Statistical Inference for Generative Models
with Maximum Mean Discrepancy

François-Xavier Briol[1,3], Alessandro Barp[2,3], Andrew B. Duncan[2,3], Mark Girolami[1,3]

[1]*University of Cambridge, Department of Engineering*
[2]*Imperial College London, Department of Mathematics*
[3]*The Alan Turing Institute*

June 17, 2019

Abstract

While likelihood-based inference and its variants provide a statistically efficient and widely applicable approach to parametric inference, their application to models involving intractable likelihoods poses challenges. In this work, we study a class of minimum distance estimators for intractable generative models, that is, statistical models for which the likelihood is intractable, but simulation is cheap. The distance considered, maximum mean discrepancy (MMD), is defined through the embedding of probability measures into a reproducing kernel Hilbert space. We study the theoretical properties of these estimators, showing that they are consistent, asymptotically normal and robust to model misspecification. A main advantage of these estimators is the flexibility offered by the choice of kernel, which can be used to trade-off statistical efficiency and robustness. On the algorithmic side, we study the geometry induced by MMD on the parameter space and use this to introduce a novel natural gradient descent-like algorithm for efficient implementation of these estimators. We illustrate the relevance of our theoretical results on several classes of models including a discrete-time latent Markov process and two multivariate stochastic differential equation models.

1 Introduction

Consider an open subset $\mathcal{X} \subset \mathbb{R}^d$ and denote by $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on this domain. We consider the problem of learning a probability measure $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ from identically and independently distributed (IID) realisations $\{y_j\}_{j=1}^{m} \overset{\text{IID}}{\sim} \mathbb{Q}$. We will focus on parametric inference with a parametrised family $\mathcal{P}_\Theta(\mathcal{X}) = \{\mathbb{P}_\theta \in \mathcal{P}(\mathcal{X}) : \theta \in \Theta\}$, for an open set $\Theta \subset \mathbb{R}^p$ i.e. we seek $\theta^* \in \Theta$ such that $\mathbb{P}_{\theta^*}$ is closest to $\mathbb{Q}$ in an appropriate sense. If $\mathbb{Q} \in \mathcal{P}_\Theta(\mathcal{X})$ we are in the *M-closed* setting, otherwise we are in the *M-open* setting. When $\mathbb{P}_\theta$ has a density $p(\cdot|\theta)$ with respect to the Lebesgue measure, then a standard approach is to use the maximum likelihood estimator (MLE):

$$\hat{\theta}_m^{\text{MLE}} = \arg\max_{\theta \in \Theta} \frac{1}{m} \sum_{j=1}^{m} \log p(y_j|\theta).$$

1

For complex models, a density may not be easily computable, or even exist and so the MLE need not be available. In some cases it is possible to approximate the likelihood; see for example pseudo likelihood [Besag, 1974], profile likelihood [Murphy and van der Vaart, 2000] and composite likelihood [Varin et al., 2011] estimation. It is sometimes also possible to access likelihoods in un-normalised forms i.e. $p(y|\theta) = \bar{p}(y|\theta)/C(\theta)$ where the constant $C(\theta)$ is unknown. This class of models is known as un-normalised models, or doubly-intractable models in the Bayesian literature, and a range of exact and approximate methods have been developed for this case; see for example the Markov chain Monte Carlo (MCMC) algorithms of Murray et al. [2006], Moller et al. [2006] or the score-based and ratio-based approaches of Hyvärinen [2006, 2007], Gutmann and Hyvarinen [2012].

However, for many models of interest in modern statistical inference, none of the methods above can be applied straightforwardly and efficiently due to the complexity of the likelihoods involved. This is most notably the case for intractable generative models, sometimes also called implicit models or likelihood-free models; see Mohamed and Lakshminarayanan [2016] for a recent overview. Intractable generative models are parametric families of probability measures for which it is possible to obtain realisations for any value of the parameter $\theta \in \Theta$, but for which we do not necessarily have access to a likelihood or approximation thereof. These models are used throughout the sciences, including in the fields of ecology [Wood, 2010], population genetics [Beaumont et al., 2002] or astronomy [Cameron and Pettitt, 2012]. They also appear in machine learning as black-box models; see for example generative adversarial networks (GANs) [Goodfellow et al., 2014] and variational auto-encoders [Kingma and Welling, 2014].

Given a Borel probability space $(\mathcal{U}, \mathcal{F}, \mathbb{U})$, we will call generative model any probability measure which is the pushforward $G_\theta^\# \mathbb{U}$ of the probability measure $\mathbb{U}$ with respect to a measurable parametric map $G_\theta : \mathcal{U} \to \mathcal{X}$ called the *generator*. To generate n independent realisations from the model, we produce IID realisations $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and apply the generator to each of these samples: $x_i = G_\theta(u_i)$ for $i = 1, \ldots, n$. While it is straightforward to generate samples from these models, a likelihood function need not be available, given that an associated positive density may not be computable or even exist. We therefore require alternatives to the MLE.

The estimators studied in this paper fall within the class of minimum divergence/distance estimators [Pardo, 2005, Basu et al., 2011]. These are estimators minimising some notion of divergence $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ (or an approximation thereof) between an empirical measure $\mathbb{Q}^m = \frac{1}{m} \sum_{j=1}^m \delta_{y_j}$ (where δ_{y_j} denotes a Dirac measure at y_j), obtained from the data $\{y_j\}_{j=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$, and the parametric model:

$$\hat{\theta}_m^D = \underset{\theta \in \Theta}{\operatorname{argmin}} \, D(\mathbb{P}_\theta || \mathbb{Q}^m) \tag{1}$$

If $\mathbb{Q}^m$ was absolutely continuous with respect to $\mathbb{P}_\theta$, maximising the likelihood would correspond to minimising the Kullback-Leibler (KL) divergence which, given $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{X})$, is defined as $D_{\text{KL}}(\mathbb{P}_1 || \mathbb{P}_2) := \int_{\mathcal{X}} \log(d\mathbb{P}_1/d\mathbb{P}_2) d\mathbb{P}_1$, where $d\mathbb{P}_1/d\mathbb{P}_2$ is the Radon-Nikodym derivative of $\mathbb{P}_1$ with respect to $\mathbb{P}_2$. This approach to inference is useful for models with complicated or intractable likelihood, since the choice of divergence can be adapted to the class of models of interest.

In previous works, minimum distance estimators for generative models have been considered based on the Wasserstein distance and its Sinkhorn relaxation; see Bassetti et al. [2006], Frogner et al. [2015], Montavon et al. [2016], Genevay et al. [2018], Frogner and Poggio [2018], Sanjabi

et al. [2018]. These have the advantage that they can leverage extensive work in the field of optimal transport. In a Bayesian context, similar ideas are used in approximate Bayesian computation (ABC) methods Marin et al. [2012], Lintusaari et al. [2017] where synthetic data sets are simulated from the model then compared to the true data using some notion of distance. There, significant work has been put into automating the choice of distance [Fearnhead and Prangle, 2011], and the use of the Wasserstein distance has also recently been studied [Bernton et al., 2019].

In this paper, we shall investigate the properties of minimal divergence estimators based on an approximation of *maximum mean discrepancy* (MMD). Such estimators have already been used extensively in the machine learning literature with generators taking the form of neural networks [Dziugaite et al., 2015, Li et al., 2015, 2017, Sutherland et al., 2017, Arbel et al., 2018, Bińkowski et al., 2018, Romano et al., 2018, dos Santos et al., 2019] where they are usually called MMD GANs, but can be used more generally. Our main objective in this paper is to present a general framework for minimum MMD estimators, to study their theoretical properties and to provide an initial discussion of the impact of the choice of kernel. This study brings insights into the favourable empirical results of previous work in MMD for neural networks, and demonstrate more broadly the usefulness of this approach for inference within the large class of intractable generative models of interest in the statistics literature. As will be discussed, this approach is significantly preferable to alternatives based on the Wasserstein distance for models with expensive generators as it comes with significantly stronger generalisation bounds and is more robust in several scenarios. Our detailed contributions can be summarised as follows:

1. In Section 2, we introduce the MMD metric, minimum MMD estimators, and the statistical Riemannian geometry the metric induces on the parameter space Θ. Through this, we rephrase the mimimum divergence estimator problem in terms of a gradient flow, thus obtaining a stochastic natural gradient descent method for finding the estimator θ^* which can significantly reduce computation cost as compared to stochastic gradient descent.

2. In Section 3, we focus on the theoretical properties of minimum MMD estimators and associated approximations. We use the information geometry of MMD to demonstrate generalisation bounds and statistical consistency, then prove that the estimator is asymptotically normal in the M-closed setting. These results give us necessary assumptions on the generator for the use of the estimators. We then analyse the robustness properties of the estimator in the M-open setting, establishing conditions for qualitative and quantitative robustness.

3. In Section 4 we study the efficiency and robustness of minimum MMD estimators based on Gaussian kernels for classes of isotropic Gaussian location and scale models. We demonstrate the effect of the kernel lengthscale on the efficiency of the estimators, and demonstrate a tradeoff between (asymptotic) efficiency and robustness. For high-dimensional problems, we demonstate that choosing the lengthscale according to the median heuristic provides an asymptotic variance independent of dimensionality. We also extend our analysis to mixtures of kernels, providing insights on settings often considered in machine learning applications.

4. In Section 5, we perform numerical simulations to support the theory detailed in the previous sections, demonstrating the behaviour of minimum MMD estimators for a number

of examples including estimation of unknown parameters for the g-and-k distribution, in a stochastic volatility model and for two systems of stochastic differential equations.

2 The Maximum Mean Discrepancy Statistical Manifold

We begin by formalising the notion of MMD and introduce the corresponding minimum MMD estimators. We then use tools from information geometry to analyse these estimators, which leads to a stochastic natural gradient descent algorithm for efficient implementation.

2.1 Maximum Mean Discrepancy

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Borel measurable kernel on $\mathcal{X}$, and consider the reproducing kernel Hilbert space $\mathcal{H}_k$ associated with k (see Berlinet and Thomas-Agnan [2004]), equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$. Let $\mathcal{P}_k(\mathcal{X})$ be the set of Borel probability measures μ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\mu(\mathrm{d}x) < \infty$. The *kernel mean embedding* $\Pi_k(\mu) = \int_{\mathcal{X}} k(\cdot, y)\mu(\mathrm{d}y)$, intepreted as a Bochner integral, defines a continuous embedding from $\mathcal{P}_k(\mathcal{X})$ into $\mathcal{H}_k$. The mean embedding pulls-back the metric on $\mathcal{H}_k$ generated by the inner product to define a pseudo-metric on $\mathcal{P}_k(\mathcal{X})$ called the maximum mean discrepancy MMD $: \mathcal{P}_k(\mathcal{X}) \times \mathcal{P}_k(\mathcal{X}) \to \mathbb{R}_+$, i.e., $\mathrm{MMD}(\mathbb{P}_1||\mathbb{P}_2) = \|\Pi_k(\mathbb{P}_1) - \Pi_k(\mathbb{P}_2)\|_{\mathcal{H}_k}$. The squared-MMD has a particularly simple expression that can be derived through an application of the reproducing property ($f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k}$):

$$
\begin{aligned}
\mathrm{MMD}^2(\mathbb{P}_1||\mathbb{P}_2) := {} & \left\| \int_{\mathcal{X}} k(\cdot, x)\mathbb{P}_1(\mathrm{d}x) - \int_{\mathcal{X}} k(\cdot, x)\mathbb{P}_2(\mathrm{d}x) \right\|_{\mathcal{H}_k}^2 \\
= {} & \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{P}_1(\mathrm{d}x)\mathbb{P}_1(\mathrm{d}y) - 2 \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{P}_1(\mathrm{d}x)\mathbb{P}_2(\mathrm{d}y) \\
& + \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{P}_2(\mathrm{d}x)\mathbb{P}_2(\mathrm{d}y),
\end{aligned}
$$

thus providing a closed form expression up to calculation of expectations. The MMD is in fact a *integral probability pseudo-metric* [Muller, 1997, Sriperumbudur et al., 2012, Sriperumbudur, 2016] since it can be expressed as:

$$
\mathrm{MMD}(\mathbb{P}_1||\mathbb{P}_2) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int_{\mathcal{X}} f(x)\mathbb{P}_1(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\mathbb{P}_2(\mathrm{d}x) \right|.
$$

Integral probability metrics are prominent in the information-based complexity literature where they correspond to the worst-case integration error [Dick et al., 2013, Briol et al., 2019]. If Π_k is injective then the kernel k is said to be characteristic [Sriperumbudur et al., 2010]. In this case MMD becomes a metric on $\mathcal{P}_k$ (and hence a statistical divergence). A sufficient condition for k to be characteristic is that k is *integrally strictly positive definite*, i.e. $\int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y)\mathbb{P}(\mathrm{d}x)\mathbb{P}(\mathrm{d}y) = 0$ implies that $\mathbb{P} = 0$ for all $\mathbb{P} \in \mathcal{P}_k$. On $\mathcal{X} = \mathbb{R}^d$, Sriperumbudur et al. [2010] showed that the Gaussian and inverse multiquadric kernels are both integrally strictly positive definite. We shall assume this condition holds throughout the paper, unless explicitly stated otherwise.

2.2 Minimum MMD estimators

This paper proposes to use MMD in a minimum divergence estimator framework for inference in intractable generative models. Given an unknown data generating distribution $\mathbb{Q}$ and a parametrised family of model distributions $\mathcal{P}_\Theta(\mathcal{X})$, we consider a minimum MMD estimator:

$$\hat{\theta}_m = \arg\min_{\theta \in \Theta} \text{MMD}^2(\mathbb{P}_\theta || \mathbb{Q}^m), \qquad (2)$$

where $\mathbb{Q}^m(\mathrm{d}y) = \frac{1}{m}\sum_{i=1}^{m} \delta_{y_i}(\mathrm{d}y)$, and $\{y_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{Q}$. In the following we will use $\mathbb{Q}^m$ to denote both the random measure $\mathbb{Q}^m$ and the measure $\mathbb{Q}^m(\omega)$, and we shall assume that $\mathcal{P}_\Theta(\mathcal{X}) \subset \mathcal{P}_k(\mathcal{X})$. Several existing methodologies fall within this general framework, including kernel scoring rules [Eaton, 1982] and MMD GANs [Dziugaite et al., 2015, Li et al., 2015]. For analogous methodology in a Bayesian context, see kernel ABC [Fukumizu et al., 2013, Park et al., 2015].

In general, the optimisation problem will not be convex and the minimiser $\hat{\theta}_m$ will not be computable analytically. If the generator G_θ is differentiable with respect to θ with a computable Jacobian matrix, the minimiser will be a fixed point of the equation $\dot{\theta} = -\nabla_\theta\text{MMD}^2(\mathbb{P}_\theta||\mathbb{Q}^m)$ where $\nabla_\theta = (\partial_{\theta_1}, \ldots, \partial_{\theta_p})$. Assuming that the Jacobian $\nabla_\theta G_\theta$ is $\mathbb{U}$-integrable then the gradient term can be written as

$$\nabla_\theta\text{MMD}^2(\mathbb{P}_\theta||\mathbb{Q}^m) = 2\int_\mathcal{U}\int_\mathcal{U} \nabla_1 k(G_\theta(u), G_\theta(v))\nabla_\theta G_\theta(u)\mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v)$$
$$- \frac{2}{m}\sum_{j=1}^{m}\int_\mathcal{U} \nabla_1 k(G_\theta(u), y_j)\nabla_\theta G_\theta(u)\mathbb{U}(\mathrm{d}u),$$

where $\nabla_1 k$ corresponds to the partial derivative with respect to the first argument. In practice it will not be possible to compute the integral terms analytically. We can introduce a U-statistic approximation for the gradient as follows:

$$\hat{J}_\theta(\mathbb{Q}^m) = = \frac{2\sum_{i \neq i'} \nabla_\theta G_\theta(u_i)\nabla_1 k(G_\theta(u_i), G_\theta(u_{i'}))}{n(n-1)} - \frac{2\sum_{j=1}^{m}\sum_{i=1}^{n} \nabla_\theta G_\theta(u_i)\nabla_1 k(G_\theta(u_i), y_j)}{nm},$$

where $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$. This is an unbiased estimator in the sense that $\mathbb{E}[\hat{J}_\theta(\mathbb{Q}^m)] = \nabla_\theta\text{MMD}^2(\mathbb{P}_\theta||\mathbb{Q}^m)$, where the expectation is taken over the independent realisations of the $u_i's$. This allows us to use a stochastic gradient descent (SGD) [Dziugaite et al., 2015, Li et al., 2015]: starting from $\hat{\theta}^{(0)} \in \Theta$, we iterate:

(i) Sample $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and set $x_i = G_{\hat{\theta}^{(k-1)}}(u_i)$ for $i = 1, \ldots, n$.

(ii) Compute $\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \eta_k \hat{J}_{\hat{\theta}^{(k-1)}}(\mathbb{Q}^m)$.

where $(\eta_k)_{k\in\mathbb{N}}$ is a step size sequence chosen to guarantee convergence (see [Robbins and Monro, 1985]) to the minimiser in Equation 2. For large values of n, the SGD should approach $\hat{\theta}_m$, but this will come at significant computational cost. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Theta \subseteq \mathbb{R}^p$. The overall cost of the gradient descent algorithm is $\mathcal{O}\left((n^2 + nm)dp\right)$ per iteration. This cost is linear in the number of data points m, but quadratic in the number of simulated samples n. It could be made linear in n by considering approximations of the maximum mean discrepancy as found in Chwialkowski

et al. [2015]. In large data settings (i.e. m large), subsampling elements uniformly at random from $\{y_j\}_{j=1}^m$ may lead to significant speed-ups.

Clearly, when the generator G_θ and its gradient $\nabla_\theta G_\theta$ are computationally intensive, letting n grow will become effectively intractable, and it will be reasonable to assume that the number of simulations n is commensurate or even smaller than the sample size. To study the behaviour of minimum MMD estimators when synthetic data is prohibitively expensive, we consider the following minimum divergence estimator: $\hat{\theta}_{n,m} = \text{argmin}_{\theta \in \Theta} \text{MMD}^2_{U,U}(\mathbb{P}^n_\theta || \mathbb{Q}^m)$ based on a U-statistic approximation of the MMD:

$$\text{MMD}^2_{U,U}(\mathbb{P}^n_\theta || \mathbb{Q}^m) = \frac{\sum_{i \neq i'} k(x_i, x_{i'})}{n(n-1)} - \frac{2\sum_{j=1}^m \sum_{i=1}^n k(x_i, y_j)}{mn} + \frac{\sum_{j \neq j'} k(y_j, y_{j'})}{m(m-1)}.$$

where $\mathbb{P}^n_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for some $\{x_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{P}_\theta$. This estimator is closely related to the method of simulated moments [Hall, 2005] and satisfies $\mathbb{E}[\text{MMD}^2_{U,U}(\mathbb{P}^n_\theta || \mathbb{Q}^m)] = \text{MMD}^2(\mathbb{P}_\theta || \mathbb{Q})$, thus providing an unbiased estimator of the square distance between $\mathbb{P}_\theta$ and $\mathbb{Q}$. While the estimator $\hat{\theta}_{n,m}$ is not used in practice (since we re-sample from the generator at each gradient iteration), it is an idealisation which gives us insights into situations where the gradient descent cannot be iterated for a large numbers of steps relative to the observed data-set size, and so we cannot appeal on the law of large numbers.

2.3 The Information Geometry induced by MMD

The two estimators $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ defined above are flexible in the sense that the choice of kernel and kernel hyperparameters will have a significant influence on the geometry induced on the space of probability measures. This section studies this geometry and develops tools which will later give us insights into the impact of the choice of kernel on the generalisation, asymptotic convergence and robustness of the corresponding estimators.

Let $\mathcal{P}_\Theta(\mathcal{X})$ be a family of measures contained in $\mathcal{P}_k(\mathcal{X})$ and parametrised by an open subset $\Theta \subset \mathbb{R}^p$. Assuming that the map $\theta \to \mathbb{P}_\theta$ is injective, the MMD distance between the elements $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ in $\mathcal{P}_k$ induces a distance between θ and θ' in Θ. Under appropriate conditions this gives rise to a Riemmanian manifold structure on Θ. The study of the geometry of such statistical manifolds lies at the center of information geometry [Amari, 1987, Barndorff-Nielsen, 1978]. Traditional information geometry focuses on the statistical manifold induced by the Kullback-Leibler divergence over a parametrised set of probability measures. This yields a Riemmanian structure on the parameter space with the metric tensor given by the Fisher-Rao metric. A classic result due to Cencov [2000] characterises this metric as the unique metric invariant under a large class of transformations (i.e. embeddings via Markov morphisms, see [Campbell, 1986, Montúfar et al., 2014]).

In this section, we study instead the geometry induced by MMD. To fix ideas, we shall consider a generative model distribution of the form $\mathbb{P}_\theta = G_\theta^\# \mathbb{U}$ for $\theta \in \Theta$, where $(\mathcal{U}, \mathcal{F}, \mathbb{U})$ is an underlying Borel measure space. We assume that (i) $G_\theta(\cdot)$ is $\mathcal{F}$-measurable for all $\theta \in \Theta$; (ii) $G_\cdot(u) \in C^1(\Theta)$ for all $u \in \mathcal{U}$; (iii) $\|\nabla_\theta G_\theta(\cdot)\| \in L^1(\mathbb{U})$, for all $\theta \in \Theta$. Suppose additionally that the kernel k has bounded continuous derivatives over $\mathcal{X} \times \mathcal{X}$. Define the map $J : \Theta \to \mathcal{H}_k$ to be the Bocher integral $J(\theta) = \Pi_k(\mathbb{P}_\theta)$. By [Hájek and Johanis, 2014, Theorem 90], assumptions

6

(i)-(iii) imply that the map J is Fréchet differentiable and

$$\partial_{\theta_i} J(\theta)(\cdot) = \int_{\mathcal{U}} \nabla_2 k(\,\cdot\,, G_\theta(u))\partial_{\theta_i} G_\theta(u)\mathbb{U}(\mathrm{d}u).$$

The map J induces a degenerate-Riemannian metric $g(\theta)$ on Θ given by the pull-back of the inner product on $\mathcal{H}_k$. In particular its components in the local coordinate-system are $g_{ij}(\theta) = \langle \partial_{\theta_i} J(\theta), \partial_{\theta_j} J(\theta)\rangle_{\mathcal{H}_k}$ for $i, j \in \{1, \ldots, p\}$. By [Steinwart and Christmann, 2008, Lemma 4.34], it follows that for $i, j \in \{1, \ldots, p\}$,

$$g(\theta) = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G_\theta(u)^\top \nabla_2 \nabla_1 k(G_\theta(u), G_\theta(v)) \nabla_\theta G_\theta(v)\mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v), \tag{3}$$

where $\nabla_1 \nabla_2 k(x, y) = \{\partial_{x_i} \partial_{y_j} k(x, y)\}_{i,j=1,\ldots,d}$. The induced metric tensor is in fact just the information metric associated to the MMD-squared divergence (see A.1). Further details about the geodesics induced by MMD can be found in Appendix A.2. This information metric will allow us to construct efficient optimisation algorithm and study the statistical properties of the minimum MMD estimators.

