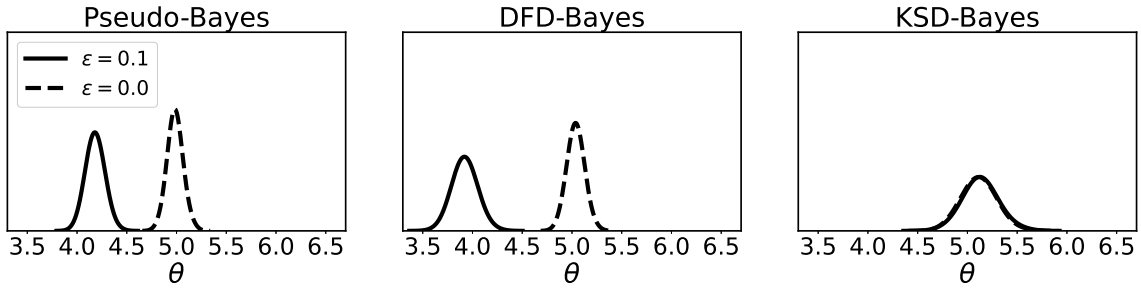


Figure B.6 Posteriors of Pseudo-Bayes (left), DFD-Bayes (centre), and KSD-Bayes (right) for the Ising model in the presence of outlier with $\epsilon = 0.1$ and no outlier with $\epsilon = 0.0$.

B.3. Details of Experimental Assessment

Finally, this section summarises deferred details of the reported experiments in Chapter 5.

B.3.1. Settings for KSD-Bayes in Section 5.4.1

KSD-Bayes is a generalised posterior constructed by taking a KSD as a loss function; see Chapter 4. The approach requires us to specify a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, based on which the KSD is constructed. In these experiments, we adopted a kernel recommended by Yang et al. (2018) for the KSD in discrete domains $\mathcal{X}$ given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(x_i = x_i')\right)$$

where $\mathbb{1}$ is an indicator function, taking values in $\{0, 1\}$. The effect of kernel choice is difficult to predict in the discrete context; for example, Yang et al. (2018) found that the closely related kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d} \mathbb{1}(x_i = x_i')$, can perform poorly in moderate-to-high dimensions d when employed in a Stein discrepancy. General principles for kernel choice in the discrete setting have not yet been established. Thus, one of the advantages of DFD-Bayes is the absence of any user-specified parameters of the method.

B.3.2. Markov Chain Monte Carlo in Section 5.4.1

A Metropolis–Hasting algorithm was employed to sample from the standard Bayesian posterior, as well as KSD-Bayes and DFD-Bayes. For computational convenience, the parametrisation $\tilde{\theta}_1 = \log(\theta_1)$ and $\tilde{\theta}_2 = \log(\theta_2)$ was applied so that parameters are defined on an unbounded domain $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) \in \mathbb{R}^2$. An isotropic Gaussian random walk proposal with covariance $\sigma^2 I$ was employed, with $\sigma = 0.1$ used for all experiments. The convergence of the Markov chain was diagnosed using univariate Gelman–Rubin statistics for each θ_1 and θ_2 computed from 10 independent chains. In total, 500 samples were obtained from each chain by thinning 5,000 samples, all after an initial burn-in of length 5,000. In all cases, the univariate Gelman–Rubin statistics were below 1.02, respectively, for θ_1 and θ_2.

B.3.3. Sales Dataset of Shmueli et al. (2005) in Section 5.4.1

This dataset consists of quarterly sales figures for a particular item of clothing, taken across the different stores of a large national retailer. The original dataset is publicly available at https://www.stat.cmu.edu/COM-Poisson/Sales-data.html; see Shmueli et al. (2005). Quarterly sales at each store can be small and result in a large proportion of 0 entries in the dataset, so that the Conway–Maxwell–Poisson model has a clear advantage against the standard Poisson model. To obtain a maximum *a posteriori* estimate for the parameters of the Conway–Maxwell–Poisson model for this sales dataset, Shmueli et al. (2005) considered a prior π defined by

$$\pi(\theta) \propto \theta_1^{a-1} \exp(-b\theta_2) \left(\sum_{j=1}^{\infty} \theta_1^j / (j!)^{\theta_2} \right)^{-c} \kappa(a, b, c) \tag{B.12}$$

where (a, b, c) is the hyperparameter and $\kappa(a, b, c)$ is the normalising constant of π. The motivation to use this prior is conjugacy, since the resulting posterior takes the same form as (B.12). However, the prior itself contains the intractable terms $(\sum_{j=1}^{\infty} \theta_1^j/(j!)^{\theta_2})^{-c}$ and $\kappa(a, b, c)$. To avoid this additional intractability, which is not a focus of the present work, we considered a simpler chi-squared prior distribution in the main text.