2.4 MMD Gradient Flow

Given the loss function $L(\theta) = \mathrm{MMD}^2(\mathbb{P}_\theta || \mathbb{Q}^m)$, a standard approach to finding a minimum divergence estimator is via gradient descent (or in our case stochastic gradient descent). Gradient descent methods aim to minimise a function L by following a curve $\theta(t)$, known as the gradient flow, that is everywhere tangent to the direction of steepest descent of L. This direction depends on the choice of Riemannian metric g on Θ, and is given by $-\nabla_g L$ where $\nabla_g L$ denotes the Riemannian gradient (or covariant derivative) of L.

A particular instance of gradient descent, based on the Fisher Information metric, was developed by Amari and collaborators [Amari, 1998]. It is a widely used alternative to standard gradient descent methods and referred to as natural gradient descent. It has been successfully applied to a variety of problems in machine learning and statistics, for example reinforcement learning [Kakade, 2002], neural network training [Park et al., 2000], Bayesian variational inference methods [Hoffman et al., 2013] and Markov chain Monte Carlo [Girolami and Calderhead, 2011]. While the classical natural gradient approach is based on the Fisher information matrix induced by the KL divergence, information geometries arising from other metrics on probabilities have also been studied in previous works, including those arising from optimal transport metrics [Chen and Li, 2018, Li and Montufar, 2018] and the Fisher divergence [Karakida et al., 2016].

As discussed above, a gradient descent method can be formulated as an ordinary differential equation for the *gradient flow* $\theta(t)$ which solves $\dot{\theta}(t) = -\nabla_g L(\theta(t))$ for some specified initial conditions. In local coordinates the Riemannian gradient can be expressed in terms of the standard gradient ∇_θ, formally $\nabla_g = g^{-1}(\theta)\nabla_\theta$, so we have $\dot{\theta}(t) = -g^{-1}(\theta)\nabla_\theta L(\theta)$. This flow can be approximated by taking various discretisations. An explicit Euler discretisation yields the scheme: $\theta^{(k)} = \theta^{(k-1)} - \eta_k g^{-1}(\theta^{(k-1)})\nabla_\theta L(\theta^{(k-1)})$. Under appropriate conditions on the step-size sequence $(\eta_k)_{k\in\mathbb{N}}$ this gradient descent scheme will converge to a local minimiser of $L(\theta)$. Provided that $\nabla_\theta L(\theta)$ and the metric tensor are readily computable, the Euler discretisation yields a gradient scheme analogous to those detailed in [Amari, 1987, 1998].

For the MMD case, we cannot evaluate g from Equation (3) exactly since it contains intractable integrals against $\mathbb{U}$. We can however use a similar approach to that used for the stochastic gradient algorithm and introduce a U-statistic approximation of the intractable integrals:

$$g_U(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \nabla_\theta G_\theta(u_i)^\top \nabla_2 \nabla_1 k(G_\theta(u_i), G_\theta(u_j)) \nabla_\theta G_\theta(u_j),$$

where $\{u_i\}_{i=1}^n$ are IID realisations from $\mathbb{U}$. We propose to perform optimisation using the following natural stochastic gradient descent algorithm: starting from $\hat{\theta}^{(0)} \in \Theta$, we iterate

(i) Sample $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and set $x_i = G_{\hat{\theta}^{(k-1)}}(u_i)$ for $i = 1, \ldots, n$.

(ii) Compute $\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \eta_k g_U\big(\hat{\theta}^{(k-1)}\big)^{-1} \hat{J}_{\hat{\theta}^{(k-1)}} (\mathbb{Q}^m)$.

The experiments in Section 5 demonstrate that this new algorithm can provide significant computational gains. This could be particularly impactful for GANs, where a large number of stochastic gradient descent are currently commonly used. The approximation of the inverse metric tensor does however yield an additional computational cost due to the inversion of a dense matrix: $\mathcal{O}(((n^2 + nm)p^2 d + p^3))$ per iteration. When the dimension of the parameter set Θ is high, the calculation of the inverse metric at every step can hence be prohibitive. The use of online methods to approximate g^{-1} sequentially without needing to compute inverses of dense matrices can be considered as in [Ollivier, 2018], or alternatively, approximate linear solvers could also be used to reduce this cost.

In certain cases, the gradient of the generator $\nabla_\theta G_\theta$ may not be available in closed form, precluding exact gradient descent inference. An alternative is the method of finite difference stochastic approximation [Kushner and Yin, 2003] can be used to approximate an exact descent direction. Alternatively, one can consider other discretisations of the gradient flow. For example, a fully implicit discretisation yields the following scheme [Jordan et al., 1998]:

$$\theta^{(k)} = \arg\min_{\theta \in \Theta} L(\theta) + \frac{1}{2\eta} \text{MMD}^2(\mathbb{P}_\theta || \mathbb{P}_{\theta^{(k-1)}}), \tag{4}$$

where $\eta > 0$ is a step-size. Therefore the natural gradient flow can be viewed as a motion towards a lower value of $L(\theta)$ but constrained to be close (in MMD) to the previous time-step. The constant η controls the strength of this constraint, and thus can be viewed as a step size. The formulation allows the possibility of a natural gradient descent approach being adopted even if $\nabla_\theta L$ and g are not readily computable. Indeed, (4) could potentially be minimised using some gradient-free optimisation method such as Nelder-Mead.

2.5 Minimum MMD Estimators and Kernel Scoring Rules

Before concluding this background section, we highlight the connection between our minimum MMD estimators and scoring rules [Dawid, 2007]. A scoring rule is a function $S : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ such that $S(x, \mathbb{P})$ quantifies the accuracy of a model $\mathbb{P}$ upon observing the realisation $x \in \mathcal{X}$ (see [Gneiting and Raftery, 2007] for technical conditions). We say a scoring rule is strictly proper if $\int_\mathcal{X} S(x, \mathbb{P}_1) \mathbb{P}_2(\text{d}x)$ is uniquely minimised when $\mathbb{P}_1 = \mathbb{P}_2$. Any strictly proper scoring rule induces

a divergence of the form $D_S(\mathbb{P}_1||\mathbb{P}_2) = \int_{\mathcal{X}} S(x, \mathbb{P}_1)\mathbb{P}_2(\mathrm{d}x) - \int_{\mathcal{X}} S(x, \mathbb{P}_2)\mathbb{P}_2(\mathrm{d}x)$. These divergences can then be used to obtain minimum distance estimators: $\hat{\theta}_m^S = \mathrm{argmin}_{\theta \in \Theta} D_S(\mathbb{P}_\theta||\mathbb{Q}^m) = \mathrm{argmin}_{\theta \in \Theta} \sum_{j=1}^m S(y_j, \mathbb{P}_\theta)$. One way to solve this problem is by setting the gradient in θ to zero; i.e. solving $\sum_{j=1}^m \nabla_\theta S(y_j, \mathbb{P}_\theta) = 0$, called estimating equations.

The minimum MMD estimators $\hat{\theta}_m$ in this paper originate from the well-known kernel scoring rule [Eaton, 1982, Dawid, 2007, Zawadzki and Lahaie, 2015, Steinwart and Ziegel, 2017, Masnadi-Shirazi, 2017], which takes the form

$$S(x, \mathbb{P}) = k(x, x) - 2 \int_{\mathcal{X}} k(x, y)\mathbb{P}(\mathrm{d}y) + \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, z)\mathbb{P}(\mathrm{d}y)\mathbb{P}(\mathrm{d}z).$$

This connection between scoring rules and minimum MMD estimators will be useful for theoretical results in the following section. Whilst the present paper focuses on minimum MMD estimators for generative models, our results also have implications for kernel scoring rules.

3 Behaviour of Minimum MMD estimators

The two estimators $\hat{\theta}_n$ and $\hat{\theta}_{n,m}$ defined above are flexible in the sense that the choice of kernel and kernel hyperparameters will have a significant influence on the geometry induced on the space of probability measures. This choice will also have an impact on the generalisation, asymptotic convergence and robustness of the estimators, as will be discussed in this section.

3.1 Concentration and Generalisation Bounds for MMD

In this section we will restrict ourselves to the case where $\mathcal{X} \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$ for $d, p \in \mathbb{N}$. Given observations $\{y_i\}_{i=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$, it is clear that the convergence and efficiency of $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ in the limit of large n and m will depend on the choice of kernel k as well as the dimensions p and d. As a first step, we obtain estimates for the out-of-sample error for each estimator, in the form of generalization bounds.

The necessary conditions in this proposition are quite natural. They are required to ensure the existence of $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$, and reclude models which are unidentifiable over a non-compact subset of parameters, i.e. models for which there are minimising sequences $\hat{\theta}_m$ of $\mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m)$ which are unbounded. While these assumptions must be verified on a case-by-case basis, for most models we would expect these conditions to hold immediately.

Assumption 1. *1. For every $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$, there exists $c > 0$ such that the set $\{\theta \in \Theta : MMD(\mathbb{P}_\theta||\mathbb{Q}) \leq \inf_{\theta' \in \Theta} MMD(\mathbb{P}_{\theta'}||\mathbb{Q}) + c\}$, is bounded.*

2. For every $n \in \mathbb{N}$ and $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$, there exists $c_n > 0$ such that the set $\{\theta \in \Theta : MMD(\mathbb{P}_\theta^n||\mathbb{Q}) \leq \inf_{\theta' \in \Theta} MMD(\mathbb{P}_{\theta'}||\mathbb{Q}) + c_n\}$, is almost surely bounded.

Theorem 1 (Generalisation Bounds)**.** *Suppose that the kernel k is bounded, and that Assumption 1 holds, then with probability at least $1 - \delta$,*

$$MMD\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) \leq \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}) + 2\sqrt{\frac{2}{m} \sup_{x \in \mathcal{X}} k(x, x)} \left(2 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right),$$

and

$$MMD\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) \leq \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}) + 2\left(\sqrt{\frac{2}{n}} + \sqrt{\frac{2}{m}}\right)\sqrt{\sup_{x \in \mathcal{X}} k(x,x)}\left(2 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right).$$

All proofs are deferred to Appendix B. An immediate corollary of the above result is that the speed of convergence in the generalisation errors decreases as $n^{-\frac{1}{2}}$ and $m^{-\frac{1}{2}}$ with the rates being independent of the dimensions p and d, and the properties of the kernel. Indeed, if the kernel is translation invariant, then $k(x,x)$ will reduce to the maximum value of the kernel. A similar generalisation result was obtained in Dziugaite et al. [2015] for minimum MMD estimation of deep neural network models. While the bounds are of the same form, Theorem 1 only requires minimal assumptions on the smoothness of the kernel. Moreover, all the constants in the bound are explicit, demonstrating clearly dimensional dependence. Assumption 1 is required to guarantee the existence of at least one minimiser, whereas this is implicitly assumed in Dziugaite et al. [2015]. The key result which determines the rate is the following concentration inequality.

Lemma 1 (Concentration Bound). *Assume that the kernel k is bounded and let $\mathbb{P}$ be a probability measure on $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\mathbb{P}^n$ be the empirical measure obtained from n independently and identically distributed samples of $\mathbb{P}$. Then with probability $1 - \delta$, we have that*

$$MMD(\mathbb{P}||\mathbb{P}^n) \leq \sqrt{\frac{2}{n}\sup_{x \in \mathcal{X}} k(x,x)}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).$$

See also [Gretton et al., 2009, Theorem 17] for an equivalent bound. We can compare this result with [Fournier and Guillin, 2015, Theorem 1] on comparing the rate of convergence of Wasserstein-1 distance (denoted W_1) to the empirical measure, which implies that for $d > 2$ and q sufficiently large, with probability $1 - \delta$ we have $W_1(\mathbb{P}||\mathbb{P}^n) \leq CM_q^{1/q}(\mathbb{P})\delta^{-1}n^{-1/d}$, where $M_q(\mu) := \int_{\mathcal{X}} |x|^q \mu(\mathrm{d}x)$ and C is a constant depending only on the constants p, q and d. This suggests that generalisation bounds analogous to Theorem 1 for Wasserstein distance would depend exponentially on dimension, at least when the distribution is absolutely continuous with respect to the Lebesgue measure. For measures support on a lower dimensional manifold, this bound has been recently tightened, see Weed and Bach [2017] and also Weed and Berthet [2019]. For Sinkhorn divergences, which interpolate between optimal transport and MMD distance Genevay et al. [2018] this curse of dimensionality can be mitigated Genevay et al. [2019] for measures on bounded domains.

3.2 Consistency and Asymptotic Normality

With additional assumptions, we can recover a classical strong consistency result.

Proposition 1 (Consistency). *Suppose that Assumption 1 holds and that there exists a unique minimiser $\theta^* \in \Theta$ such that $MMD(\mathbb{P}_{\theta^*}||\mathbb{Q}) = \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q})$. Then $\lim_{m \to \infty} \hat{\theta}_m = \theta^*$ and $\lim_{m,n \to \infty} \hat{\theta}_{m,n} = \theta^*$ as $n, m \to \infty$, almost surely.*

Theorem 1 provides fairly weak probabilistic bounds on the convergence of the estimators $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ in terms of their MMD distance to the data distribution $\mathbb{Q}$. Proposition 1 provides

conditions under which these bounds translate to convergence of the estimators, however it is not clear how to extract quantitative information about the speed of convergence, and the efficiency of the estimator in general. A classical approach to this is to establish the asymptotic normality of the estimators and characterise the efficiency in terms of the asymptotic variance. We do this now, assuming that we are working in the M-close setting, i.e. assuming that $\mathbb{Q} = \mathbb{P}_{\theta^*}$ for some θ^*.

Theorem 2 (Central Limit Theorems). *Suppose that $\mathbb{Q} = \mathbb{P}_{\theta^*}$ for some $\theta^* \in \Theta$ and that the conclusions of Proposition 1 hold. Suppose that:*

1. *There exists an open neighbourhood $O \subset \Theta$ of θ^* such that G_θ is three times differentiable in O with respect to θ.*

2. *The information metric $g(\theta)$ is positive definite at $\theta = \theta^*$.*

3. *There exists a compact neighbourhood $K \subset O$ of θ^* such that $\int_{\mathcal{U}} \sup_{\theta \in K} \left\| \nabla^{(i)} G_\theta(u) \right\| \mathbb{U}(\mathrm{d}u) < \infty$ for $i = 1, 2, 3$ where $\nabla^{(i)}$ denotes the mixed derivatives of order i and $\|\cdot\|$ denotes the spectral norm.*

4. *The kernel $k(\cdot, \cdot)$ is translation invariant, with bounded mixed derivatives up to order 2.*

Then as $k \to \infty$:

$$\sqrt{m} \left(\hat{\theta}_m - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, C),$$

where $\xrightarrow{d}$ denotes convergence in distribution. The covariance matrix is given by the Godambe *matrix $C = g(\theta^*)^{-1} \Sigma g(\theta^*)^{-1}$ where*

$$\Sigma = \int_{\mathcal{U}} \left(\int_{\mathcal{U}} \left(\nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) - \overline{M} \right) \mathbb{U}(\mathrm{d}u) \right)^{\otimes 2} \mathbb{U}(\mathrm{d}v)$$

and

$$\overline{M} = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v).$$

Here, $A \otimes B$ denotes the tensor product and $A^{\otimes 2} := A \otimes A$. Furthermore, suppose that:

5 *The kernel $k(\cdot, \cdot)$ has bounded mixed derivatives up to order 3.*

6 *The indices satisfy $n = n_k$, $m = m_k$ where $n_k/(n_k + m_k) \to \lambda \in (0, 1)$,*

Then, as $k \to \infty$,

$$\sqrt{n_k + m_k} \left(\hat{\theta}_{n,m} - \theta^* \right) \xrightarrow{d} \mathcal{N}(0, C_\lambda),$$

where $C_\lambda = (1/(1 - \lambda)\lambda)C$.

We remark that the asymptotic covariance C_λ is minimised when $\lambda = 1/2$, that is, when the number of samples n generated from the model equals that of the data m (at which point $C_\lambda = 4C$). This means that it will be computationally inefficient to use n much larger than m. We note that the variance also does not depend on any amplitude parameter of the kernel, or any location parameters in $\mathbb{U}$. To the best of our knowledge, there are no known analogous result for minimum Wasserstein or Sinkhorn estimators (except a one-dimensional result for the minimum Wasserstein estimator in the supplementary material of [Bernton et al., 2019]).

Theorem 2 raises the question of efficiency of the estimator. The Cramer-Rao bound provides a lower bound on the variance of any unbiased estimator for $\mathbb{P}_\theta$, and it is well-known that it is attained by maximum likelihood estimators. The following result is an adaptation of the Cramer-Rao bound in Godambe [1960] for our estimators, which are biased.

Theorem 3 (Cramer-Rao Bounds). *Suppose that the CLTs in Theorem 2 hold and that the data distribution $\mathbb{Q}$ satisfies $\mathbb{Q} = \mathbb{P}_{\theta^*}$, where $\mathbb{P}_{\theta^*} = G_{\theta^*}^{\#}\mathbb{U}$ is assumed to have density $p(x|\theta^*)$. Furthermore, suppose that the MMD information metric $g(\theta^*)$ and the Fisher information metric $F(\theta) = \int_{\mathcal{X}} \nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^\top \mathbb{P}_\theta(\mathrm{d}x)$ are positive definite when $\theta = \theta^*$. Then the asymptotic covariances C and C_λ of the estimator $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ satisfy the Cramer-Rao bound, i.e. $C - F(\theta^*)^{-1}$ and $C_\lambda - F(\theta^*)^{-1}$ are non-negative definite.*

The results above demonstrate that we cannot expect our (biased) estimators to outperform maximum likelihood in the M-closed case. The efficiency of these estimators is strongly determined by the choice of kernel, in particular on the kernel bandwidth l. The following result characterises the efficiency as $l \to \infty$.

Proposition 2 (Efficiency with Large Lengthscales). *Suppose that k is a radial basis kernel, i.e. $k(x, y) = r(|x - y|^2/2l^2)$, where $\lim_{s\to 0} r'(s) < \infty$ and $\lim_{s\to 0} r''(s) < \infty$. Let C^l and C_λ^l denote the asymptotic variance as a function of the bandwidth l of $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ respectively. Then*

$$\lim_{l\to\infty} C^l = (\nabla_\theta M(\theta))^\dagger V(\theta) (\nabla_\theta M(\theta))^{\dagger\top}, \tag{5}$$

where $M(\theta)$ and $V(\theta)$ are the mean and covariance of $p(x|\theta)$ respectively and $A^\dagger$ denotes the Moore-Penrose inverse of A. As a result, $\lim_{l\to\infty} C_\lambda^l = (1/(1-\lambda)\lambda)(\nabla_\theta M(\theta))^\dagger V(\theta) (\nabla_\theta M(\theta))^{\dagger\top}$.

In general, the minimum MMD estimators may not achieve the efficiency of maximum likelihood estimators in the limit $l \to \infty$, however in one dimension, the limiting covariance in Equation 5 is a well known approximation for the inverse Fisher information [Jarrett, 1984, Stein and Nossek, 2017], which is optimal.

Before concluding this section on efficiency of minimum MMD estimators, we note that the asymptotic covariances C and C_λ of Theorem 2 could be used to create confidence intervals for the value of θ^* (only for the M-closed case). Although these covariances cannot be estimated exactly since they depend on θ^* and contain intractable integrals, they can be approximated using the generator at the current estimated value of the parameters.

3.3 Robustness

This concludes our theoretical study of the M-closed case and we now move on to the M-open case. A concept of importance to practical inference is robustness when subjected to corrupted

data [Huber and Ronchetti, 2009]. As will be seen below, minimum MMD estimators have very favourable robustness properties for this case.

Our first objective is to demonstrate qualitative robustness in the sense of Hampel [1971]. More specifically, given some parametrized probability measure $\mathbb{P}_\theta$, we show that if two measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are close in Prokhorov metric, then the distributions of the minimum distance estimators $\hat{\theta}_m^i \in \operatorname{argmin}_{\theta \in \Theta} \mathrm{MMD}^2(\mathbb{P}_\theta || \mathbb{Q}_i^m)$ and $\hat{\theta}_{n,m}^i \in \operatorname{argmin}_{\theta \in \Theta} \mathrm{MMD}^2(\mathbb{P}_\theta^n || \mathbb{Q}_i^m)$ for $i = 1, 2$ are respectively close.

Theorem 4 (Qualitative Robustness). *Suppose that (i) $\forall \mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$ there exists a unique $\theta^{\mathbb{Q}}$ such that $\inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta || \mathbb{Q}) = MMD(\mathbb{P}_{\theta^{\mathbb{Q}}} || \mathbb{Q})$ and (ii) $\forall \epsilon > 0$, $\exists \delta > 0$ such that $\| \theta - \theta^{\mathbb{Q}} \| \geq \epsilon$ implies that $MMD(\mathbb{P}_\theta || \mathbb{Q}) > MMD(\mathbb{P}_{\theta^{\mathbb{Q}}} || \mathbb{Q}) + \delta$. Then $\hat{\theta}_m$ is qualitatively robust in the sense of Hampel [1971].*

Additionally, suppose that for any empirical measure $\mathbb{U}^n$ on n points, that (i') $\forall \mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$ there exists a unique $\theta^{\mathbb{Q}}$ such that $\inf_{\theta \in \Theta} MMD(G_\theta^\# \mathbb{U}^n || \mathbb{Q}) = MMD(G_{\theta^{\mathbb{Q}}}^\# \mathbb{U}^n || \mathbb{Q})$ and (ii') $\forall \epsilon > 0$, $\exists \delta > 0$ such that $\| \theta - \theta^{\mathbb{Q}} \| \geq \epsilon$ implies that $MMD(G_\theta^\# \mathbb{U}^n || \mathbb{Q}) > MMD(G_{\theta^{\mathbb{Q}}}^\# \mathbb{U}^n || \mathbb{Q}) + \delta$. Then $\exists N$ such that $\hat{\theta}_{n,m}$ is qualitatively robust for $n \geq N$.

The result above characterises the qualitative robustness of the estimators, but does not provide a measure of the degree of robustness which can be used to study the effect of corrupted data on the estimated parameters. An important quantity used to quantify robustness is the *influence function* IF : $\mathcal{X} \times \mathcal{P}_\Theta(\mathcal{X}) \to \mathbb{R}$ where $\mathrm{IF}(z, \mathbb{P}_\theta)$ measures the impact of an infinitesimal contamination of the data generating model $\mathbb{P}_\theta$ in the direction of a Dirac measure δ_z located at some point $z \in \mathcal{X}$. The influence function of a minimum distance estimator based on a scoring rule S is given by [Dawid and Musio, 2014]: $\mathrm{IF}_S(z, \mathbb{P}_\theta) := (\int_{\mathcal{X}} \nabla_\theta \nabla_\theta S(x, \mathbb{P}_\theta) \mathbb{P}_\theta(\mathrm{d}x))^{-1} \nabla_\theta S(z, \mathbb{P}_\theta)$. The supremum of the influence function over $z \in \mathcal{X}$ is called the gross-error sensitivity, and if it is finite, we say that an estimator is *bias-robust* [Hampel, 1971]. We can use the connection with kernel scoring rules to study bias robustness of our estimators.

Theorem 5 (Bias Robustness). *The influence function corresponding to the maximum mean discrepancy is given by $IF_{MMD}(z, \mathbb{P}_\theta) = g^{-1}(\theta) \nabla_\theta MMD(\mathbb{P}_\theta, \delta_z)$. Furthermore, suppose that $\nabla_1 k$ is bounded and $\int_{\mathcal{U}} \| \nabla_\theta G_\theta(u) \| \mathbb{U}(\mathrm{d}u) < \infty$, then the MMD estimators are bias-robust.*

Note that the conditions for this theorem to be valid are less stringent than assumptions required for the CLT in Theorem 2. As we shall see in the next section, there will be a trade-off between efficiency and robustness as the kernel bandwidth is varied. We shall demonstrate this through the influence function.

Overall, these results demonstrating the qualitative and bias robustness of minimum MMD estimators provides another strong motivation for their use. For complex generative model, it is common to be in the M-open setting; see for example all of the MMD GANs applications in machine learning where neural networks are used as models of images. Although it is not realistically expected that neural networks are good models for this, our robustness results can help explain the favourable experimental results observed in that case. Note that, to the best of our knowledge, the robustness of Wasserstein and Sinkhorn estimators has not been studied.

4 The Importance of Kernel Selection: Gaussian Models

As should be clear from the previous sections, the choice of kernel will strongly influence the characteristics of minimum MMD estimators, including (but not limited to) the efficiency of the estimators, their robustness to misspecification and the geometry of the loss function. In this section, we highlight some of these consequences for two particular models: a location and scale model for a Gaussian distribution. These models are illustrative problems for which many quantities of interest (such as the asymptotic variance and influence function) can be computed in closed form, allowing for a detailed study of the properties of minimum MMD estimators.

4.1 Kernel Selection in the Literature

A number of approaches for kernel selection have been proposed in the literature, most based on radial basis kernels of the form $k(x, y; l) = r(\|x - y\|/l)$, for some function $r : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We now highlight each of these approaches, and later discuss the consequences of our theoretical results in the case of Gaussian location and scale models.

Dziugaite et al. [2015] proposed to set the lengthscale using the median heuristic proposed in Gretton et al. [2008] for two-sample testing with MMD, and hence picked their lengthscale to be $\sqrt{\text{median}(\|y_i - y_j\|_2^2/2)}$ where $\{y_j\}_{j=1}^m$ is the data. This heuristic has previously been demonstrated to lead to high power in the context of two-sample testing for location models in Ramdas et al. [2015], Reddi et al. [2015]. See also Garreau et al. [2017] for an extensive investigation. Li et al. [2015], Ren et al. [2016], Sutherland et al. [2017] have demonstrated empirically that a mixture of squared-exponential kernels yields good performance, i.e. a kernel of the form $k(x, y) = \sum_{s=1}^S \gamma_s k(x, y; l_s)$ where $\gamma_1, \ldots, \gamma_S \in \mathbb{R}_+$ and the lengthscales $l_1, \ldots, l_S > 0$ are chosen to cover a wide range of bandwidth. The weights can either be fixed, or optimised; see Sutherland et al. [2017] for more details. As the sum of characteristic kernels is characteristic (see Sriperumbudur et al. [2010]) this is a valid choice of kernel.

Another approach orginating from the use of MMD for hypothesis testing consists of studying the asymptotic distribution of the test statistics, and choose kernel parameters so as to maximise the power of the test. This was for example used in [Sutherland et al., 2017]. A similar idea could be envisaged in our case: we could minimise the asymptotic variance of the CLT obtained in the previous section. Unfortunately, this will not be tractable in general since computing the asymptotic variance requires knowing the value of θ^*, but an approximation could be obtained using the current estimate of the parameter.

Finally, recent work [Li et al., 2017] also proposed to include the problem of kernel selection in the objective function, leading to a minimax objective. This renders the optimisation problem delicate to deal with in practice [Bottou et al., 2017]. The introduction of several constraints on the objective function have however allowed significant empirical success [Arbel et al., 2018, Bińkowski et al., 2018]. We do not consider this case, but it will be the subject of future work.

4.2 Gaussian Location Model

To focus ideas we shall focus on a Gaussian location model for a d-dimensional istropic Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_{d \times d})$ with unknown mean $\theta \in \mathbb{R}^d$ and known standard deviation $\sigma > 0$. In this case, we take $\mathcal{U} = \mathcal{X} = \mathbb{R}^d$, $\mathbb{U}$ is a standard Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{d \times d})$ and

$G_\theta(u) = u + \theta$. The derivative of the generator is given by $\nabla_\theta G_\theta(u) = I_{d \times d}$. Although this is of course a fairly simple model which could be estimated by MLE, it will be useful to illustrate some of the important points for the implementation of MMD estimators. In the first instance, we consider the M-closed case where the data consists of samples $\{y_j\}_{j=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$ where $\mathbb{Q} = \mathbb{P}_{\theta*}$ and the kernel is given by $k(x, y; l) = \phi(x; y, l^2)$, where $\phi(x; y, l^2)$ is the probability density function of a Gaussian $\mathcal{N}(y, l^2 I_{d \times d})$.

Proposition 3 (Asymptotic Variance for Gaussian Location Models). *Consider the minimum MMD estimator for the location θ of a Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_{d \times d})$ using a Gaussian kernel $k(x, y) = \phi(x; y, l^2)$, then the estimator $\hat{\theta}_m$ has asymptotic variance given by*

$$C = \sigma^2((l^2 + \sigma^2)(3\sigma^2 + l^2))^{-\frac{d}{2}-1}(l^2 + 2\sigma^2)^{d+2} I_{d \times d}. \tag{6}$$

The Fisher information for this model is given by $1/\sigma^2 I_{d \times d}$, and so in the regime $l \to \infty$ we recover the efficiency of the MLE, so that the Cramer-Rao bound in Theorem 3 is attained. On the other hand, for finite values of l, the minimum MMD estimator will be less efficient than the MLE. For $l \to \infty$, the asymptotic variance is $O(1)$ with respect to d, but we notice that the asymptotic variance is $O(\alpha^{d+2})$ as $l \to 0$, where $\alpha = 2/\sqrt{3} \approx 1.155 > 1$. This demonstrates a curse of dimensionality in this regime. This transition in behaviour suggests that there is a critical scaling of l with respect to d which results in asymptotic variance independent of dimension.

Proposition 4 (Critical Scaling for Gaussian Location Models). *Consider the minimum MMD estimator for the location θ of a Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_{d \times d})$ using a single Gaussian kernel $k(x, y) = \phi(x; y, l^2)$ where $l = d^\alpha$. The asymptotic variance is bounded independently of dimension if and only $\alpha \geq 1/4$.*

As previously mentioned, it has been demonstrated empirically that choosing the bandwidth according to the median heuristic results in good performance in the context of MMD hypothesis tests [Reddi et al., 2015, Ramdas et al., 2015]. These works note that the median heuristic yields $l = O(d^{1/2})$, which lies within the dimension independent regime in Proposition 4. Our CLT therefore explains some of the favourable properties of this choice.

Clearly, the choice of lengthscale can have a significant impact on the efficiency of the estimator, but it can also impact other aspects of the problem. For example, the loss landscape of the MMD, and hence our ability to perform gradient-based optimisation, is severely impacted by the choice of kernel. This is illustrated in Figure 1 (top left) in $d = 1$, where choices of lengthscale between 5 and 25 will be preferable for gradient-based optimisation routines since they avoid large regions where the loss function will have a gradient close to zero. Using a mixture of kernels with a range of lengthscale regimes could help avoid regients of near-zero gradient and hence be generally desirable for the gradient-based optimization. A third aspect of the inference scheme which is impacted by the lengthscale is the robustness. We can quantify the influence of the kernel choice on robustness using the influence function. Similar plots for different classes of kernels can be found in the Appendix in Figures 8, 9 and 10.

Proposition 5 (Influence Function for Gaussian Location Models). *Consider the MMD estimator for the location θ of a Gaussian model $\mathcal{N}(\theta, \sigma^2 I_{d \times d})$ based on a Gaussian kernel $k(x, y) =$*

$\phi(x; y, l^2)$. *Then the influence function is given by:*

$$IF_{MMD}(\mathbb{P}_\theta, z) = 2 \left(\frac{l^2 + 2\sigma^2}{l^2 + \sigma^2} \right)^{\frac{d}{2}+1} \exp\left(-\frac{\|z - \theta\|_2^2}{2(l^2 + \sigma^2)} \right) (z - \theta).$$

Note that the asymptotic variance (6) is minimised by taking l arbitrarily large. Despite this, in practice we do not want to choose l to be larger than necessary as this will poorly influence the robustness of the estimator. Clearly, for the location model, we see that l controls the sensitivity of our estimators. For every finite l, we have the following uniform bound for the influence function

$$\sup_{z \in \mathbb{R}^d} |IF_{MMD}(\mathbb{P}_{\theta^*}, z)| = 2e^{-1/2} \sqrt{l^2 + \sigma^2} \left(\frac{l^2 + 2\sigma^2}{l^2 + \sigma^2} \right)^{\frac{d}{2}+1}.$$

Taking $l \to \infty$ we have $\text{IF}_{MMD}(\mathbb{P}_{\theta^*}, z) \to (\theta^* - z)$, thus losing robustness in the limit. As with asymptotic variance, the sensitivity will depend exponentially on dimension when l is small. Contrary to intuition, the uniform influence function minimum will not be attained when l approaches zero, but rather at an intermediate point, when $l^2 = d\sigma^2$, after which the influence to contamination will increase as $l \to \infty$. The middle plot in Figure 1 (top) illustrates the effect of kernel bandwidth on robustness. The figure plots the l_1 error between the estimated parameter $\hat{\theta}_m$ (for n) based on a polluted data sample $\mathbb{Q}(\mathrm{d}x) = (1 - \epsilon)\phi(x; 0, 1)\mathrm{d}x + \epsilon\delta z$, for some $z \in \mathbb{R}^d$ where $\epsilon = 0.2$. While the estimator is qualitatively robust, for higher kernel bandwidths, the estimator will undergo increasingly large excursions from $\theta^* = 0$ as the position of the contaminent point z moves to infinity. The second plot demonstrates the behaviour of the estimators as the pollution strength ϵ is increased from 0 to 1 and $z = (10, \ldots, 10)^\top$. We observe that for small kernel bandwidths, the estimator undergoes a rapid transition around $\epsilon = 0.5$. However, as the lengthscale is increased the estimator becomes increasingly sensitive to distance sample points to the extent that the error grows linearly with respect to ϵ. Interestingly, additional experiments presented in Figure 11 of the Appendix indicate that Wasserstein-based estimators may not be robust.

4.3 Gaussian Scale Model

The second model we consider is a d-dimensional isotropic Gaussian distribution $\mathcal{N}(\mu, \exp(2\theta)I_{d \times d})$ with known location parameter $\mu \in \mathbb{R}^d$. Since the asymptotic variance does not depend on any location parameters of $\mathbb{U}$, we will assume without loss of generality that the base measure $\mathbb{U}$ is a d-dimensional Gaussian with mean zero and identity covariance matrix, and that $G_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is defined by $G_\theta(u) = \exp(\theta)u$. For simplicity, we assume that we are in the M-closed situation, where the true data distribution $\mathbb{Q}$ is given by $\mathbb{P}_{\theta^*}$ and the kernel is $k(x, y; l) = \phi(x; y, l^2)$. For this model, the conclusions in terms of efficiency, robustness and loss landscape are similar to those of the Gaussian location model. For example, we can once again compute the asymptotic variance of the CLT:

Proposition 6 (Asymptotic Variance for Gaussian Scale Models). *Consider the minimum MMD estimator for the scale θ of a Gaussian distribution $\mathcal{N}(\mu, \exp(\theta)I_{d \times d})$ using a Gaussian RBF kernel $k(x, y; l) = \phi(x; y, l^2)$. The asymptotic variance of the minimum MMD estimator $\hat{\theta}_m$ satisfies $C^l = g^{-1}(\theta^*)\Sigma g^{-1}(\theta^*)$ where the metric tensor at θ^* satisfies*

$$g(\theta^*) = (2\pi)^{-d/2} \left(l^2 + 2e^{2\theta^*} \right)^{-d/2} d^2 K\left(d, l, e^{2\theta^*} \right),$$
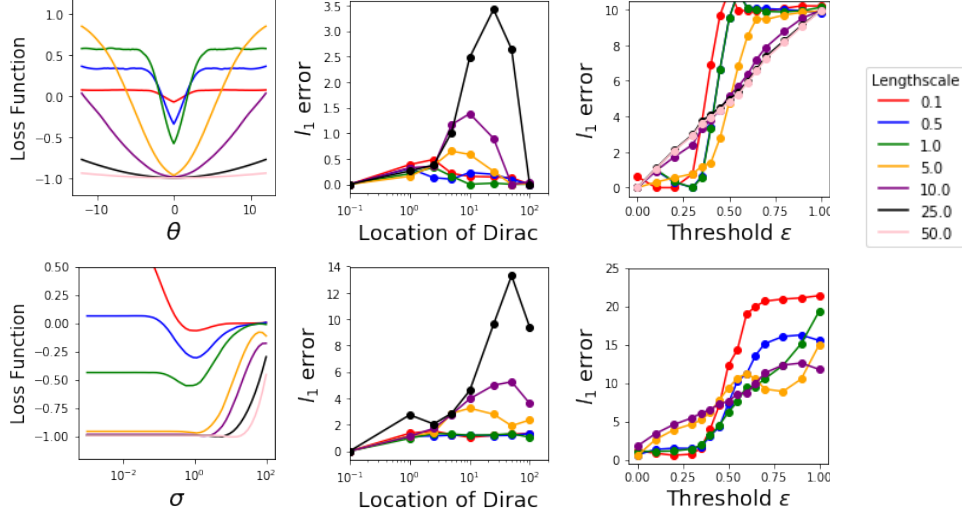
16

Figure 1: *Gaussian location and scale models - Performance of the Gaussian RBF kernel (for n, m large).* The top plots refer to the Gaussian location model, whilst the bottom plots refer to the Gaussian scale model. *Left:* Comparison of the loss landscape for various lengthscale values in $d = 1$. *Center:* Robustness problem with varying location x for the Dirac but threshold fixed to $\epsilon = 0.2$ in $d = 1$. *Right:* Robustness problem with varying threshold but fixed location for the Dirac at $x = 10$ in $d = 1$.

for a $K(d, l, s)$ is bounded with respect to d, l, and s and $K(d, 0, s) = \frac{1}{4}(1 + 2d^{-1})$; and

$$\Sigma = (2\pi)^{-d} d^2 e^{4\theta^*} \left(e^{\theta^*} + l^2\right)^{-2} \left(C_1 \left(l^2 + 3e^{2\theta^*}\right)^{-d/2} \left(l^2 + e^{2\theta^*}\right)^{-d/2} + C_2 \left(l^2 + 2e^{2\theta^*}\right)^{-d}\right),$$

where C_1 and C_2 are bounded uniformly with respect to the parameters. Asymptotically, the asymptotic variance behaves as $C^l \sim (2/\sqrt{3})^d/d^2$ for $l \ll 1$ and $C^l \sim l^4/d^2$ for $l \gg 1$.

This result indicates that the asymptotic variance grows exponentially with dimension as l small. In the other extreme, for l going to infinity results in an asymptotic variance which is bounded independent of dimension. In fact, the following result characterises the choice of lengthscale required for dimension independent efficiency.

Proposition 7 (Critical Scaling for Gaussian Scale Models). *Consider the minimum MMD estimator for the scale θ of a Gaussian distribution $\mathcal{N}(\mu, \exp(\theta)I_{d \times d})$ with a single Gaussian kernel $k(x, y) = \phi(x; y, l^2)$ where $l = d^\alpha$. The asymptotic variance is bounded independently of dimension if and only if $\alpha \geq 1/4$.*

This scaling is the same as for the Gaussian location model, indicating that a more general result on critical scaling for MMD estimators may exists. We reserve this issue for future work. Once again, we notice (Figure 1, bottom left) that the choice of lengthscale has a significant impact on the loss landscape. However, an interesting point is that values of the lengthscale which render the loss landscape easily amenable to gradient-based optimisation are different in the two cases.
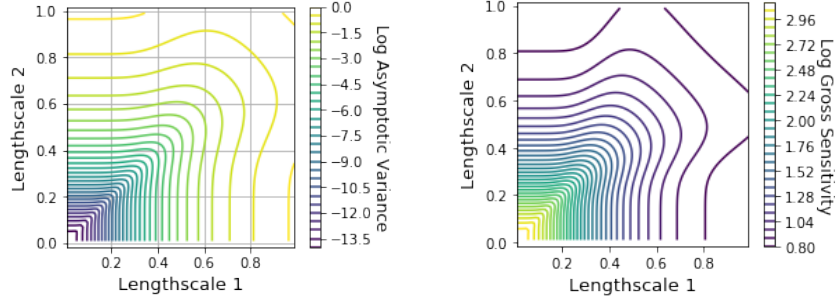
17

Figure 2: *Gaussian location model - Efficiency and Robustness for a Mixture of Kernels.* Asymptotic variance and gross sensitivity for minimum MMD estimators of the Gaussian location model as a function of l_1 and l_2 for a mixture of squared exponential kernels: $k(x, y) = \exp(-\|x - y\|_2^2/2l_1^2) + \exp(-\|x - y\|_2^2/2l_2^2)$.

Numerical experiments clearly indicate that the choice of lengthscale may need to be adapted based on the parameters of interest. The lengthscale has, once again, a significant impact on the robustness of the estimator, as demonstrated in the following result.

Proposition 8 (Influence Function for Gaussian Scale Models). *Consider the minimum MMD estimator for the scale θ of a Gaussian model $\mathcal{N}(\mu, \exp(\theta)I_{d \times d})$ based on a single Gaussian kernel $k(x, y) = \phi(x; y, l^2)$. The influence function associated with this estimator is:*

$$IF_{MMD}(\mathbb{P}_\theta, z) = \left(\frac{l^2 + 2e^{2\theta^*}}{l^2 + e^{2\theta^*}} \right)^{\frac{d}{2}+2} \frac{\left(l^2 + e^{2\theta^*} - z^2 \right)}{d(d+2)e^{2\theta^*}} \exp\left(-\frac{z^2}{2\left(l^2 + e^{2\theta^*} \right)} \right).$$

In particular, for every finite l we have that

$$\sup_{z \in \mathbb{R}^d} |IF_{MMD}(\mathbb{P}_{\theta^*}, z)| = \frac{4e^{-3/2} \left(l^2 + 2e^{2\theta^*} \right)}{d(d+2)e^{2\theta^*}} \left(\frac{l^2 + 2e^{2\theta^*}}{l^2 + e^{2\theta^*}} \right)^{\frac{d}{2}+1},$$

independently of z, so that l controls the sensitivity of the estimator. Once again, we see exponential dependence on dimension for l small, with the minimum uniform influence at an intermediate point, with the dependence increasing as $l \to \infty$.

4.4 Using Mixtures of Gaussian Kernels

In Li et al. [2015], Ren et al. [2016], Sutherland et al. [2017] it was observed empirically that using mixtures of distributions offers advantageous performance compared to making single choices. In particular, it circumvents issues arising from gradient descent due to *vanishing gradients*, which can occur if the lengthscale of the kernel chosen to be too small, as can be seen in Figure 1. While multiple kernels offer advantage for gradient descent, we aim to understand where mixture kernels offer any advantages in terms of asymptotic efficiency and robustness. Focusing on the Gaussian location model case, we have the following result.

Proposition 9 (Efficiency and Robustness with Mixture Kernels for Gaussian Location). *Consider the minimum MMD estimator for the location of a Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_{d\times d})$ using a Gaussian mixture kernel $k(x, y) = \sum_{s=1}^{S} \gamma_s \phi(x; y, l_s^2)$. Then the minimimum MMD estimator has asymptotic variance given by*

$$\sigma^2 \frac{\sum_{s=1}^{S} \sum_{s'=1}^{S} \gamma_s \gamma_{s'} \left((l_s^2 + \sigma^2)(l_{s'}^2 + \sigma^2) + \sigma^2(2\sigma^2 + l_s^2 + l_{s'}^2)\right)^{-\frac{d}{2}-1}}{\left(\sum_{s=1}^{S} \gamma_s (l_s^2 + 2\sigma^2)^{-\frac{d}{2}-1}\right)^2} I_{d\times d}. \tag{7}$$

Furthermore, the influence function is given by:

$$IF_{MMD}(z, \mathbb{P}_\theta) = \frac{2 \sum_{s=1}^{S} \gamma_s (l_s^2 + \sigma^2)^{-\frac{d}{2}-1} \exp\left(-\frac{\|z-\theta\|_2^2}{2(l_s^2+\sigma^2)}\right)(z-\theta)}{\sum_{s=1}^{S} \gamma_s (l_s^2 + 2\sigma^2)^{-\frac{d}{2}-1}}.$$

In Figure 2 we plot the log asymptotic variance and log gross sensitivity for the Gaussian location model based on a mixture kernel composed of two Gaussian kernels with lengthscales l_1 and l_2. What is interesting to note that there are choices of (l_1, l_2) which give rise to higher efficiency and robustness than their individual counterparts. Indeed, for example, choosing $l_1 = 0.8$ then the asymptotic variance will be minimised when $l_2 \approx 0.6$, although this choice will reduce bias-robustness. This figure appears to support the claim that mixture kernels can also provide increased performance beyond assisting gradient descent, and merits further investigation.