B.3.4. Simulating Data from the Ising Model in Section 5.4.2

Samples from the Ising model were obtained using the same Metropolis–Hasting algorithm used in Yang et al. (2018). First, all coordinates x_i of $\boldsymbol{x}$ were randomly initialised to either -1 or 1 with equiprobability $1/2$. Then, at each iteration, we randomly select one coordinate x_i of $\boldsymbol{x}$ and flip the value of x_i either from -1 to 1 or from 1 to -1, where the flipped value $\tilde{x}_i$ is accepted with probability $\min(1, \exp(-2\tilde{x}_i \sum_{j \in \mathcal{N}_i} x_j/\theta))$ and otherwise rejected. For the experiments in this paper, we ran $n = 1,000$ chains in parallel, in each case taking the final state at iteration $100,000$. This algorithm was used due to its implementational simplicity, rather than its efficiency, and we note that more sophisticated Markov chain Monte Carlo algorithms are available (e.g. Elçi et al., 2018).

B.3.5. Settings for KSD-Bayes in Section 5.4.2

The same choice of kernel as Section B.3.1 is used.

B.3.6. Markov Chain Monte Carlo in Section 5.4.2

The same Metropolis–Hasting algorithm as Section B.3.2 was used, in this case in dimension $p = 1$ with proposal standard deviation $\sigma = 0.1$. The convergence of the Markov chain was again diagnosed using univariate Gelman–Rubin statistics computed from 10 independent chains. In total, 100 samples were obtained after thinning from 2000 samples, with an

initial burn-in of length 2000. In all cases, the univariate Gelman–Rubin statistics were below 1.002.

B.3.7. Description of the Dataset in Section 5.4.3

The original data were gathered by the Cancer Genome Atlas Program, run by the National Cancer Institute in the United States, who have built large-scale genomic profiles of cancer patients with the aim to discover the genetic substructures of cancer (Wan et al., 2015). It contains molecular profiles of biological samples of more than 30 cancer types, e.g. measured via RNA sequencing technology. The raw data were pre-processed using the TCGA2STAT software developed by Wan et al. (2015). Inouye et al. (2017) studied a subset of these data relevant to breast cancer, consisting of a total count of each gene profile found in biological samples. They applied a "log-count" transform, a common preprocessing technique for RNA sequencing data, for every datum, that is a floor function of a log transformed value of the datum. Gene profiles were then sorted by variance of the counts in descending order, with the top 10 gene profiles constituting the final dataset. The preprocessed data studied in Inouye et al. (2017) can be found in https://github.com/davidinouye/sqr-graphical-models.

B.3.8. Markov Chain Monte Carlo in Section 5.4.3

The Metropolis-Hasting Markov Chain Monte Carlo was applied for this experiment. The detail for the Conway–Maxwell–Poisson graphical model is described first as the Poisson graphical model is the special case. For computational convenience, we work with the square of the interaction and dispersion parameters, i.e. $\tilde{\theta}_{i,j} := \theta_{i,j}^2$ and $\tilde{\theta}_{0,i} = \theta_{0,i}$, which modify the model as

$$p_\theta(\boldsymbol{x}) \propto \exp\left(\sum_{i=1}^{d} \theta_i x_i - \sum_{i=1}^{d} \sum_{j \in \mathcal{M}_i} \tilde{\theta}_{i,j}^2 x_i x_j - \sum_{i=1}^{d} \tilde{\theta}_{0,i}^2 \log(x_i!)\right)$$

The domain of each original parameter $\theta_{i,j}$ and $\theta_{0,j}$ is $[0, \infty)$. With this modification, $\tilde{\theta}_{i,j}$ and $\tilde{\theta}_{0,i}$ can be extended to $\mathbb{R}$, making the model $p_\theta(\boldsymbol{x})$ differentiable with respect to $\theta \in \mathbb{R}^p$. The derivatives of the corresponding DFD-Bayes posterior is then available to implement an efficient gradient-based Markov chain Monte Carlo method. We place a standard normal distribution as a prior on each θ_i, a normal distribution with mean 0 and scale $(d(d-1)/2)^{-1}$ as a prior on each $\tilde{\theta}_{i,j}$, and a standard normal distribution as a prior on each $\tilde{\theta}_{0,i}$, that corresponds to the original priors of each θ_i, $\theta_{i,j}$, and $\theta_{0,j}$. The small scale of the half normal distribution prior on $\tilde{\theta}_{i,j}$ was chosen to suppress rapid increase of the quadratic term $x_i x_j$ as opposed to the linear term x_i in the first summation. After the Markov chain finished, the absolute value was taken for the sampled values of $\tilde{\theta}_{i,j}$ and $\tilde{\theta}_{0,i}$ to convert them to the original parameters $\theta_{i,j}$ and $\theta_{0,j}$. The same setting is applied for the Poisson graphical model by fixing the dispersion parameter $\tilde{\theta}_{0,i} = \theta_{0,i} = 1$.

A No-U-Turn Sampler was used to approximate the DFD-Bayes posterior of both the models. In total, 100 points were obtained thinning from $5,000$ samples, with an initial burn-in of length $5,000$. The posterior predictive of each model $p_\theta(\boldsymbol{x})$ was computed by generating $500,000$ samples from $p_\theta(\boldsymbol{x})$ at every θ sampled from the DFD-Bayes posterior. Each $500,000$ predictive samples were thinned to 878 points to make it comparable with the original data of $n = 878$. The number of bootstrap minimisers B used to calibrate β for this experiment was $B = 100$.