5 Numerical Experiments

In this final section, we examine the impact of the choice of kernel on several applications. In particular, we highlight the importance of working with estimators which are robust to model misspecification. We start with two applications which are popular test-beds for inference for intractable generative models: the g-and-k distribution and a stochastic volatility model. We then move on to a problem of parameter inference for systems of stochastic differential equations, where we consider a parameter-prey model and a multiscale model. These examples allow us to demonstrate the advantage of our natural gradient descent algorithm, and the favourable robustness properties of the estimators.

5.1 G-and-k distribution

A common synthetic model in the literature on generative models is the g-and-k distribution [Bernton et al., 2019, Prangle, 2017]. For this model, we only have access to the quantile function $G_\theta : [0, 1] \to \mathbb{R}$ (also called inverse cumulative distribution function) given by:

$$G_\theta(u) := a + b\left(1 + 0.8\frac{(1 - \exp(-c\Phi^{-1}(u; 0, 1)))}{(1 + \exp(-c\Phi^{-1}(u; 0, 1)))}\right)\left(1 + (\Phi^{-1}(u; 0, 1))^2\right)^k \Phi^{-1}(u; 0, 1)$$

where $\Phi^{-1}(u; 0, 1)$ refers to the u'th quantile of the standard normal distribution. The parameter of interest is $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ where $\theta_1 = a$ controls location, $\theta_2 = b$ controls scale, $\theta_3 = c$ controls skewness and $\theta_4 = \exp(k)$ controls kurtosis. Although this is a model defined on a
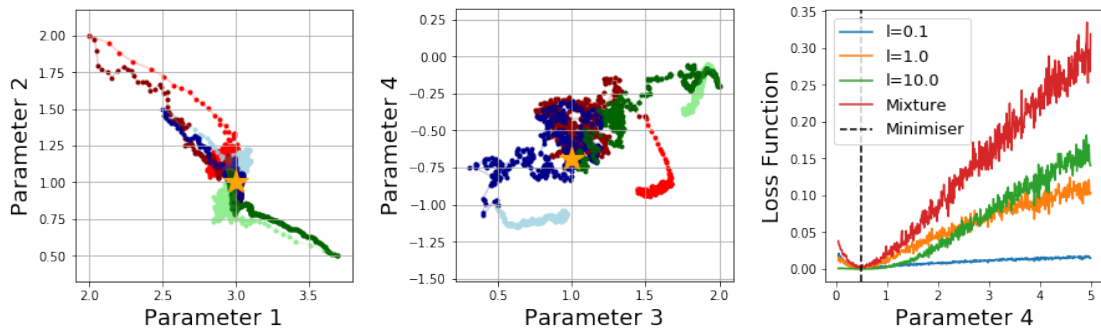
Figure 3: *Inference for the parameters of the g-and-k distribution using a maximum mean discrepancy estimator. Left & Center:* Several runs of a stochastic gradient descent (light blue, light red and light green) and a stochastic natural gradient descent (dark blue, dark red and dark green) algorithm on the MMD loss function with Gaussian RBF kernel with lengthscale $l = 2$. The black dot corresponds to the minimiser. *Right:* Estimate of the MMD loss function around the minimum as a function of θ_4 for a Gaussian RBF kernel with varying choices of lengthscales and a mixture of all the Gaussian kernels.

one-dimensional space, the four parameters allow for a very flexible family of distributions. A rescaling of the last parameter is used to avoid instabilities. Since the quantile function is available, we can easily simulate from this model using inverse transform sampling.

We study the behaviour of the MMD estimators for this model in Figure 3. For the left and center plots, we used both stochastic gradient descent and stochastic gradient descent with preconditioner to obtain an estimate of $\hat{\theta}_m$. We used a constant step-size for both algorithms (tuned for good performance) and ran each algorithm for 500 iterations. The data is of size $m = 30000$ but we used minibatches of size 200, the simulated data was of size $n = 200$, the kernel was Gaussian RBF with lengthscale $l = 2$ and $\theta^* = (3, 1, 1, -\log(2))$. The large number of data points is used to guarantee that the minimiser can be recovered. We notice that both the stochastic gradient descent and stochastic natural gradient descent are able to recover θ_1^* and θ_2^* for a variety of initial conditions in the neighbourhood of the minimiser. On the other hand, as observed in the center plot, the stochastic gradient descent algorithm is very slow for θ_3^* and θ_4^*, whereas the natural stochastic gradient descent algorithm is able to recover both of these parameters in a small number of steps. This clearly highlights the advantage of the rescaling of the parameter space provided by the preconditionner based on the geometry induced by the information metric.

We also plot the MMD loss function as a function of θ_4 in the neighborhood of θ^*. The estimator is sensitive to the choice of lengthscale: in the case where we use a Gaussian RBF kernel, a lengthscale smaller of equal to $l = 0.1$ or greater or equal to $l = 10$ led to a loss function which is flat on a wide range of the space. In those cases, it will be difficult to obtain an accurate estimate of the parameter due to the noise in our estimates of the gradient. On the other hand, a lengthscale of $l = 1$ allows us to provide more accurate results. Furthermore, the use of a mixture of all of these kernels also allows us to obtain accurate results without having to manually tune the choice of lenghtscale.

20

5.2 Stochastic Volatility Model with Gaussian and Cauchy Noise

Our second model is a stochastic volatility model [Kim et al., 1998], popular in the econometrics literature as a model of the returns on assets over time. The model can be simulated from by sampling the first hidden variable $h_1 \sim \mathcal{N}(0, \sigma^2/(1 - \phi^2))$ representing the initial volatility, then following the following set of equations:

$$
\begin{aligned}
h_t &= \phi h_{t-1} + \eta_t, &\quad \eta_t &\sim \mathcal{N}(0, \sigma^2), \\
y_t &= \epsilon_t \kappa \exp(0.5 h_t), &\quad \epsilon_t &\sim \mathcal{N}(0, 1).
\end{aligned}
$$

where y_t is the mean corrected return on holding an asset at time t, and h_t the log-volatility at time t. The $\{y_t\}_{t=1}^T$ are observed data and $\{h_t\}_{t=1}^T$ are unobserved latent variables. This is therefore a generative model with parameters (ϕ, κ, σ), which we reparameterised with $\theta_1 = \log((1 + \phi)/(1 - \phi)), \theta_2 = \log \kappa, \theta_3 = \log(\sigma^2)$ to avoid numerical issues so that we want to recover $\theta = (\theta_1, \theta_2, \theta_3)$. The data dimension is $d = T$ and the parametric dimension is $p = 3$. The likelihood of these models is usually not available in closed form due to the presence of latent variables and hence given by $p(y_1, \ldots, y_T | \theta) = \int p(y_1, \ldots, y_T | h_1, \ldots, h_T, \theta) p(h_1, \ldots, h_T | \theta) \mathrm{d} h_1 \ldots \mathrm{d} h_T$ which is a high-dimensional intractable integral. Alternative approaches based on quasi-likelihood estimation or expectation-maximisation can be considered, but the approximation obtained may be unreliable. Furthermore, it may be preferable to make use of minimum MMD estimators since these will allow for robust inference, which is not the case for alternative approaches.

In our experiments, we choose $T = 30$ and considered inference with minimum MMD estimators with Gaussian kernels. Initially, we considered the M-closed case and generated $m = 20000$ data points for $\theta^* = (0.98, 0.65, 0.15)$, which we then tried to infer by minimising the MMD loss function with a wide range of kernels. For the experimental results, we used stochastic gradient descent and stochastic natural gradient descent with minibatches of size 2000, and used $n = 45$. Results in Figure 4 (top) demonstrate that our natural gradient algorithm is able to recover the parameters in around five thousand iterations whereas the gradient descent algorithm isn't close to convergence after 30000 steps. Note that even though the dimension $d = 30$, the parameter space has dimension $p = 3$ so that the additional computational cost of the preconditioner is negligeable for this problem (and completely dwarfed by the cost of the generator).

We then considered the M-open case, and introduced misspecification by simulating the ϵ_t values using IID realisations of a Cauchy distribution with location parameter 0 and scale parameter $\sqrt{2/\pi}$. This distribution has the same median as the Gaussian distribution, and their probability density functions match at that point, but the Cauchy has much fatter tails. The results of these experiments are available in Figure 4, and in each case we repeated the experiments with 4 different stepsize choices and plot the best result. In the well-specified case, we notice that the natural gradient descent algorithm is able to take advantage of the local geometry of the problem and converges to θ^* in a small number of iterations. Further experiments with a larger range of kernels is available in Appendix D.3, but the mixture kernel tended to work best.

In the misspecified case, we notice (as expected) that while none of the minimum MMD estimators is able to recover the true value of θ^*, but that the inferred results remain stable, i.e. close to the truth. The choice of kernel has a clear impact on the output. For Gaussian RBF kernels with lengthscales $l = 1$ or $l = 5$, the loss function is too flat for gradient descent and we are not able to move much from the initial parameter. For larger values of the lengthscale (e.g.
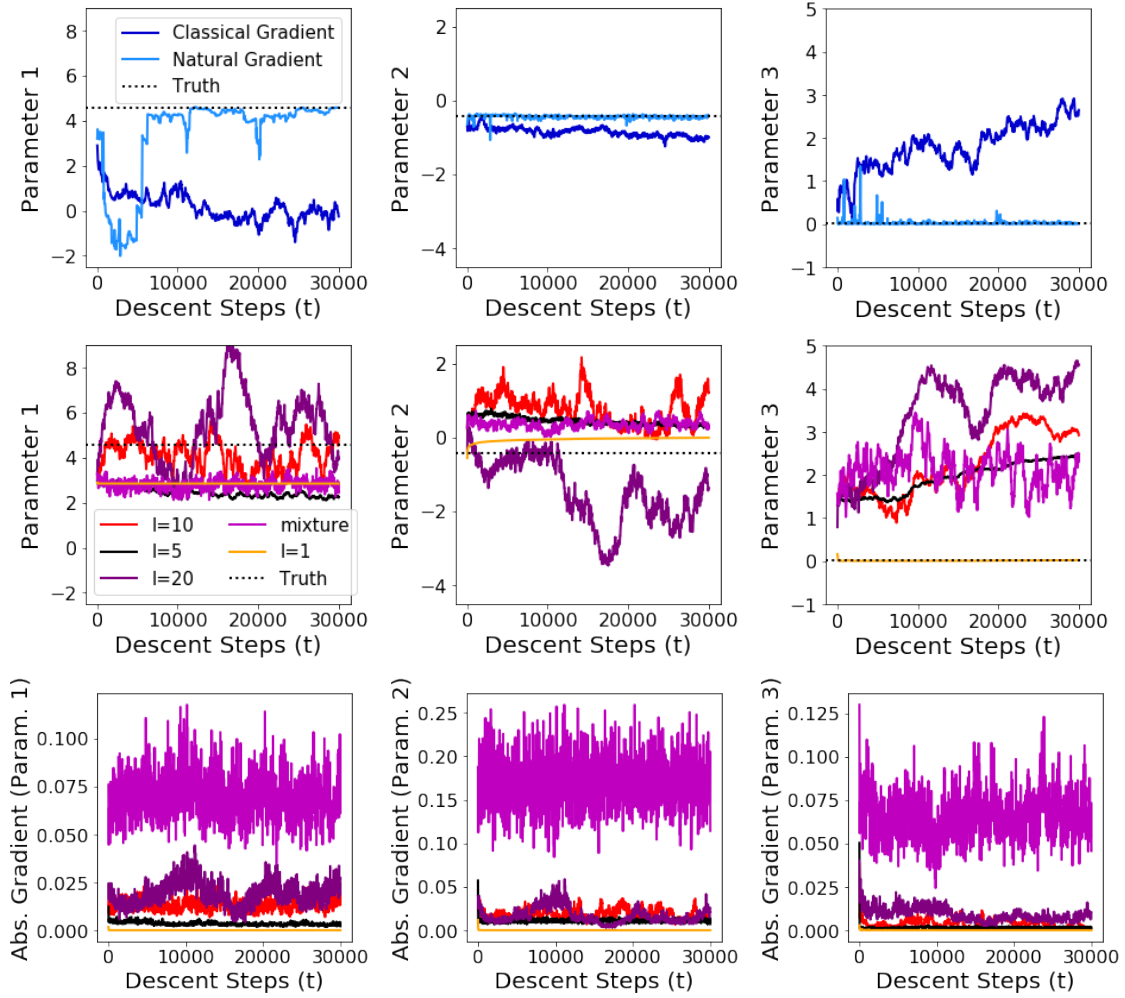
Figure 4: *Inference for stochastic volatility models. Top:* Well-specified case - Gradient descent and natural gradient descent on the MMD loss function with a mixture of Gaussian RBF kernels with lengthscales $1, 5, 10, 20, 40$. *Bottom:* Misspecified case - Gradient descent on the MMD loss function with a variety of kernels including a Gaussian RBF kernel with lengthscales 1, 5, 10, 20, as well as a mixture of all these kernels and a Gaussian RBF kernel with lengthscale $l = 40$.

$l = 10, 20, 40$) and for the mixture kernel, we are able to use gradient-based optimisation but that it demonstrates increased sensitivity to model mispecification. We note that the single Gaussian RBF kernels with large lengthscale are able to learn θ_1^* well in the sense that there is negligeable bias as compared to θ_2^* and θ_3^*. This is likely due to the improvement in bias robustness expected for kernels with large lengthscale.

5.3 Inference for Systems of Stochastic Differential Equations

For our third set of experiments, we use minimum MMD estimation to infer the initial condition and parameters for coupled systems of stochastic differential equations (SDEs). In general, will will consider a d-dimensional Itô stochastic differential equation of the form

$$dX_t = b(X_t; \theta_1) \, dt + \sigma(X_t; \theta_1) \, dW_t, \tag{8}$$

where $b : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \Theta \to \mathbb{R}^{d \times k}$, W_t is a k dimensional standard Brownian motion and with initial value $X_0 = \theta_2$ and where $(\theta_1, \theta_2) \in \Theta$ is a vector of unknown parameters to be determined. We assume that for each θ there is a unique solution to (8) which depends continuously on the initial condition.

For any fixed θ, provided we can simulate $X(t)$, at points $0 = t_0 < t_1 < \ldots < t_K = T$, then we can consider the generative model defined by $\mathbb{P}_\theta = G_\theta^\# \mathbb{U}$, where $\mathbb{U}$ is the Wiener path measure for a k dimensional standard Wiener process on $C[0, T]$ and $G_\theta = O_\theta \circ I_\theta$, where $I_\theta : C[0, T] \to C[0, T]$ is the Itô map, transforming the Wiener process to the solution $X(\cdot)$ of the SDE. Here, O_θ is an observation operator, for example mapping $w \in C[0, T]$ to $(w(t_1), \ldots, w(t_K))$ or any other smooth functional of the path which depends smoothly on θ. Note that it is trivial to incorporate observational noise and volatility parameters into the observation operator.

To perform MMD gradient descent for this model we must calculate the gradient of the forward map with respect to the parameters θ. Pathwise derivatives of the solution of (8) with respect to initial conditions and coefficient parameters are well established [Kunita, 1997, Gobet and Munos, 2005, Friedman, 2012] and are detailed in the following result; see also Tzen and Raginsky [2019] for a similar result arising in a similar context.

Proposition 10. *[Kunita, 1997, Theorem 2.3.1] Suppose that the drift $b(x; \theta_1)$ and diffusion tensor $\sigma(x; \theta_1)$ are Lipschitz with Lipschitz derivatives with respect to x and θ_1. Then the pathwise derivative of X_t with respect to the parameters θ_1 is given by the solution of the Ito process,*

$$d\left(\nabla_{\theta_1} X_t\right) = \left(\nabla_x b(X_t; \theta_1)\nabla_{\theta_1} X_t + \nabla_{\theta_1} b(X_t; \theta_1)\right) \, dt + \left(\nabla_x \sigma(X_t; \theta)\nabla_{\theta_1} X_t + \nabla_{\theta_1} \sigma(X_t; \theta_1)\right) \, dW_t$$

with initial condition $\nabla_{\theta_1} X_0 = \mathbf{0}$ and the derivative of X_t with respect to θ_2 is given by

$$d\left(\nabla_{\theta_2} X_t\right) = \left(\nabla_x b(X_t; \theta_1)\nabla_{\theta_2} X_t\right) \, dt + \left(\nabla_x \sigma(X_t; \theta)\nabla_{\theta_2} X_t\right) \, dW_t,$$

where $\nabla_{\theta_2} X_0 = I$. In particular, given a differentiable function F of $X_{t_0}, \ldots, X_{t_K}$,

$$\nabla_{\theta_i} \mathbb{E}\left[F(X_{t_0}, \ldots, X_{t_K})\right] = \mathbb{E}\left[\sum_{k=1}^{K} \nabla_{x_k} F(X_{t_0}, \ldots, X_{t_K})\nabla_{\theta_i} X_{t_k}\right], \quad \text{for } i = 1, 2.$$

Before moving on to the experiments, we note that Abbati et al. [2019] proposed an alternative method, performing Gaussian process-based gradient matching for ODEs and SDEs with additive noise, by using MMD to fit a GP process inferred from the data to the SDE. However, the approach we propose permits parametric estimation for more general SDEs and noise models.
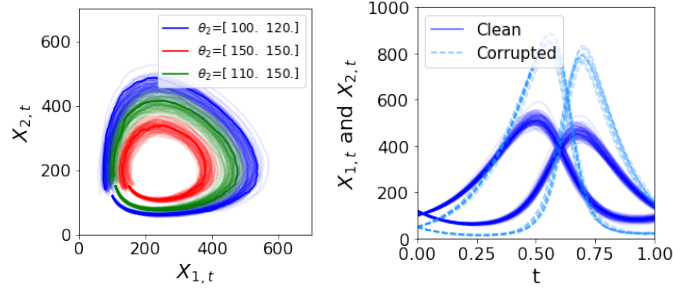
Figure 5: *Inference for the initial conditions of a Lotka-Volterra model with noisy dynamics.* *Left:* $n = 100$ realisations from the coupled stochastic differential equations for several initial conditions. *Right:* $n = 100$ realisations used for inference, including 90 realisations from the correct model and 10 which are corrupted.

5.3.1 Noisy Lotka-Volterra Model with Unknown Initial Conditions

As an example, we consider the stochastic Lotka-Volterra model [Volterra, 1926], which consists of a pair of nonlinear differential equations describing the evolution of two species through time:

$$d\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} \theta_{11}X_{1,t} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \theta_{12}X_{1,t}X_{2,t} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} \theta_{13}X_{2,t} \right] dt$$

$$+ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sqrt{\theta_{11}X_{1,t}}dW_t^{(1)} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \sqrt{\theta_{12}X_{1,t}X_{2,t}}dW_t^{(2)} + \begin{pmatrix} 0 \\ -1 \end{pmatrix} \sqrt{\theta_{13}X_{2,t}}dW_t^{(3)},$$

where the initial conditions $\theta_2 = (X_{1,0}, X_{2,0})$ are unknown, but the parameters $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ governing the dynamics are known. While exact sampling methods for diffusions exist, see Beskos and Roberts [2005], for simplicity we shall employ an inexact Euler-Maruyama discretisation, choosing the step size sufficiently small to ensure stability of the discretisation. We choose the "true" initial condition to be deterministic with value $\theta_2^* = (X_{1,0}, X_{2,0})$. We fix a-priori the time horizon to $T = 1$ and the parameters governing the equation to $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13}) = (5, 0.025, 6)$. In this case, $p = 2$, $d = 2$ and n tends to be small (in the tens or hundreds). We consider the case where $n = 50$.

Typical realisations for the system of coupled stochastic differential equations can be found in Figure 5 (left) for several values of the initial conditions. As we would expect, the closer the initial conditions, the closer the realisations of stochastic differential equations will be. This clearly motivates the use of minimum MMD estimators. We are particularly interested interested in the behaviour of the estimators as a proportion of the data is corrupted. In particular, we will consider the problem of inferring initial conditions $\theta_2^* = (100, 120)$ given realisations from this model which are corrupted by realisations from the model initialised at $\theta_2^\dagger = (50, 50)$. Realisations are provided in Figure 5 (right) for the case with 10% misspecification.

We expect this type of misspecification to lead to severe issues for non-robust inference algorithms, but the bias robustness of minimum MMD estimators allows us to provide reasonable estimates of the parameter. This can be seen in Figure 6 (left) where we plot estimates provided by MMD estimators for θ_{22} as a function of natural gradient steps for various proportion levels of
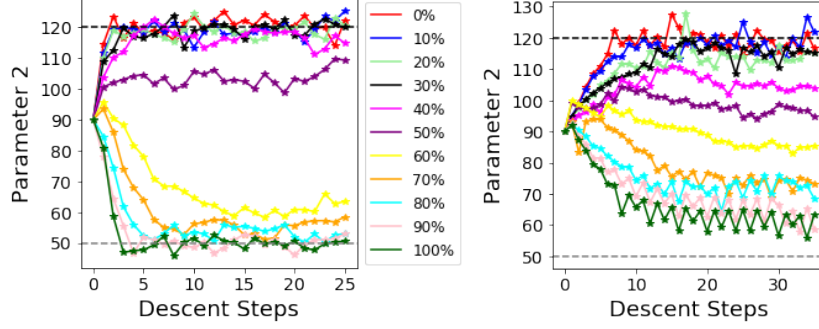
24

Figure 6: *Inference for the initial conditions of a Lotka-Volterra model with corrupted observations based on $m = 100$ realisations and $n = 50$ simulated data.* Each color correspond to a different percentage of corrupted observations. *Left:* Stochastic gradient descent steps for minimum Sinkhorn estimator with l_2 cost and $\epsilon = 1$ regularisation. *Right:* Stochastic gradient descent steps for minimum MMD estimator with Gaussian RBF kernel and lengthscale $l = 30$.

corruption. This is compared to the Sinkhorn algorithm of Genevay et al. [2018]. As can be seen, the MMD estimator can recover the truth for a large proportion of corrupted samples whereas Wasserstein-based estimators are very sensitive to corrupted data.

5.3.2 Parametric Inference for a System of SDEs with Multiple Scales

We consider a second example where we observe realisations of the following two-dimensional multiscale system

$$dX_t^\epsilon = \left(\frac{\sqrt{\theta_{12}}}{\epsilon} Y_t^\epsilon + \theta_{11} X_t^\epsilon \right) dt, \qquad dY_t^\epsilon = -\frac{1}{\epsilon^2} Y_t^\epsilon \, dt + \frac{\sqrt{2}}{\epsilon} dW_t, \tag{9}$$

where W_t is a standard Brownian motion, $0 < \epsilon \ll 1$ is a small length-scale parameter, $\theta_1 = (\theta_{11}, \theta_{12})$ are unknown parameters governing the dynamics, and the initial conditions θ_2 are known. Such systems arise naturally in atmosphere/ocean science [Majda et al., 2001], materials science [Weinan, 2011] and biology [Erban et al., 2006], and the inference of such stochastic multiscale systems has been widely studied, see [Pavliotis and Stuart, 2007, Krumscheid, 2018].

The process Y_t^ϵ is an Ornstein-Uhlenbeck process with vanishing autocorrelation controlled by ϵ. Formally, in the limit of $\epsilon \to 0$ it will behave as the derivative of Brownian motion. One can formulate minimum MMD problem for estimating the parameters θ_{11} and θ_{12}, appealing to Proposition 10 to compute the MMD gradient. However, a direct approach which involves integrating the SDEs in (9) multiple times is computationally infeasible, due to the fact that the simulation step-size would need to be commensurate to the small scale parameter ϵ. This motivates us to use a coarse grained model for estimating the unknown parameters. As $\epsilon \to 0$, the process $X_\cdot^\epsilon$ will converge weakly in $C[0, T]$ to a process $\overline{X}_\cdot$, given by the solution of the Itô SDE:

$$d\overline{X}_t = \theta_{11} \overline{X}_t + \sqrt{2\theta_{12}} \, dW_t, \quad t \in [0, T], \tag{10}$$
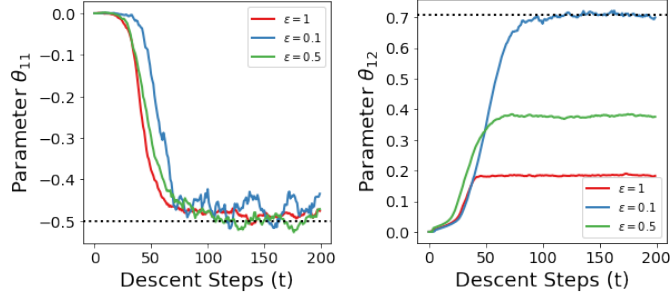
25

Figure 7: *Inference for the parameters of a two-scale stochastic process using a coarse grained model.* The plots show the convergence of the estimators to the truth values (dashed-lines) as the number of gradient descent steps increase, for data coming from (9).

see [Pavliotis and Stuart, 2008, Chapter 11]. As the coefficients of this SDE do not depend on the small scale parameter, we are able to generate realisations far more efficiently than for (9). We consider the minimum MMD estimator for θ_{11} and θ_{12} using (10) as a model. This introduced model misspecification of an interesting nature: for ϵ small, the path measures associated with (9) and (10) on $C[0, T]$ will be close with respect to the Levy-Prokhorov metric (which metrizes weak convergence) but not with respect to stronger divergences such as total variation or KL divergence. Indeed, the KL divergence between both measures will diverge as $\epsilon \to 0$. As MMD induces a coarser topology than the Levy-Prokhorov metric, we expect that the MMD estimators will be robust with respect to this misspecification for ϵ small, whereas maximum likelihood estimators are known to be biased in this case [Pavliotis and Stuart, 2007].

Suppose that we observe 100 realisations of (9) at discrete times $0.1, 0.2, \ldots, 1.0$ over a time horizon of $T = 1$ with known initial conditions $\theta_2 = (1.0, 0.0)$ with true values of the parameters given by $\theta_1^* = (-1/2, \sqrt{1/2})$. We construct a minimum MMD estimator for θ_1 using the coarse grained SDEs as a model. In this case, $p = 2$, $d = 1$. To simulate the coarse-grained model, we use an Euler-Maruyama discretisation with a step-size of 10^{-2}. We use natural gradient descent to minimise MMD, generating $n = 100$ synthetic realisations of the coarse SDE (10) per gradient step. In Figure 7 we plot the natural gradient descent trajectory for the estimators of θ_1 for $\epsilon = 1, 0.5, 0.1$, respectively. For $\epsilon = 1$, where we anticipate the misspecification to be high, the minimum MMD estimator converges to the true value of θ_{11}, but fails to recover the θ_{12} parameter (though remains within an order of magnitude). Taking ϵ smaller we observe that the accuracy of the estimators increases, indicating that the MMD estimators capture the weak convergence of $\{X_t^\epsilon, t \in [0, T]\}$ to $\{\overline{X}_t, t \in [0, T]\}$. We also note however that the volatility in the estimator for parameter θ_{11} is increasing as ϵ decreases, which suggests that the size of the simulated data (and perhaps also the size of the minibatches) must be increased as ϵ goes to 0 to maintain a constant mean square error.

26

6 Conclusion

This paper studied a class of statistical estimators for models for which the likelihood is unknown, but for which we can simulate realisations given parameter values. Our estimators are based on minimising U-statistic approximations of the maximum mean discrepancy squared. We provided several results on their asymptotic properties and robustness, as well as a novel natural-gradient descent algorithm for efficient implementation. As demonstrated first in our theory, then later in the experiments, the choice of reproducing kernel allows for great flexibility and can help us trade-off statistical efficiency with robustness.

This methodology clearly provides a rigorous approach to parametric estimation of complex black-box models for which we can only evaluate the forward map and its gradient. The natural robustness properties of these estimators make them a clear candidate for fitting models to engineering systems which are subject to intermittent sensor failures. Our theory also provides insights into the behaviour of MMD estimators for neural networks such as MMD GANs.

There are several directions in which this work could be extended. Firstly, we note this methodology can be readily applied to other continuum models such as ordinary differential equations and (stochastic) partial differential equations with noisy parameters. In these cases, adjoint based methods can be exploited to reduce the cost of computing gradients.

A second direction which is promising relates to model reduction or *coarse graining*, where a complex, very expensive model is replaced by a series of smaller models which are far cheaper to simulate. We believe that minimum MMD based estimators are an excellent candidate for effecting these coarse graining approaches thanks to their robustness properties.

Finally, we note that a drawback of this methodology is the poor scaling as a function of data-size. Indeed, the cost of computing MMD grows quadratically with data-size. This clearly motivates a second direction of research involving the use of cheaper approximate estimators for maximum mean discrepancy, such as [Chwialkowski et al., 2015].

7 Acknowledgements

References

G. Abbati, P. Wenk, S. Bauer, M. A. Osborne, A. Krause, and B. Scholkopf. AReS and MaRS - Adversarial and MMD-Minimizing Regression for SDEs. In *International Conference on Machine Learning*, pages 1–10, 2019.

S.-I. Amari. *Differential Geometrical Methods in Statistics*. Springer-Verlag, 1987.

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

M. Arbel, D. J. Sutherland, M. Binkowski, and A. Gretton. On gradient regularizers for MMD GANs. In *Neural Information Processing Systems*, pages 6700–6710, 2018.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, 1978.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76:1298–1302, 2006.

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. CRC Press, 2011.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162:2025–2035, 2002.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(2):192–236, 1974.

A. Beskos and G. O. Roberts. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representation*, 2018.

L. Bottou, M. Arjovsky, D. Lopez-Paz, and M. Oquab. Geometrical insights for implicit generative modeling. *Braverman Readings in Machine Learning: Key Ideas from Inception to Current State*, pages 229–268, 2017.

F-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? (with discussion). *Statistical Science*, 34(1):1–22, 2019.

D. Burago, I. Burago, and S. Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

E. Cameron and A. N. Pettitt. Approximate Bayesian Computation for astronomical model analysis: A case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012.

L. L. Campbell. An extended Cencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

N. N. Cencov. *Statistical Decision Rules and Optimal Inference*. Number 53. American Mathematical Society, 2000.

Y. Chen and W. Li. Natural gradient in Wasserstein statistical manifold. *arXiv:1805.08380*, 2018.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

A. Cuevas. Qualitative robustness in abstract inference. *Journal of Statistical Planning and Inference*, 18(3):277–289, 1988.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathe-*

matics, 59(1):77–93, 2007.

A. P. Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2): 169–183, 2014.

J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22(April 2013):133–288, 2013.

C. N. dos Santos, Y. Mroueh, I Padhi, and P. Dognin. Learning implicit generative models by matching perceptual features. *arXiv:1904.02762*, 2019.

B. A. Dubrovin, A. T. Fomenko, S. P. Novikov, and R. G. Burns. *Modern Geometry - Methods and Applications. Part I: The Geometry of Surfaces, Transformation Groups, and Fields*. Springer.

R. M. Dudley. *Real Analysis and Probability*. Chapman and Hall/CRC, 2018.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

M. L. Eaton. A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III*, pages 329–352, 1982.

R. Erban, I. G. Kevrekidis, and H. G. Othmer. An equation-free computational approach for extracting population-level behavior from individual-based models of biological dispersal. *Physica D: Nonlinear Phenomena*, 215(1):1–24, 2006.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic ABC. *Journal of the Royal Statistical Society B: Statistical Methodology*, 74(3):419–474, 2011.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

A. Friedman. *Stochastic Differential Equations and Applications*. Courier Corporation, 2012.

C. Frogner and T. Poggio. Approximate inference with Wasserstein gradient flows. *arXiv:1806.04542*, 2018.

C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. *arXiv:1707.07269*, 2017.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR 84*, pages 1608–1617, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2): 123–214, 2011.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

E. Gobet and R. Munos. Sensitivity analysis using Itô–Malliavin calculus and martingales, and application to stochastic optimal control. *SIAM Journal on Control and Optimization*, 43(5):

1676–1713, 2005.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel method for the two-sample problem. *Journal of Machine Learning Research*, 1(157):0–43, 2008.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, pages 673–681, 2009.

M. U. Gutmann and A. Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13: 307–361, 2012.

P. Hájek and M. Johanis. *Smooth Analysis in Banach Spaces*, volume 19. Walter de Gruyter GmbH & Co KG, 2014.

A. R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of mathematical statistics*, pages 293–325, 1948.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, 2009.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2006.

A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5):2499–2512, 2007.

R. G. Jarrett. Bounds and expansions for Fisher information when the moments are known. *Biometrika*, 71(1):101–113, 1984.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

R. Karakida, M. Okada, and S.-I. Amari. Adaptive natural gradient learning algorithms for unnormalized statistical models. *Artificial Neural Networks and Machine Learning - ICANN*, 2016.

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393, 1998.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

S. Krumscheid. Perturbation-based inference for diffusion processes: Obtaining effective models from multiscale data. *Mathematical Models and Methods in Applied Sciences*, 28(08):1565–1597, 2018.

H. Kunita. *Stochastic Flows and Stochastic Differential Equations*, volume 24. Cambridge Uni-

versity Press, 1997.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

S. Lang. *Fundamentals of Differential Geometry*, volume 191. Springer Science & Business Media, 2012.

E. Lehmann. Consistency and unbiasness of certain nonparametric tests. *Annals of Mathematical Statistics*, 22:165–179, 1951.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

W. Li and G. Montufar. Natural gradient via optimal transport. *Information Geometry*, 1(2): 181–214, 2018.

Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. In *Proceedings of the International Conference on Machine Learning*, volume 37, pages 1718–1727, 2015.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):66–82, 2017.

A. J. Majda, I. Timofeyev, and E. Vanden Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics*, 54(8):891–974, 2001.

J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

H. Masnadi-Shirazi. Strictly proper kernel scoring rules and divergences with an application to kernel two-sample hypothesis testing. *arXiv:1704.02578*, 2017.

C. McDiarmid. *On the Method of Bounded Differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. *arXiv:1610.03483v4*, 2016.

J. Moller, A. N. Pettitt, and R. Reeves. An efficient Markov Chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3718–3726, 2016.

G. Montúfar, J. Rauh, and N. Ay. On the Fisher metric of conditional probability polytopes. *Entropy*, 16(6):3207–3233, 2014.

A. Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.

N. Neumeyer. A central limit theorem for two-sample U-processes. *Statistics and Probability Letters*, 67(1):73–85, 2004.

D. Nolan and D. Pollard. U-Processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.

Y. Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2): 2930–2961, 2018.

A. Papadopoulos. *Metric Spaces, Convexity and Nonpositive Curvature*, volume 6. European Mathematical Society, 2014.

L. Pardo. *Statistical Inference Based on Divergence Measures*, volume 170. Chapman and Hall/CRC, 2005.

H. Park, S-I. Amari, and K. Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: approximate Bayesian computation with kernel embeddings. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, PMLR 51:398–407, 2015.

G. Pavliotis and A. Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer Science & Business Media, 2008.

G. A. Pavliotis and A. M. Stuart. Parameter estimation for multiscale diffusions. *Journal of Statistical Physics*, 127(4):741–781, 2007.

D. Prangle. gk: An R Package for the g-and-k and generalised g-and-h Distributions. *arXiv:1706.06889*, 2017.

A. Ramdas, S. J. Reddi, A. Singh, and L. Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high-dimensional two sample testing. *arXiv:1508.00655*, 2015.

S. J. Reddi, A. Ramdas, B. Poczos, A. Singh, and L. Wasserman. On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives. In *International Conference on Artificial Intelligence and Statistics*, pages 772–780, 2015.

Y. Ren, J. Li, Y. Luo, and J. Zhu. Conditional generative moment-matching networks. In *Advances in Neural Information Processing Systems*, pages 2928–2936, 2016.

H. Robbins and S. Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.

Y. Romano, M. Sesia, and E. J. Candès. Deep knockoffs. *arXiv:1811.06687*, 2018.

M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7088–7098, 2018.

B. K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11, 2010.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and . R G Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

M. S. Stein and J. A. Nossek. A pessimistic approximation for the Fisher information measure. *IEEE Transactions on Signal Processing*, 65(2):386–396, 2017.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact

spaces. *arXiv:1712.05279*, 2017.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *Proceedings of the International Conference on Learning Representation*, 2017.

B. Tzen and M. Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118 (2972):558–560, 1926.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli (to appear)*, 2017.

J. Weed and Q. Berthet. Estimation of smooth densities in Wasserstein distance. *arXiv:1902.01778*, 2019.

E. Weinan. *Principles of Multiscale Modeling*. Cambridge University Press, 2011.

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466 (7310):1102–1104, 2010.

E. Zawadzki and S. Lahaie. Nonparametric scoring rules. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI 2015)*, pages 3635–3641, 2015.

Supplementary Material for "Statistical Inference for Generative Models with Maximum Mean Discrepancy"

The supplementary materials are structured as follows. Section A provides further discussion on the geometry induced by MMD on parametric families of probability distributions, and in particular derives the corresponding metric tensor, gradient flow and geodesics. Section B contains all the proofs of results in the paper, including asymptotic results and results on robustness. Section C contains the derivation of important quantities for the Gaussian models. Finally, Section D contains further details on the numerical experiments.

A   Geometry of the MMD Statistical Manifold

In this appendix we complement Section 2 and provide additional details on the Riemmanian manifold induced by the MMD metric.

A.1   Identification of the Information Metric Tensor

Identifying $\mathbb{P}_\theta$ as the pushforward $G_\theta^\# \mathbb{U}$, we have:

$$\mathrm{MMD}^2(\mathbb{P}_\alpha || \mathbb{P}_\beta) = \int_\mathcal{U} \int_\mathcal{U} k(G_\alpha(u), G_\alpha(v)) \mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v) - 2 \int_\mathcal{U} \int_\mathcal{U} k(G_\alpha(u), G_\beta(v)) \mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v)$$
$$+ \int_\mathcal{U} \int_\mathcal{U} k(G_\beta(u), G_\beta(v)) \mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v)$$

Taking the derivative with respect to α and β, and noticing that:

$$\partial_{\beta^k} \partial_{\alpha^j} k(G_\alpha(u), G_\beta(v)) = \sum_{l,i} \partial_{2^l} \partial_{1^i} k(G_\alpha(u), G_\beta(v)) \partial_{\alpha^j} G_\alpha^i(u) \partial_{\beta^k} G_\beta^l(v)$$
$$= \left(\nabla_\alpha G_\alpha(u)^\top \nabla_2 \nabla_1 k(G_\alpha(u), G_\beta(v)) \nabla_\beta G_\beta(v) \right)_{jk}$$

which yields the expression for the information metric associated to the MMD^2 divergence. Let $\mathcal{H}$ be a Hilbert space viewed as a Hilbert manifold. As usual we identify the tangent spaces $T_p\mathcal{H} \cong \mathcal{H}$, and the Riemannian metric is $m(f, g) = \langle f, g \rangle$ for any $f, g \in \mathcal{H}$. Let $\Psi : S \to \mathcal{H}$ be a differentiable injective immersion (i.e., its derivative is injective), from a finite-dimensional manifold S. Then Ψ induces a Riemannian structure on S given by the pull-back Riemannian metric $g = \Psi^* m$. If x^i are local coordinates on S, and ∂_{x^i} is the associated local basis of vector fields, then the components of g are defined by

$$g_{ij} = g(\partial_{x^i}, \partial_{x^j}) = m\big(d\Psi(\partial_{x^i}), d\Psi(\partial_{x^j})\big),$$

where $d\Psi : TS \to T\mathcal{H}$ is the differential/tangent map (here TS denotes the tangent bundle of S, or, roughly, the set of vectors tangent to S). When S is an open subset of $\mathbb{R}^n$, since the Frechet partial derivative $\nabla_{x^j} \Psi(x)$ is the derivative of the function $t \mapsto \Psi(x^1, \ldots, x^{j-1}, t, x^{j+1}, \ldots, x^n)$, of the curve $t \mapsto (x^1, \ldots, x^{j-1}, t, x^{j+1}, \ldots, x^n)$ is precisely the curve tangent to the vector $\partial_{x^j}|_x$, we have $\nabla_{x^j} \Psi = d\Psi(\partial_{x^j})$ (see [Lang, 2012] page 28). Hence $g_{ij} = m\big(\nabla_{x^i} \Psi, \nabla_{x^j} \Psi\big)$.

Note that if Ψ is not an immersion, the pullback Riemannian metric will in general just be a degenerate quadratic form rather than a positive definite one.

Let S be a statistical manifold, i.e., $x \in S$ is associated to a probability measure P_x (we assume the map $x \mapsto P_x$ is a bijection). We can define a divergence on S by $D\big(P_\alpha, P_\beta\big) = \|\Psi(\alpha) - \Psi(\beta)\|^2$, which is the pull-back of the square-metric $(f, g) \mapsto \|f - g\|^2$ on $\mathcal{H}$ induced by the inner product. The corresponding information metric has components I_{ij} in a local coordinate chart given by

$$I_{ij} = -\frac{\partial}{\partial \alpha^k \partial \beta^j} D\big(p(\alpha), p(\beta)\big)|_{\alpha=\beta=\theta} = 2\partial_{\beta^j}\partial_{\alpha^k}\langle \Psi(\alpha), \Psi(\beta)\rangle|_{\alpha=\beta=\theta}.$$

Suppose now that $\mathcal{H}_k$ is a RKHS, and Ψ is defined as the mean-embedding. Then

$$\langle \Psi(\alpha), \Psi(\beta)\rangle = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) P_\alpha(\mathrm{d}x) P_\beta(\mathrm{d}y).$$

In particular if the measures in S can be written as either $P_\alpha = G_\alpha^\# \mu$, or $P_\alpha(\mathrm{d}x) = p_\alpha(x)\mu(\mathrm{d}x)$ for some fixed measure μ, then $\partial_{\beta^j}\partial_{\alpha^k}\langle \Psi(\alpha), \Psi(\beta)\rangle|_{\alpha=\beta=\theta} = \langle \partial_{\theta^j}\Psi(\theta), \partial_{\theta^k}\Psi(\theta)\rangle$ and we recover the pullback Riemannian metric.

A.2   Geodesics of the MMD metric

The following result summarises the properties of the geodesics induced by the MMD metric on $\mathcal{P}_k$.

Proposition 11 (The MMD Information Metric). *Suppose that k is a characteristic kernel with a bounded continuous derivative and that assumptions (i)-(iv) stated above hold. If the matrix $g(\theta) = (g_{ij}(\theta))_{i,j=1,\dots,p}$ is positive definite on Θ, then the MMD metric on $\mathcal{P}_k$ induces a Riemannian geometry (Θ, g) on Θ. The metric induced on Θ is given by*

$$d_{MMD}^2(\theta|\theta') = \inf_{\theta(t) \in C^1(0,1)} \left[\int_0^1 \dot{\theta}(t)^\top g(\theta)\dot{\theta}(t) dt : \theta(0) = \theta, \theta(1) = \theta' \right], \qquad (11)$$

for all $\theta, \theta' \in \Theta$. Geodesics in (Θ, g) are given by infimisers by (11) and satisfy the following system of ODEs Dubrovin et al.:

$$\dot{\theta}(t) - g^{-1}(\theta(t))S(t) = 0$$
$$\dot{S}(t) - \frac{1}{2}S(t)^\top \nabla_\theta g(\theta(t))^{-1} S(t) = 0. \qquad (12)$$

Sufficient conditions for g being positive definite need to be verified on a case by case basis. Since $(\mathcal{P}_k, \mathrm{MMD})$ is a length space [Papadopoulos, 2014, Burago et al., 2001], it follows immediately that geodesics in this metric is via teleportation of mass, i.e. a geodesic connecting $\mathbb{P}_1$ and $\mathbb{P}_2$ in $\mathcal{P}_k$ is defined by $\mathbb{P}_t = (1 - t)\mathbb{P}_1 + t\mathbb{P}_2, t \in [0, 1]$. This will not be the case for (Θ, g) as geodesics $\theta(t)$ must be constrained to ensure that $\mathbb{P}_{\theta(t)} \in \mathcal{P}_\Theta$.

B   Proofs of Main Results

In this appendix, we give the proofs of all lemmas, propositions and theorems in the main text.

B.1 Proof of Theorem 1

Before moving on to Theorem 1, we show the following result, which proves that the there is a uniform bound between the different versions of the MMD discrepancy. First, for convenience we define the following approximation to MMD between a measure $\mathbb{P}$ and a empirical measure $\mathbb{Q}^m(\mathrm{d}y) = \frac{1}{m}\sum_{i=1}^{m}\delta_{y_i}(\mathrm{d}y)$:

$$\mathrm{MMD}_U^2(\mathbb{P}||\mathbb{Q}^m) = \int_{\mathcal{X}}\int_{\mathcal{X}}k(x,y)\mathbb{P}(\mathrm{d}x)\mathbb{P}(\mathrm{d}y) - \frac{2}{m}\int_{\mathcal{X}}\sum_{i=1}^{m}k(x,y_i)\mathbb{P}(\mathrm{d}x) + \frac{1}{m(m-1)}\sum_{i\neq i'}k(y_i,y_{i'}).$$

Note that if $\{y_j\}_{j=1}^{m} \overset{\mathrm{IID}}{\sim} \mathbb{Q}$ then $\mathbb{E}[\mathrm{MMD}_U^2(\mathbb{P}||\mathbb{Q}^m)] = \mathrm{MMD}^2(\mathbb{P}||\mathbb{Q})$.

Lemma 2. *Suppose that k is bounded, then for any two $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$ and empirical distribution $\mathbb{Q}^m = \frac{1}{m}\sum_{i=1}^{m}\delta_{y_j}$ in $\mathcal{P}_k(\mathcal{X})$ made of independently and identically distributed realisations of $\mathbb{Q}$, we have: $\left|MMD_U^2(\mathbb{P}||\mathbb{Q}^m) - MMD^2(\mathbb{P}||\mathbb{Q}^m)\right| \leq 2m^{-1}\sup_{x\in\mathcal{X}}k(x,x)$ and:*

$$MMD^2(\mathbb{P}||\mathbb{Q}) = \mathbb{E}[MMD^2(\mathbb{P}||\mathbb{Q}^m)] + m^{-1}\left(\int_{\mathcal{X}}\int_{\mathcal{X}}k(x,y)\mathbb{Q}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y) - \int_{\mathcal{X}}k(x,x)\mathbb{Q}(\mathrm{d}x)\right).$$

Similarly, when computing the MMD squared between $\mathbb{Q}^m$ and $\mathbb{P}^n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \in \mathcal{P}_k(\mathcal{X})$ (made out of IID realisations from $\mathbb{P}$) $\left|MMD_{U,U}^2(\mathbb{P}^n||\mathbb{Q}^m) - MMD^2(\mathbb{P}^n||\mathbb{Q}^m)\right| \leq 2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)$, and similarly:

$$MMD^2(\mathbb{P}||\mathbb{Q}) = \mathbb{E}[MMD^2(\mathbb{P}^n||\mathbb{Q}^m)] + \left(m^{-1}+n^{-1}\right)\left(\int_{\mathcal{X}}\int_{\mathcal{X}}k(x,y)\mathbb{Q}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y) - \int_{\mathcal{X}}k(x,x)\mathbb{Q}(\mathrm{d}x)\right).$$

Proof. We see that

$$\mathrm{MMD}_U^2(\mathbb{P}||\mathbb{Q}^m) - \mathrm{MMD}^2(\mathbb{P}||\mathbb{Q}^m)$$

$$= (m(m-1))^{-1}\sum_{i\neq j}k(y_i,y_j) - m^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(y_i,y_j)$$

$$= (m(m-1))^{-1}(1-(m(m-1))m^{-2})\sum_{i\neq j}k(y_i,y_j) - m^{-2}\sum_{i=1}^{m}k(y_i,y_i)$$

$$= m^{-1}\left((m(m-1))^{-1}\sum_{i\neq j}k(y_i,y_j) - m^{-1}\sum_{i=1}^{m}k(y_i,y_i)\right).$$

Since the kernel is bounded, it follows that $\left|\mathrm{MMD}_U^2(\mathbb{P}||\mathbb{Q}^m) - \mathrm{MMD}^2(\mathbb{P}||\mathbb{Q}^m)\right| \leq 2m^{-1}\sup_{x\in\mathcal{X}}k(x,x)$ as required. The second statement follows in a similar fashion and from the fact that MMD_U^2 is an unbiased estimator of MMD^2. Similarly for the discrepancy $\mathrm{MMD}_{U,U}^2$:

$$\mathrm{MMD}_{U,U}^2(\mathbb{P}^n||\mathbb{Q}^m) - \mathrm{MMD}^2(\mathbb{P}^n||\mathbb{Q}^m) = m^{-1}(m(m-1))^{-1}\sum_{i\neq j}k(y_i,y_j) - m^{-1}\sum_{i=1}^{m}k(y_i,y_i))$$

$$+ n^{-1}((n(n-1))^{-1}\sum_{i\neq j}k(x_i,x_j) - n^{-1}\sum_{i=1}^{n}k(x_i,x_i)),$$

so $\left|\mathrm{MMD}_{U,U}^2(\mathbb{P}^n||\mathbb{Q}^m) - \mathrm{MMD}^2(\mathbb{P}^n||\mathbb{Q}^m)\right| \leq 2(m^{-1}+n^{-1})\sup_{x\in\mathcal{X}} k(x,x)$ and the final equation holds similarly. $\qquad\square$

We now establish conditions under which a minimiser of the empirical loss always exists.

Lemma 3. *Suppose that the kernel k is continuous and bounded and that the map $\theta \to G_\theta(u)$ continuous for almost every $u \in \mathcal{U}$ and $\theta \in \Theta$. Then given $n, m \in \mathbb{N}$ the following statements hold.*

1. *Let $\epsilon^* = \inf_{\theta\in\Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}^m)$. Then if for some $\epsilon = \epsilon(m,\omega) > 0$ the set*

$$\{\theta \in \Theta : MMD(\mathbb{P}_\theta||\mathbb{Q}^m) \leq \epsilon^* + \epsilon\} \subset \Theta,$$

   is bounded then $\arg\inf_{\theta\in\Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}^m) \neq \emptyset$.

2. *Let $\epsilon^* = \inf_{\theta\in\Theta} MMD(\mathbb{P}_\theta^n||\mathbb{Q}^m)$, if for some $\epsilon = \epsilon(n,m,\omega) > 0$ the set*

$$\{\theta \in \Theta : MMD(\mathbb{P}_\theta^n||\mathbb{Q}^m) \leq \epsilon^* + \epsilon\} \subset \Theta,$$

   is bounded then $\arg\inf_{\theta\in\Theta} MMD(\mathbb{P}_\theta^n||\mathbb{Q}^m) \neq \emptyset$.

Proof. The continuity assumption on G_θ implies that the map $\theta \to \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m)$ is continuous from Θ to $[0,\infty)$. By definition of the infimum, it follows that $\{\theta \in \Theta : \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m) \leq \epsilon^* + \epsilon\} \neq \emptyset$. Moreover, by continuity of the map, the set is closed and bounded in Θ and thus compact in Θ. The map $\theta \to \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m)$ therefore will attain its minimum within the set, and so the first statement follows. The result for the second estimator follows in an analogous fashion. $\qquad\square$

We now provide the key concentration inequality.

Proof of Lemma 1. Let $\mathcal{F}_k = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$. By definition, we have $\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n) = \sup_{f\in\mathcal{F}_k} |\int_\mathcal{X} f(x)\mathbb{P}(\mathrm{d}x) - \frac{1}{n}\sum_{i=1}^n f(x_i)|$. Define $h(x_1,\ldots,x_n) = \sup_{f\in\mathcal{F}_k} |\frac{1}{n}\sum_{i=1}^n (f(x_i) - \int_\mathcal{X} f(x)\mathbb{P}(\mathrm{d}x))|$. By definition, for all $\{x_i\}_{i=1}^n, x_i' \in \mathcal{X}$,

$$|h(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - h(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|$$
$$\leq 2n^{-1}\sup_{x\in\mathcal{X}} k(x,x)^{1/2}.|h(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - h(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|$$
$$\leq 2n^{-1}\sup_{x\in\mathcal{X}} k(x,x)^{1/2}.$$

By McDiarmid's inequality [McDiarmid, 1989] we have that for any $\varepsilon > 0$: $Pr(\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n) - \mathbb{E}[\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n)] \geq \varepsilon) \leq \exp(-2\varepsilon^2/4n^{-1}\sup_{x\in\mathcal{X}} k(x,x))$. Setting the RHS to be δ, it follows that with probability greater than $1 - \delta$,

$$\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n) - \mathbb{E}[\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n)] < \sqrt{2n^{-1}\sup_{x\in\mathcal{X}} k(x,x)\log(1/\delta)}.$$

From Jensen's inequality and Lemma 2, we obtain that

$$\mathbb{E}[\mathrm{MMD}(\mathbb{P}||\mathbb{P}^n)] \leq \mathbb{E}[\mathrm{MMD}^2(\mathbb{P}||\mathbb{P}^n)]^{1/2} \leq \sqrt{2n^{-1}}\sup_{x\in\mathcal{X}} k(x,x)^{1/2},$$

so that the advertised result holds. $\qquad\square$

We now prove Theorem 1:

Proof. From Lemma 2 and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain $\forall \mathbb{P} \in \mathcal{P}_k(\mathcal{X})$, $|\mathrm{MMD}_U(\mathbb{P}||\mathbb{Q}^m) - \mathrm{MMD}(\mathbb{P}||\mathbb{Q}^m)| \leq \sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)}$. In particular, since $\theta \to \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m)$ is bounded from below, using the above inequality and the definition of $\hat{\theta}_m$, we obtain that:

$$
\begin{aligned}
\mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}^m\right) &\leq \mathrm{MMD}_U\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}^m\right) + \sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)} \\
&= \inf_{\theta \in \Theta} \mathrm{MMD}_U(\mathbb{P}_\theta||\mathbb{Q}^m) + \sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)} \\
&\leq \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m) + 2\sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)}.
\end{aligned}
$$

We can then write:

$$
\begin{aligned}
&\mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \\
&\leq \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) - \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}^m\right) + \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}^m\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \\
&\leq \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) - \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}^m\right) + \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m) \\
&\quad - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) + 2\sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)}.
\end{aligned}
$$

Since the θ-indexed family $\mathrm{MMD}(\mathbb{P}_\theta||\cdot)$ is uniformly bounded (since k is bounded), and using that for bounded functions $f, g : \mathbb{R} \to \mathbb{R}$, $|\inf_\theta f(\theta) - \inf_\theta g(\theta)| \leq \sup_\theta |f-g|$ and the reverse triangle inequality, we further obtain that

$$
\begin{aligned}
&\mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \\
&\leq 2\sup_{\theta \in \Theta} |\mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) - \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}^m)| + 2\sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)} \\
&\leq 2\sup_{\theta \in \Theta} \mathrm{MMD}(\mathbb{Q}||\mathbb{Q}^m) + 2\sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)}.
\end{aligned}
$$

Applying Lemma 1, with probability $1 - \delta$,

$$
\mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \leq 2\sqrt{2m^{-1} \sup_{x \in \mathcal{X}} k(x,x)}(2 + \sqrt{\log(1/\delta)}),
$$

as required. For the second generalisation bound, note that

$$
\begin{aligned}
&\mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \\
&\leq \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) - \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}^n||\mathbb{Q}\right) + \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}^n||\mathbb{Q}\right) - \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}^n||\mathbb{Q}^m\right) \\
&\quad + \mathrm{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}^n||\mathbb{Q}^m\right) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta^n||\mathbb{Q}) + \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta^n||\mathbb{Q}) - \inf_{\theta \in \Theta} \mathrm{MMD}(\mathbb{P}_\theta||\mathbb{Q}).
\end{aligned}
$$

38

We can bound the individual terms on the RHS as follows via the triangle inequality,

$$\text{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{Q})$$

$$\leq \text{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{P}^n_{\hat{\theta}_{n,m}}\right) + \text{MMD}(\mathbb{Q}^m||\mathbb{Q}) + \text{MMD}(\mathbb{P}^n_{\hat{\theta}_{n,m}}||\mathbb{Q}^m)$$

$$- \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}) + \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{Q}).$$

Similarly as above,

$$\text{MMD}\left(\mathbb{P}^n_{\hat{\theta}_{n,m}}||\mathbb{Q}^m\right) \leq \text{MMD}_{U,U}\left(\mathbb{P}^n_{\hat{\theta}_{n,m}}||\mathbb{Q}^m\right) + \sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}$$

$$= \inf_{\theta\in\Theta}\text{MMD}_{U,U}(\mathbb{P}^n_\theta||\mathbb{Q}^m) + \sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}$$

$$\leq \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}^m) + 2\sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}.$$

Similarly we obtain that

$$\text{MMD}\left(\mathbb{P}^n_{\hat{\theta}_{n,m}}||\mathbb{Q}^m\right) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q})$$

$$\leq \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}^m) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}) + 2\sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}$$

$$\leq \sup_{\theta\in\Theta}|\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}^m) - \text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q})| + 2\sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}$$

$$\leq \text{MMD}(\mathbb{Q}||\mathbb{Q}^m) + 2\sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)},$$

and $\inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{Q}) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{Q}) \leq \sup_{\theta\in\Theta}\text{MMD}(\mathbb{P}^n_\theta||\mathbb{P}_\theta)$. Combining these inequalities we obtain,

$$\text{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{Q})$$

$$\leq 2\sup_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{P}^n_\theta) + 2\text{MMD}(\mathbb{Q}^m||\mathbb{Q}) + 2\sqrt{2\left(m^{-1}+n^{-1}\right)\sup_{x\in\mathcal{X}}k(x,x)}.$$

Applying Lemma 1 with probability $1-2\delta$,

$$\text{MMD}\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) - \inf_{\theta\in\Theta}\text{MMD}(\mathbb{P}_\theta||\mathbb{Q})$$

$$\leq 2(\sqrt{2n^{-1}}+\sqrt{2m^{-1}})\sqrt{\sup_{x\in\mathcal{X}}k(x,x)}(1+\sqrt{\log(1/\delta)}) + 2\sqrt{2(m^{-1}+n^{-1})\sup_{x\in\mathcal{X}}k(x,x)}$$

$$\leq 2(\sqrt{2n^{-1}}+\sqrt{2m^{-1}})\sqrt{\sup_{x\in\mathcal{X}}k(x,x)}(2+\sqrt{\log(1/\delta)}).$$

$\square$

B.2 Proof of Proposition 1

Proof. Given $m \in \mathbb{N}$ define the event

$$A_m = \left\{ \left| \text{MMD} \left(\mathbb{P}_{\hat{\theta}_m} || \mathbb{Q} \right) - \inf_{\theta' \in \Theta} \text{MMD}(\mathbb{P}_{\theta'} || \mathbb{Q}) \right| > 2\sqrt{\frac{2}{m} \sup_x k(x,x)} (2 + \sqrt{2 \log m}) \right\}.$$

From Theorem 1 (where we have set $\delta = 1/m^2$), $\mathbb{Q}(A_m) \leq \frac{1}{m^2}$, and so $\sum_m \mathbb{Q}(A_m) < \infty$. The Borel Cantelli lemma implies that $\mathbb{Q}$-almost surely, there exists $M \in \mathbb{N}$ such that for all $m \geq M$,

$$\text{MMD} \left(\mathbb{P}_{\hat{\theta}_m} || \mathbb{Q} \right) - \inf_{\theta' \in \Theta} \text{MMD}(\mathbb{P}_{\theta'} || \mathbb{Q}) \leq 2\sqrt{\frac{2}{m} \sup_x k(x,x)} (2 + \sqrt{2 \log m}).$$

Since the right hand side converges to zero, it follows that $\text{MMD}(\mathbb{P}_{\hat{\theta}_m} || \mathbb{Q}) \to \inf_{\theta' \in \Theta} \text{MMD}(\mathbb{P}_{\theta'} || \mathbb{Q}) = \text{MMD}(\mathbb{P}_{\theta^*} || \mathbb{Q})$, $\mathbb{Q}$-almost surely. By assumption (ii), the set $\{\hat{\theta}_m\}_{m \in \mathbb{N}}$ is bounded almost surely and thus possesses at least one limit point in Θ. Moreover each subsequence $(\hat{\theta}_{n_k})_{k \in \mathbb{N}}$ satisfies $\text{MMD}(\mathbb{P}_{\hat{\theta}_{m_k}} || \mathbb{Q}) \to \text{MMD}(\mathbb{P}_{\theta^*} || \mathbb{Q})$, so that any limit point must equal θ^*, thus establishing almost sure convergence. The consistency for the estimator $\hat{\theta}_{m,n}$ follows in an analogous manner. $\square$

B.3 Proof of Theorem 2

Proof. We shall prove the result only for the estimator $\hat{\theta}_{n,m}$ since the proof of the central limit theorem for $\hat{\theta}_m$ follows in an entirely analogous way. Recall that

$$\text{MMD}^2_{U,U}(\mathbb{P}^n_\theta || \mathbb{Q}^m) = \frac{1}{n(n-1)} \sum_{i \neq j} k(G_\theta(u_i), G_\theta(u_j)) +$$

$$- \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} k(G_\theta(u_i), y_j) + \frac{1}{m(m-1)} \sum_{i \neq j} k(y_i, y_j)$$

For $n, m \in \mathbb{N}$ define $F_{n,m}(\theta) = F_{n,m}(\theta, \omega)$ by $F_{n,m}(\theta) = \text{MMD}^2_{U,U}(\mathbb{P}^n_\theta || \mathbb{Q}^m)$. By definition $\hat{\theta}_{n,m}$ is a local minimum for $F_{n,m}$, so the first order optimality condition implies that $\nabla_\theta F_{n,m}(\hat{\theta}_{n,m}) = 0$. Since Θ is open, by applying the mean value theorem to $\nabla_\theta F_{n,m}$ we obtain $0 = \nabla_\theta F_{n,m}(\theta^*) + \nabla_\theta \nabla_\theta F_{n,m}(\tilde{\theta})(\hat{\theta}_{n,m} - \theta^*)$, where $\tilde{\theta}$ lies on the line between $\hat{\theta}_{n,m}$ and θ^*. Let $\{u_i\}_{i=1}^{n}$ be independently and identically distributed realisations from $\mathbb{U}$. Since $\mathbb{Q} = G^{\#}_{\theta^*} \mathbb{U}$, there exist $\{\tilde{u}_1, \ldots, \tilde{u}_m\}$ which are $\mathbb{U}$ distributed and independent from $\{u_i\}$ such that

$$\nabla_\theta F_{n,m}(\theta^*) = \frac{2}{n(n-1)} \sum_{i \neq j} \nabla_1 k(G_{\theta^*}(u_i), G_{\theta^*}(u_j)) \nabla_\theta G_{\theta^*}(u_i)$$

$$- \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \nabla_1 k(G_{\theta^*}(u_i), G_{\theta^*}(\tilde{u}_j)) \nabla_\theta G_{\theta^*}(u_i).$$

Note that $\mathbb{E}[\nabla_\theta F_{n,m}(\theta^*)] = 0$. We wish to characterise the fluctuations of $\nabla_\theta F_{n,m}(\theta^*)$ as $n, m \to \infty$. Define the U-statistic $U_1 = (n(n-1))^{-1} \sum_{i \neq j} h(u_i, u_j)$, where

$$h(u,v) = \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) + \nabla_1 k(G_{\theta^*}(v), G_{\theta^*}(u)) \nabla_\theta G_{\theta^*}(v),$$

and the U-statistic $U_2 = (nm)^{-1} \sum_{i,j=1}^{n,m} g(u_i, \tilde{u}_j)$, where

$$g(u, v) = 2\nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v))\nabla_\theta G_{\theta^*}(u).$$

From the calculations above we have $\nabla_\theta F_{n,m}(\theta^*) = U_1 - U_2$. Following van der Vaart [1998] we make use of the Hajek projection principle to identify $U_1 - U_2$ as small perturbation of a sum of independently and identically distributed random variables, from which a central limit theorem can be obtained, see also Hoeffding [1948], Lehmann [1951]. To this end, we look for a projection onto the set of all random variables of the form $\sum_{i=1}^n \hat{h}_i(u_i) - \sum_{j=1}^m \hat{g}_i(\tilde{u}_j)$, where $\hat{h}_i$ and $\hat{g}_i$ are square-integrable measurable functions. Let $M = \mathbb{E}[U_1] = \mathbb{E}[U_2]$, the Hajek projection principle [van der Vaart, 1998, Chap. 11 & 12] implies that $U_1 - M$ has projection $\hat{U}_1 = \frac{2}{n} \sum_{i=1}^n h_1(u_i)$, where $h_1(u) = \mathbb{E}_X h(u, X) - M$. Similarly, $U_2 - M$ has projection $\hat{U}_2 = \frac{1}{n} \sum_{i=1}^n g_1(u_i) + \frac{1}{m} \sum_{i=1}^m g_2(\tilde{u}_i)$, where $g_1(u) = \mathbb{E}g(u, Y) - M$ and $g_2(y) = \mathbb{E}g(X, y) - M$. By the central limit theorem for identically and independently distributed random variables, $\sqrt{n+m}(\hat{U}_1 - \hat{U}_2) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma = A + B - 2C$ and

$$A = \lim_{k\to\infty} 4(n_k + m_k)n_k^{-2} \sum_{i=1}^{n_k} \mathrm{Cov}[h_1(u_i)]$$

$$= 4\lambda^{-1} \int_{\mathcal{U}} \left(\int_{\mathcal{U}} (h(u, v) - M)\,\mathbb{U}(dv) \otimes \int_{\mathcal{U}} (h(u, w) - M)\,\mathbb{U}(dw) \right) \mathbb{U}(du),$$

where λ is defined in Assumption 4. Similarly,

$$B = \lim_{k\to\infty} (n_k + m_k/n_k^2) \sum_{i=1}^{n_k} \mathrm{Cov}[g_1(u_i)] + (m_k + n_k/m_k^2) \sum_{i=1}^{m_k} \mathrm{Cov}[g_2(\tilde{u}_i)]$$

$$= \lambda^{-1} \int_{\mathcal{U}} \left(\int_{\mathcal{U}} (g(u, v) - M)\mathbb{U}(dv) \otimes \int_{\mathcal{U}} (g(u, w) - M)\mathbb{U}(dw) \right) \mathbb{U}(du)$$

$$+ (1 - \lambda)^{-1} \int_{\mathcal{U}} \left(\int_{\mathcal{U}} (g(u, v) - M)\mathbb{U}(du) \otimes \int_{\mathcal{U}} (g(w, v) - M)\mathbb{U}(dw) \right) \mathbb{U}(dv),$$

$$C = 2 \lim_{k\to\infty} (n_k + m_k)n_k^{-2}\mathrm{Cov}\left[\sum_{i=1}^{n_k} h_1(u_i), \sum_{i=1}^{n_k} g_1(u_i) \right]$$

$$= 2\lambda^{-1} \int_{\mathcal{U}} \int_{\mathcal{U}} (h(u, v) - M)\,\mathbb{U}(dv) \otimes \int_{\mathcal{U}} (g(u, w) - M)\,\mathbb{U}(dw)\mathbb{U}(du),$$

Substituting the values of g and h we arrive at Σ. We will show that the remainder term $R_k = \sqrt{n_k + m_k}((U_1 - \hat{U}_1) + (U_2 - \hat{U}_2))$ converges to 0 in probability, as $k \to \infty$, which will imply the desired result, by Slutsky's theorem. This term has expectation zero for all $k \in \mathbb{N}$. Moreover

$$\mathbb{E}[\|R_k\|]^2 \le 2(n_k + m_k)n_k^{-1}n_k\mathrm{Tr}(\mathrm{Cov}[U_1 - \hat{U}_1]) + 2(n_k + m_k)\mathrm{Tr}(\mathrm{Cov}[U_2 - \hat{U}_2]).$$

Using the fact that $n_k(n_k + m_k)^{-1} \to \lambda$ as $k \to \infty$, and by [van der Vaart, 1998, Theorem 12.3], the first term converges on the right hand side converges to 0. For the second term, from [van der

Vaart, 1998, Theorem 12.6] both $(n_k + m_k)\text{Tr}\,(\text{Cov}[U_2])$ and $(n_k + m_k)\text{Tr}(\text{Cov}[\hat{U}_2])$ converge to

$$\lambda^{-1}\text{Tr}\left(\int_{\mathcal{U}}\left(\int_{\mathcal{U}}(g(u,v) - M)\,\mathbb{U}(\mathrm{d}v)\right)^{\otimes 2}\mathbb{U}(\mathrm{d}u)\right) \tag{13}$$

$$+ (1-\lambda)^{-1}\text{Tr}\left(\int_{\mathcal{U}}\left(\int_{\mathcal{U}}(g(u,v) - M)\,\mathbb{U}(\mathrm{d}u)\right)^{\otimes 2}\mathbb{U}(\mathrm{d}v)\right). \tag{14}$$

It remains to consider $\text{Cov}[U_2, \hat{U}_2]$ which is given by

$$(n_k + m_k)\mathbb{E}\left[\left(n_k^{-1}\sum_{i=1}^{n_k}g_1(u_i) + m_k^{-1}\sum_{i=1}^{m}g_2(\tilde{u}_i) - 2M\right)\otimes\left((n_k m_k)^{-1}\sum_{i,j=1}^{n_k,m_k}g(u_i, \tilde{u}_j) - M\right)\right]$$

$$= ((n_k + m_k)n_k^{-1})(n_k + m_k)^{-1}\sum_{i=1}^{n_k}\mathbb{E}[g_1(u_i)^{\otimes 2}]$$

$$+ \left(\frac{n_k + m_k}{m_k}\right)\frac{1}{n_k + m_k}\sum_{i=1}^{m_k}\mathbb{E}[g_2(\tilde{u}_i)^{\otimes 2}] - 2(n_k + m_k)M \otimes M,$$

so that $\text{Tr}(\text{Cov}[U_2, \hat{U}_2])$ converges to (13) as $k \to \infty$, and so $\text{Cov}[U_2 - \hat{U}_2] \to 0$ as required. Now consider the term $H_{m,n} = \nabla_\theta\nabla_\theta F_{n,m}(\tilde{\theta})$ in the first order Taylor expansion, where $\tilde{\theta}$ lies along the line between θ^* and $\hat{\theta}_{n,m}$. We show that $\nabla_\theta\nabla_\theta F_{n,m}(\tilde{\theta})$ converges to $g(\theta^*)$ as $n, m \to \infty$. To this end, consider $H_{m,n}^{a,b}(\tilde{\theta}) - g_{ab}(\theta^*)$, where

$$H_{m,n}^{a,b}(\theta) = (n(n-1))^{-1}\partial_{\theta_a}\partial_{\theta_b}\sum_{i\neq j}k(G_\theta(u_i), G_\theta(u_j)) - 2(nm)^{-1}\partial_{\theta_a}\partial_{\theta_b}\sum_{i,j=1}^{n,m}k(G_\theta(u_i), y_j).$$

Then we have that $|H_{m,n}^{a,b}(\tilde{\theta}) - g_{ab}(\theta^*)| \leq |H_{m,n}^{a,b}(\tilde{\theta}) - g_{ab}(\tilde{\theta})| + |g_{ab}(\tilde{\theta}) - g_{ab}(\theta^*)|$. Since $\hat{\theta}_{n,m} \to \theta^*$ almost surely, it follows that $\tilde{\theta} \to \theta^*$, and so for n, m sufficiently large, $\tilde{\theta}$ almost surely lies in the compact set K. Thus $|H_{m,n}^{a,b}(\tilde{\theta}) - g_{ab}(\theta^*)| \leq \sup_{\theta \in K}|H_{m,n}^{a,b}(\theta) - g_{ab}(\theta)| + |g_{ab}(\tilde{\theta}) - g_{ab}(\theta^*)|$.

It follows from the assumptions and the dominated convergence theorem that $\theta \to g_{ab}(\theta)$ is continuous on K. By Assumption 3, the first three θ-derivatives of G_θ are bounded in K and so the conditions of Lemma 4 hold, so that the first term goes to zero in probability. The second term converges to zero by continuity on K. Since g is assumed to be invertible, there exists $m = m(\omega), n = n(\omega)$ after which $H_{m,n}(\tilde{\theta})$ is also invertible, so that by Slutsky's theorem

$$\sqrt{n_k + m_k}(\hat{\theta}_{n,m} - \theta^*) = -(H_{m,n})^{-1}\sqrt{n_k + m_k}\nabla_\theta F_{n,m}(\theta^*) \xrightarrow{d} \mathcal{N}(0, g(\theta^*)^{-1}\Sigma g(\theta^*)^{-1}).$$

$\square$

Lemma 4. *Let K be a compact set and $q_1(x, y, \theta) = \partial_{\theta_a}\partial_{\theta_b}k(G_\theta(x), G_\theta(y))$ and $q_2(x, y, \theta) = 2\partial_{\theta_a}\partial_{\theta_b}k(G_\theta(x), y)$. Suppose that for $\theta_1, \theta_2 \in K$ we have, $|q_1(x, y, \theta_1) - q_1(x, y, \theta_2)| \leq (\theta_1 - \theta_2)Q_1(x, y)$ and $|q_1(x, y, \theta_1) - q_1(x, y, \theta_2)| \leq (\theta_1 - \theta_2)Q_2(x, y)$, where $\int_{\mathcal{X}}\int_{\mathcal{X}}Q_1(x, y)\mathbb{U}(\mathrm{d}x)\mathbb{U}(\mathrm{d}y) < \infty$ and $\int_{\mathcal{X}}\int_{\mathcal{X}}Q_2(x, y)\mathbb{U}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y) < \infty$. Then $\sup_{\theta \in K}|H_{m,n}^{a,b}(\theta) - g_{ab}(\theta)| \xrightarrow{p} 0$ as $m \wedge n \to \infty$.*

Proof. We show that the spaces of functions $\mathcal{Q}_1 = \{q_1(\cdot, \cdot, \theta) : \theta \in K\}$ and $\mathcal{Q}_2 = \{q_2(\cdot, \cdot, \theta) : \theta \in K\}$ are Euclidean in the sense of Nolan and Pollard [1987]. Let $\epsilon > 0$ and let $\theta_1, \ldots, \theta_M \in K$ be centers of an ϵ–cover of K, where $M = \mathrm{diam}(K)/\epsilon$. Given $q_i \in \mathcal{Q}_i$, $i = 1, 2$, there exists θ_k such that $|q_i(\cdot, \cdot, \theta_k) - q_i(\cdot, \cdot, \theta)| \leq \epsilon Q_i(\cdot, \cdot)$, and so, given a measure μ on $(\Theta, \mathcal{B}(\Theta))$ such that $\mu(Q_i) < \infty$ we have $\mu|q_i(\cdot, \cdot, \theta_k) - q_i(\cdot, \cdot, \theta)| \leq \epsilon \mu(Q_i)$, therefore $N_1(\epsilon, \mu, \mathcal{Q}_i) \leq \mathrm{diam}(K)\epsilon^{-1}$. Invoking [Nolan and Pollard, 1987, Theorem 7] for q_1 and [Neumeyer, 2004, Theorem 2.9] for q_2, we obtain the required result. $\qquad\square$

B.4 Proof of Theorem 3

Proof. Define the function

$$
h(x, \theta) = 2 \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_\theta(u), G_\theta(v)) \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v)
$$
$$
- 2 \int_{\mathcal{U}} \nabla_1 k(G_\theta(u), x) \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}u),
$$

which satisfies $\int_{\mathcal{X}} h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) = 0$ for all $\theta \in \Theta$. Differentiating this integral with respect to θ yields $\int_{\mathcal{X}} \nabla_\theta h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) = - \int_{\mathcal{X}} h(x, \theta) \otimes \nabla_\theta p(x \mid \theta) \mathrm{d}x$, where $p(x \mid \theta)$ is the density of $\mathbb{P}_\theta$ with respect to the Lebesgue measure on $\mathcal{X}$.

Let $X \sim \mathbb{P}_\theta$. Consider the covariance of $(h(X, \theta), \nabla \log p(X|\theta))^\top$, then

$$
\mathrm{Cov}(h(X, \theta), \nabla \log p_\theta(X))^\top
$$
$$
= \begin{pmatrix} \int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) p(x|\theta)\mathrm{d}x & \int_{\mathcal{X}} h(x, \theta) \otimes \nabla \log p(x|\theta) p(x|\theta)\mathrm{d}x \\ \int_{\mathcal{X}} h(x, \theta) \otimes \nabla \log p(x|\theta) p(x|\theta)\mathrm{d}x & \int_{\mathcal{X}} \nabla \log p(x|\theta) \otimes \nabla \log p(x|\theta) p(x|\theta)\mathrm{d}x \end{pmatrix}
$$
$$
= \begin{pmatrix} \int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) p(x|\theta)\mathrm{d}x & - \int_{\mathcal{X}} \nabla_\theta h(x, \theta) p(x|\theta)\mathrm{d}x \\ - \int_{\mathcal{X}} \nabla_\theta h(x, \theta) p(x|\theta)\mathrm{d}x & F(\theta) \end{pmatrix},
$$

where $F(\theta)$ is the Fisher information matrix. Since this is a covariance matrix, the determinant is non-negative, and so

$$
\det(F(\theta)) \det \left(\int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) - \left(\int_{\mathcal{X}} \nabla_\theta h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) \right) F^{-1}(\theta) \left(\int_{\mathcal{X}} \nabla_\theta h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) \right) \right) \geq 0.
$$

Since the Fisher information is positive at $\theta = \theta^*$ this implies that $\det F(\theta) > 0$ and so

$$
\int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) - \left(\int_{\mathcal{X}} \nabla_\theta h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) \right) F^{-1}(\theta) \left(\int_{\mathcal{X}} \nabla_\theta h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) \right)
$$

is non-negative definite. We note that $\int_{\mathcal{X}} \nabla_\theta h(x, \theta) p(x)\mathrm{d}x = g(\theta)$ is the information metric associateed with the MMD induced distance and is positive definite at $\theta = \theta^*$. It follows that $(1/4)g^{-1}(\theta)(\int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x))g^{-1}(\theta) - F^{-1}(\theta)$ is non-negative definite at $\theta = \theta^*$. Since

$$
\int_{\mathcal{X}} h(x, \theta) \otimes h(x, \theta) \mathbb{P}_\theta(\mathrm{d}x) = 4 \int_{\mathcal{U}} \left(\int_{\mathcal{U}} \nabla_1 k(G_\theta(u), G_\theta(v))^\top \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}u) - \mathcal{M} \right)^{\otimes 2} \mathbb{U}(\mathrm{d}v),
$$

where $\mathcal{M} = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_\theta(u), G_\theta(v)) \mathbb{U}(\mathrm{d}u)^\top \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}v)$ we see that $\frac{1}{4} g^{-1}(\theta)(\int_{\mathcal{X}} h(x,\theta) \otimes$
$h(x,\theta)\mathbb{P}_\theta(\mathrm{d}x)) g^{-1}(\theta)$ equals the asymptotic variance C for the estimator $\hat{\theta}_m$ and so $C - F^{-1}(\theta)$
is positive definite when $\theta = \theta^*$ giving the advertised inequality.

Now since $C_\lambda = (1/(1-\lambda)\lambda)C \succeq C$, it follows that $C_\lambda - F^{-1}(\theta)$ is also positive definite
when $\theta = \theta^*$ and the Cramer-Rao bound also holds for the estimator $\hat{\theta}_{n,m}$. $\qquad\square$

B.5   Proof of Proposition 2

Proof. We have that $\nabla_1 k(x,y) = ((x-y)/l^2) r'(|x-y|^2/2l^2)$, and $\nabla_1 \nabla_2 k(x,y) = -l^{-2} r'(|x-y|^2/2l^2) - l^{-4}(x-y)^2 r''(|x-y|^2/2l^2))$. We first note that the metric tensor g satisfies

$$l^2 g(\theta) \xrightarrow{l^2 \to \infty} R \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla G_\theta(u) \nabla G_\theta(v)^\top \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v) = \nabla_\theta M(\theta) \nabla_\theta M(\theta)^\top,$$

where $M(\theta) = \int_{\mathcal{X}} x p(x|\theta)\, \mathrm{d}x$ and $R = \lim_{s \to \infty} r'(s)$. Defining $S(\theta) = \int_{\mathcal{U}} |G_\theta(u)|^2 \mathbb{U}(du)$ we
obtain:

$$l^4 \Sigma \xrightarrow{l^2 \to \infty} R^2 \int_{\mathcal{U}} \left[\left(\int_{\mathcal{U}} \nabla_\theta G(u) \cdot (G_\theta(u) - G_\theta(v)) \mathbb{U}(\mathrm{d}u) \right) \right.$$
$$\left. \otimes \left(\int_{\mathcal{U}} \nabla_\theta G(w) \cdot (G_\theta(w) - G_\theta(v)) \mathbb{U}(\mathrm{d}w) \right) \right] \mathbb{U}(\mathrm{d}v)$$
$$- R^2 \left(\int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G(u)(G_\theta(u) - G_\theta(v)) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v) \right)^{\otimes 2}$$
$$= R^2 \nabla_\theta M(\theta) \cdot (V(\theta) + M(\theta)\, M(\theta)) \nabla_\theta M(\theta)^\top - \frac{R^2}{4} \left(\nabla_\theta |M(\theta)|^2 \right) \left(\nabla_\theta |M(\theta)|^2 \right)^\top$$
$$= R^2 \nabla_\theta M(\theta) \cdot V(\theta) \nabla_\theta M(\theta) + \frac{R^2}{4} \left(\nabla_\theta |M(\theta)|^2 \right) \left(\nabla_\theta |M(\theta)|^2 \right)^\top - \frac{R^2}{4} \left(\nabla_\theta |M(\theta)|^2 \right) \left(\nabla_\theta |M(\theta)|^2 \right)^\top$$
$$= R^2 \nabla_\theta M(\theta) \cdot V(\theta) \nabla_\theta M(\theta).$$

Combining we obtain

$$\lim_{l \to \infty} C^l = \left(\nabla_\theta M(\theta)\, \nabla_\theta M(\theta)^\top \right)^{-1} \nabla_\theta M(\theta) \cdot V \nabla_\theta M(\theta) \left(\nabla_\theta M(\theta)\, \nabla_\theta M(\theta)^\top \right)^{-1}$$
$$= (\nabla_\theta M(\theta))^\dagger V(\theta) (\nabla_\theta M(\theta))^{\dagger\top}.$$

$\qquad\square$

B.6   Proof of Theorem 4

Proof. Let $\epsilon > 0$, and let δ be as in assumption (ii). Suppose that $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}_k$ satisfy $d_{BL}(\mathbb{Q}_1, \mathbb{Q}_2) < (1 + \sqrt{k(0,0)})\delta/2$, where d_{BL} denotes the Bounded Lipschitz or Dudley metric [Dudley, 2018].

By [Sriperumbudur et al., 2010, Theorem 21] it follows that $\text{MMD}(\mathbb{Q}_1||\mathbb{Q}_2) < \delta/2$. Let $\theta^{(1)}$ and $\theta^{(2)}$ be the minimum MMD estimators which exist by assumption (i). By the triangle inequality:

$$\text{MMD}(\mathbb{P}_{\theta^{(2)}}||\mathbb{Q}_1) \leq \text{MMD}(\mathbb{P}_{\theta^{(2)}}||\mathbb{Q}_2) + \text{MMD}(\mathbb{Q}_1||\mathbb{Q}_2) \leq \text{MMD}(\mathbb{P}_{\theta^{(2)}}||\mathbb{Q}_2) + \delta/2.$$

Suppose that $\left|\theta - \theta^{(2)}\right| > \epsilon$, then:

$$\text{MMD}(\mathbb{Q}_1||\mathbb{P}_\theta) \geq \text{MMD}(\mathbb{Q}_2||\mathbb{P}_\theta) - \text{MMD}(\mathbb{Q}_1||\mathbb{Q}_2) \geq \text{MMD}(\mathbb{Q}_2||\mathbb{P}_\theta) - \delta/2$$
$$> \text{MMD}(\mathbb{Q}_2||\mathbb{P}_{\theta^{(2)}}) + \delta/2 \geq \text{MMD}(\mathbb{Q}_1||\mathbb{P}_{\theta^{(2)}})$$

This implies that $\theta^{(1)}$ must be in the ball $\{\theta : |\theta - \theta^{(2)}| < \epsilon\}$, i.e. that $|\theta^{(1)} - \theta^{(2)}| < \epsilon$ as required. This implies that the map $T : \mathcal{P}_k \to \Theta$ defined by $T(\mathbb{Q}) = \arg\inf_{\theta \in \Theta} \text{MMD}^2(\mathbb{P}_\theta||\mathbb{Q})$ is continuous with respect to the weak topology on $\mathcal{P}_k$. In particular, for $\mathbb{Q}^m = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$, since $T(\mathbb{Q}^m) = \arg\inf_{\theta \in \Theta} \text{MMD}_U^2(\mathbb{P}_\theta||\mathbb{Q}^m)$, by [Cuevas, 1988, Theorem 2] it follows that the estimator $\hat{\theta}_m$ is qualtiatively robust. The proof of that $\hat{\theta}_{n,m}$ is eventually qualitatively robust follows similarly. $\qquad\square$

B.7 Proof of Theorem 5

Proof. Consider the influence function obtained from the kernel scoring rule: $\text{IF}_{\text{MMD}}(z, \mathbb{P}_\theta) = \left(\int_{\mathcal{X}} \nabla_\theta \nabla_\theta S_{\text{MMD}}(x, \mathbb{P}_\theta)\mathbb{P}_\theta(\mathrm{d}x)\right)^{-1} \nabla_\theta S_{\text{MMD}}(z, \mathbb{P}_\theta)$. It is straightforward to show that under assumptions (i-iii), both the first and the second term are bounded in z, which directly implies that the whole influence function is bounded and hence the estimator is bias-robust. $\qquad\square$

C   Gaussian Location and Scale Models

Throughout this section, we will repeatedly use the fact that the product of Gaussian densities can be obtained in closed form using the following expression:

$$\phi(x; m_1, \sigma_1^2)\phi(x; m_2, \sigma_2^2) = \phi(m_1; m_2, \sigma_1^2 + \sigma_2^2)\phi\left(x; \frac{m_1\sigma_2^2 + m_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$$

where we denote by $\phi(x; m, \sigma)$ the density of a d-dimensional Gaussian with mean equals to some $m \in \mathbb{R}$ times a vector of ones, and covariance σ^2 times a d-by-d identity matrix. Furthermore, we also use the following identities: $\int_{\mathcal{U}} u^\top A u \phi(u, 0, \sigma)\mathrm{d}u = \sigma\text{Tr}(A)$ and $\int_{\mathcal{U}} \|u\|_2^4 \phi(u, 0, \sigma)\mathrm{d}u = (d^2 + 2d)\sigma^2$.

C.1   Gaussian Location Model - Asymptotic Variance in high dimensions

Proof. The generator is given by $G_\theta(u) = u + \theta$ and $\mathbb{U}$ is $\mathcal{N}(0, \sigma^2 I_{d \times d})$ distributed. Assume θ^* is the truth. We wish to compute the asymptotic variance of the estimator $\hat{\theta}_m$ of θ^*. First, we observe that the mean term satisfies: $\overline{M} = 0$ since $k(u, v) = \phi(u; v, l^2)$ is symmetric with respect

to u and v. Consider the term:

$$\int_{\mathcal{U}} \nabla_\theta G_{\theta^*}(u) \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \mathbb{U}(\mathrm{d}u)$$

$$= -\int_{\mathcal{U}} (u-v)l^{-2} \exp(-(u-v)^2/2l^2)(2\pi l^2)^{-\frac{d}{2}} \exp(-u^2/2\sigma^2)(2\pi\sigma^2)^{-\frac{d}{2}} \mathrm{d}u$$

$$= -\int_{\mathcal{U}} (u-v)l^{-2}\phi(u; v, l^2)\phi(u; 0, \sigma^2) \mathrm{d}u$$

$$= -\int_{\mathcal{U}} (u-v)l^{-2}\phi(u; v\sigma^2(l^2+\sigma^2)^{-1}, l^2\sigma^2(l+\sigma^2)^{-1})\phi(v; 0, l^2+\sigma^2) \mathrm{d}u$$

$$= -l^{-2}[(v\sigma^2(l^2+\sigma^2)^{-1} - v)\phi(v; 0, l^2+\sigma^2)] = -(l^2+\sigma^2)^{-1}v \; \phi(v; 0, l^2+\sigma^2)$$

Then, we have:

$$\Sigma = \int_{\mathcal{U}} \left[\int_{\mathcal{U}} \nabla_1 k(u, v) \mathbb{U}(\mathrm{d}u) \right] \otimes \left[\int_{\mathcal{U}} \nabla_1 k(w, v) \mathbb{U}(\mathrm{d}w) \right] \mathbb{U}(\mathrm{d}v)$$

$$= (l^2+\sigma^2)^{-2} \int_{\mathcal{U}} v \otimes v \phi(v; 0, l^2+\sigma^2)\phi(v; 0, l^2+\sigma^2)\phi(v; 0, \sigma^2) \mathrm{d}v$$

$$= (l^2+\sigma^2)^{-2} \int_{\mathcal{U}} v \otimes v \phi\left(v; 0, \frac{l^2+\sigma^2}{2} \right) \phi(v; 0, 2(l^2+\sigma^2))\phi(v; 0, \sigma^2) \mathrm{d}v$$

$$= (l^2+\sigma^2)^{-2}\phi(0; 0, 2(l^2+\sigma^2))\phi\left(0; 0, \frac{3\sigma^2+l^2}{2} \right) \int_{\mathcal{U}} v \otimes v \phi\left(v; 0, \frac{\sigma^2(l^2+\sigma^2)}{(3\sigma^2+l^2)} \right) \mathrm{d}v$$

$$= (l^2+\sigma^2)^{-2}(2\pi)^{-\frac{2d}{2}}(2(l^2+\sigma^2))^{-\frac{d}{2}} \left(\frac{3\sigma^2+l^2}{2} \right)^{-\frac{d}{2}} \frac{\sigma^2(l^2+\sigma^2)}{(3\sigma^2+l^2)} I_{d \times d}$$

$$= \sigma^2(2\pi)^{-d}(l^2+\sigma^2)^{-\frac{d}{2}-1}(3\sigma^2+l^2)^{-\frac{d}{2}-1} I_{d \times d}$$

We can compute the metric tensor $g(\theta^*)$ similarly:

$$
\begin{aligned}
g(\theta^*) &= \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G_{\theta^*}(u) \nabla_1 \nabla_2 k(u,v) \nabla_\theta G_{\theta^*}(u)^\top \mathbb{U}(du) \mathbb{U}(dv) \\
&= \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 \nabla_2 k(u,v) \phi(u; 0, \sigma^2) \phi(v; 0, \sigma^2) du dv \\
&= \int_{\mathcal{U}} \int_{\mathcal{U}} k(u,v) \nabla \phi(u; 0, \sigma^2) \otimes \nabla \phi(v; 0, \sigma^2) du dv \\
&= \int_{\mathcal{U}} \int_{\mathcal{U}} k(u,v) u \otimes v \sigma^{-4} \phi(u; 0, \sigma^2) \phi(v; 0, \sigma^2) du dv \\
&= \sigma^{-4} \int_{\mathcal{U}} \int_{\mathcal{U}} \phi(v; u, l^2) u \otimes v \phi(u; 0, \sigma^2) \phi(v; 0, \sigma^2) du dv \\
&= \sigma^{-4} \int_{\mathcal{U}} \int_{\mathcal{U}} \phi\left(v; \frac{u\sigma^2}{(l^2+\sigma^2)}, \frac{l^2\sigma^2}{(l^2+\sigma^2)}\right) u \otimes v \phi(u; 0, l^2+\sigma^2) \phi(u; 0, \sigma^2) du dv \\
&= \frac{1}{(\sigma^2+l^2)\sigma^2} \int_{\mathcal{U}} u \otimes u \phi(u; 0, l^2+\sigma^2) \phi(u; 0, \sigma^2) du \\
&= \frac{1}{(\sigma^2+l^2)\sigma^2} \phi(0; 0, l^2+2\sigma^2) \int_{\mathcal{U}} u \otimes u \phi\left(u; 0, \frac{(l^2+\sigma^2)\sigma^2}{l^2+2\sigma^2}\right) du \\
&= (2\pi)^{-\frac{d}{2}} (l^2+2\sigma^2)^{-\frac{d}{2}-1} I_{d\times d}.
\end{aligned}
$$

Combining the results above we get the advertised result. $\qquad\square$

C.2   Proof of Proposition 4

Proof. The asymptotic variance satisfies

$$
C = \sigma^2((d^{2\alpha}+\sigma^2)(3\sigma^2+d^{2\alpha}))^{-\frac{d}{2}-1}(d^{2\alpha}+2\sigma^2)^{d+2} = \sigma^2\left(\frac{1+4\sigma^2 d^{-2\alpha}+3\sigma^4 d^{-4\alpha}}{1+4\sigma^2 d^{-2\alpha}+4\sigma^4 d^{-4\alpha}}\right)^{-d/2-1}.
$$

Taking logarithms, we obtain:

$$
\log C = 2\log\sigma - \left(\frac{d}{2}+1\right)\left(\log(1+4\sigma^2 d^{-2\alpha}+3d^{-4\alpha}\sigma^4) - \log(1+4\sigma^2 d^{-2\alpha}+4d^{-4\alpha}\sigma^4)\right).
$$

By l'Hopital's rule

$$
\lim_{d\to\infty} \log C = \lim_{d\to\infty} 2\left(1+\frac{d}{2}\right)^2 \frac{4\alpha\sigma^4}{d\left(6\sigma^6 d^{-2\alpha}+6\sigma^2 d^{2\alpha}+d^{4\alpha}+11\sigma^4\right)},
$$

which is converges to σ^4 if $\alpha = 1/4$, converges to 0 if $\alpha > 1/4$ and converges to infinity if $\alpha < 1/4$. The critical scaling C^l follows immediately from this. $\qquad\square$

C.3 Gaussian Location Model - Asymptotic Variance for estimator with a Mixture of Gaussian RBF Kernels

Proof. A straightforward calculation yields $\overline{M}_T = \sum_{s=1}^{S} \gamma_s \overline{M}_s = 0$. Also, $g(\theta^*) = (2\pi)^{-\frac{d}{2}} \sum_{s=1}^{S} \gamma_s (l_s^2 + 2\sigma^2)^{-\frac{d}{2}-1} I_{d \times d}$. Furthermore, we have:

$$
\begin{aligned}
\Sigma_T &= \int_{\mathcal{U}} \left[\sum_{s=1}^{S} \gamma_s \int_{\mathcal{U}} \nabla_1 k_s(u,v) \mathbb{U}(\mathrm{d}u) \right] \otimes \left[\sum_{s'=1}^{S} \gamma_{s'} \int_{\mathcal{U}} \nabla_1 k_{s'}(w,v) \mathbb{U}(\mathrm{d}w) \right] \mathbb{U}(\mathrm{d}v) \\
&= \int_{\mathcal{U}} \left[\sum_{s=1}^{S} \gamma_s \frac{v\phi(v;0,l_s^2+\sigma^2)}{(l_s^2+\sigma^2)} \right] \left[\sum_{s'=1}^{S} \gamma_{s'} \frac{v\phi(v;0,l_s^2+\sigma^2)}{(l_{s'}^2+\sigma^2)} \right] \phi(v;0,\sigma^2) \mathrm{d}v \\
&= \sum_{s=1}^{S} \sum_{s'=1}^{S} \frac{\gamma_s \gamma_{s'}}{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)} \int_{\mathcal{U}} (v \otimes v) \phi(v;0,l_s^2+\sigma^2)\phi(v;0,l_s^2+\sigma^2)\phi(v;0,\sigma^2) \mathrm{d}v \\
&= \sum_{s=1}^{S} \sum_{s'=1}^{S} \frac{\gamma_s \gamma_{s'} \phi(0;0,2\sigma^2+l_s^2+l_{s'}^2)\phi\left(0;0,\frac{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)}{(2\sigma^2+l_s^2+l_{s'}^2)}+\sigma^2\right)}{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)} \\
&\quad \times \int_{\mathcal{U}} (v \otimes v) \phi\left(v;0,\frac{\sigma^2(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)}{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)+\sigma^2(2\sigma^2+l_s^2+l_{s'}^2)}\right) \mathrm{d}v \\
&= \sum_{s=1}^{S} \sum_{s'=1}^{S} \frac{\gamma_s \gamma_{s'} (2\pi)^{-d} \left((l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)+\sigma^2(2\sigma^2+l_s^2+l_{s'}^2) \right)^{-\frac{d}{2}}}{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)} \\
&\quad \times \left(\frac{\sigma^2(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)}{(l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)+\sigma^2(2\sigma^2+l_s^2+l_{s'}^2)} \right) I_{d \times d} \\
&= \sum_{s=1}^{S} \sum_{s'=1}^{S} \gamma_s \gamma_{s'} (2\pi)^{-d} \sigma^2 \left((l_s^2+\sigma^2)(l_{s'}^2+\sigma^2)+\sigma^2(2\sigma^2+l_s^2+l_{s'}^2) \right)^{-\frac{d}{2}-1} I_{d \times d}
\end{aligned}
$$

Combining the above in the formula $C_T = g(\theta^*)^{-1} \Sigma_T g(\theta^*)^{-1}$ gives the answers. $\qquad\square$

C.4 Gaussian Location Model - Robustness with Mixture of Gaussian RBF Kernels

Proof. Following the lines of Proposition 5 we obtain

$$\nabla_\theta \text{MMD}_T^2(\mathbb{P}_\theta, \delta_z) = \nabla_\theta \sum_{s=1}^S \gamma_s \left[\int_{\mathcal{U}} \int_{\mathcal{U}} k_s(G_\theta(u), G_\theta(v)) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v) - 2 \int_{\mathcal{U}} k_s(G_\theta(u), z) \mathbb{U}(\mathrm{d}u) \right]$$

$$= -2 \sum_{s=1}^S \gamma_s \int_{\mathcal{U}} \nabla_\theta k_s(G_\theta(u), z) \mathbb{U}(\mathrm{d}u) = -2 \sum_{s=1}^S \gamma_s \int_{\mathcal{U}} \nabla_\theta G_\theta(u) \nabla_1 k_s(G_\theta(u), z) \mathbb{U}(\mathrm{d}u)$$

$$= -2 \sum_{s=1}^S \gamma_s \int_{\mathcal{U}} \frac{(u - (z - \theta))}{l_s^2} \phi(u, z - \theta, l_s^2) \phi(u; 0, \sigma^2) \mathrm{d}u$$

$$= -2 \sum_{s=1}^S \gamma_s \int_{\mathcal{U}} \frac{(u - (z - \theta))}{l_s^2} \phi(z, \theta, l_s^2 + \sigma^2) \phi\left(u; \frac{(z - \theta)\sigma^2}{l_s^2 + \sigma^2}, \frac{l_s^2 \sigma^2}{l_s^2 + \sigma^2} \right) \mathrm{d}u$$

$$= 2 \sum_{s=1}^S \gamma_s \phi\left(z; \theta, l_s^2 + \sigma^2 \right) \frac{1}{(l_s^2 + \sigma^2)} (z - \theta)$$

$$= 2(2\pi)^{-\frac{d}{2}} \sum_{s=1}^S \gamma_s (l_s^2 + \sigma^2)^{-\frac{d}{2}-1} \exp\left(-\frac{\|z - \theta\|_2^2}{2(l_s^2 + \sigma^2)} \right) (z - \theta)$$

We conclude using the derivation of $g(\theta^*)$ in the previous proof and using the definition of influence function.

$\square$

C.5 Gaussian Scale Model: Asymptotic Variance Calculation for a single Gaussian kernel

Proof. Clearly, $\nabla_\theta G(u) = e^\theta u$. Define $s = e^{2\theta^*}$, then the metric tensor at θ^* is given by

$$g(\theta^*) = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G_{\theta^*}(u) \cdot \nabla_1 \nabla_2 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(v) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v)$$

$$= \int_{\mathcal{U}} \int_{\mathcal{U}} e^{\theta^*} u \cdot \nabla_1 \nabla_2 k(e^{\theta^*} u, e^{\theta^*} v) e^{\theta^*} v \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v)$$

$$= \int_{\mathcal{U}} \int_{\mathcal{U}} x \cdot \nabla_1 \nabla_2 k(x, y) y \phi(x; 0, s) \phi(y; 0, s) \mathrm{d}x \mathrm{d}y$$

$$= \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla \cdot (x\phi(x; 0, s)) \, k(x, y) \nabla \cdot (y\phi(y; 0, s)) \, \mathrm{d}x \mathrm{d}y$$

$$= \int_{\mathcal{U}} \int_{\mathcal{U}} \left(d - \frac{|x|^2}{s} \right) \phi(x; y, l^2) \left(d - \frac{|y|^2}{s} \right) \phi(x; 0, s) \phi(y; 0, s) \mathrm{d}x \mathrm{d}y$$

$$= A_1 + A_2 + A_3,$$

where

$$A_1 = d^2 \int_{\mathcal{U}} \int_{\mathcal{U}} \phi(x; y, l^2)\phi(x; 0, s)\phi(y; 0, s)\mathrm{d}x\mathrm{d}y$$

$$= d^2 \int_{\mathcal{U}} \int_{\mathcal{U}} \phi\left(x; \frac{ys}{l^2+s}, \frac{l^2 s}{l^2+s}\right) \phi\left(y; 0, l^2+s\right) \phi(y; 0, s)\mathrm{d}x\mathrm{d}y$$

$$= d^2 \int_{\mathcal{U}} \phi\left(y; 0, l^2+s\right) \phi(y; 0, s)\mathrm{d}y$$

$$= d^2 \int_{\mathcal{U}} \phi\left(y; 0, \frac{(l^2+s)s}{l^2+2s}\right) \phi(0; 0, l^2+2s)\mathrm{d}y$$

$$= d^2(2\pi)^{-\frac{d}{2}}(l^2+2s)^{-\frac{d}{2}}.$$

$$A_2 = -\frac{2d}{s} \int_{\mathcal{U}} \int_{\mathcal{U}} |x|^2 \phi\left(x; \frac{ys}{l^2+s}, \frac{l^2 s}{l^2+s}\right) \phi\left(y; 0, l^2+s\right) \phi\left(y; 0, s\right) \mathrm{d}x\mathrm{d}y$$

$$= -\frac{2d}{s} \int_{\mathcal{U}} \left(\frac{dl^2 s}{l^2+s} + \frac{|y|^2 s^2}{(l^2+s)^2}\right) \phi\left(y; 0, l^2+s\right) \phi\left(y; 0, s\right) \mathrm{d}y$$

$$= -\frac{2d}{l^2+s} \int_{\mathcal{U}} \left(dl^2 + \frac{|y|^2 s}{(l^2+s)}\right) \phi\left(y; 0, \frac{(l^2+s)s}{l^2+2s}\right) \phi\left(0; 0, l^2+2s\right) \mathrm{d}y$$

$$= -\frac{2d^2}{l^2+s}(l^2+2s)^{-d/2}(2\pi)^{-d/2}\left[l^2 + \frac{s^2}{l^2+2s}\right].$$

$$A_3 = \frac{1}{s^2} \int_{\mathcal{U}} \int_{\mathcal{U}} |x|^2 |y|^2 \phi\left(x; \frac{ys}{l^2+s}, \frac{l^2 s}{l^2+s}\right) \phi\left(y; 0, l^2+s\right) \phi(y; 0, s)\mathrm{d}y\mathrm{d}x$$

$$= \frac{1}{s^2} \int_{\mathcal{U}} \left(\frac{dl^2 s}{l^2+s} + \frac{|y|^2 s^2}{(l^2+s)^2}\right) |y|^2 \phi\left(y; 0, l^2+s\right) \phi(y; 0, s)\mathrm{d}y$$

$$= \frac{1}{s^2} \int_{\mathcal{U}} \left(\frac{dl^2 s}{l^2+s} + \frac{|y|^2 s^2}{(l^2+s)^2}\right) |y|^2 \phi\left(y; 0, \frac{(l^2+s)s}{l^2+2s}\right) \phi(0; 0, l^2+2s)\mathrm{d}y$$

$$= (2\pi)^{-d/2}(l^2+2s)^{-d/2}\left[\frac{d^2 l^2}{l^2+2s} + (d^2+2d)\frac{s^2}{(l^2+2s)^2}\right].$$

It follows that $g(\theta^*) = (2\pi)^{-d/2}(l^2+2s)^{-d/2}d^2 K(d, l, s)$, where $K(d, l, s)$ is bounded with respect to d, l and s and $K(d, 0, s) = (1 + 2d^{-1})/4$.

Now consider the term

$$\int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) \mathbb{U}(\mathrm{d}v)$$

$$= \int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(x), G_{\theta^*}(y)) e^{\theta^*} x \phi(x; 0, 1) \mathrm{d}x$$

$$= \int_{\mathcal{U}} \nabla_1 k(x, G_{\theta^*}(y)) x \phi(x; 0, s) \mathrm{d}x$$

$$= - \int_{\mathcal{U}} k(x, G_{\theta^*}(y)) \nabla \cdot (x \phi(x; 0, s)) \, \mathrm{d}x$$

$$= - \int_{\mathcal{U}} \phi(x; G_{\theta^*}(y), l^2) \left(d - \frac{|x|^2}{s} \right) \phi(x; 0, s) \mathrm{d}x$$

$$= - \int_{\mathcal{U}} \left(d - \frac{|x|^2}{s} \right) \phi \left(x; \frac{G_{\theta^*}(y)s}{s + l^2}, \frac{sl^2}{s + l^2} \right) \phi(G_{\theta^*}(y); 0, s + l^2) \mathrm{d}x$$

$$= - \left(\frac{ds}{s + l^2} - \frac{|G_{\theta^*}(y)|^2 s}{(s + l^2)^2} \right) \phi(G_{\theta^*}(y); 0, s + l^2).$$

Then

$$\overline{M} = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(x), G_{\theta^*}(y)) \nabla_\theta G_{\theta^*}(x) \mathbb{U}(\mathrm{d}x) \mathbb{U}(\mathrm{d}y)$$

$$= - \int_{\mathcal{U}} \left(\frac{ds}{s + l^2} - \frac{|G_{\theta^*}(y)|^2 s}{(s + l^2)^2} \right) \phi(G_{\theta^*}(y); 0, s + l^2) \mathbb{U}(\mathrm{d}y)$$

$$= - \int_{\mathcal{U}} \left(\frac{ds}{s + l^2} - \frac{|y|^2 s}{(s + l^2)^2} \right) \phi(y; 0, s + l^2) \phi(y; 0, s) \mathrm{d}y$$

$$= - \int_{\mathcal{U}} \left(\frac{ds}{s + l^2} - \frac{|y|^2 s}{(s + l^2)^2} \right) \phi \left(y; 0, \frac{(s + l^2)s}{l^2 + 2s} \right) \phi \left(0; 0, l^2 + 2s \right) \mathrm{d}y$$

$$= - (2\pi)^{-d/2} (l^2 + 2s)^{-d/2} \frac{ds}{s + l^2} \left(1 - \frac{s}{l^2 + 2s} \right).$$

It follows that

$$\int_{\mathcal{U}} \left(\int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) \mathbb{U}(\mathrm{d}u) \right)^2 \mathbb{U}(\mathrm{d}v)$$

$$= \int_{\mathcal{U}} \left(\frac{ds}{s + l^2} - \frac{|G_{\theta^*}(y)|^2 s}{(s + l^2)^2} \right)^2 \phi^2(G_{\theta^*}(y); 0, s + l^2) \phi(y; 0, 1) \mathrm{d}y$$

$$= \int_{\mathcal{U}} \left(\frac{ds}{s + l^2} - \frac{|y|^2 s}{(s + l^2)^2} \right)^2 \phi^2(y; 0, s + l^2) \phi(y; 0, s) \mathrm{d}y$$

Expanding the terms we have

$$\int \left(\frac{ds}{s+l^2} - \frac{|y|^2 s}{(s+l^2)^2} \right)^2 \phi^2(y;0,s+l^2)\phi(y;0,s)\mathrm{d}y$$

$$= \int \left(\frac{ds}{s+l^2} - \frac{|y|^2 s}{(s+l^2)^2} \right)^2 \phi\left(y;0,\frac{s+l^2}{2} \right) \phi(y;0,s)\mathrm{d}y\phi\left(0;0,2(l^2+s) \right)$$

$$= \int \left(\frac{ds}{s+l^2} - \frac{|y|^2 s}{(s+l^2)^2} \right)^2 \phi\left(y;0,\frac{(s+l^2)s}{l^2+3s} \right) \mathrm{d}y\phi\left(0;0,(l^2+3s)/2 \right) \phi\left(0;0,2(l^2+s) \right)$$

$$= \left(\frac{s^2d^2}{(s+l^2)^2} - 2\frac{d^2 s^3}{(l^2+3s)(s+l^2)^2} + \frac{(d^2+2d)s^2}{(s+l^2)^2}\frac{s^2}{(l^2+3s)^2} \right) \phi\left(0;0,(l^2+3s)/2 \right) \phi\left(0;0,2(l^2+s) \right)$$

$$= \frac{d^2 s^2}{(s+l^2)^2}\left(1 - 2\frac{s}{l^2+3s} + \frac{(1+2d^{-1})s^2}{(l^2+3s)^2} \right)(2\pi)^{-d}(l^2+3s)^{-d/2}(l^2+s)^{-d/2}.$$

It follows that

$$\Sigma = \int_{\mathcal{U}} \left(\int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(u),G_{\theta^*}(v))\nabla_\theta G_{\theta^*}(u)\mathbb{U}(\mathrm{d}u) \right)^2 \mathbb{U}(\mathrm{d}v) - \overline{M}^2$$

$$= (2\pi)^{-d}\frac{d^2 s^2}{(s+l^2)^2}\left[C_1(s,l,d)(l^2+3s)^{-d/2}(l^2+s)^{-d/2} - C_2(s,l,d)(l^2+2s)^{-d} \right],$$

where the terms

$$C_1(s,l,d) = \left(1 - 2\frac{s}{l^2+3s} + \frac{(1+2d^{-1})s^2}{(l^2+3s)^2} \right) \quad \text{and} \quad C_2(s,l,d) = \left(1 - \frac{s}{l^2+2s} \right)^2,$$

are bounded uniformly with respect to s,l,d. The asymptotic variance of the estimator $\hat\theta_m$ is then given by $C = g^{-1}(\theta^*)\Sigma g^{-1}(\theta^*)$, as stated. $\qquad\square$

C.6 Proof of Proposition 7

Proof. The asymptotic variance can be written as

$$C = \frac{(l^2+2s)^2\left((l^2+s)^{-\frac{d}{2}-2}(l^2+2s)^{d+2}(l^2+3s)^{-\frac{d}{2}-2}\left((l^2+2s)^2+2s^2/d \right)-1 \right)}{(d+2)^2 s^2}.$$

Let $l = d^\alpha$, then it is a straightforward calculation to show that the term

$$E(d) := \left(\frac{1+\frac{s}{d^{2\alpha}}}{1+\frac{2s}{d^{2\alpha}}} \right)^{-d/2}\left(\frac{1+\frac{3s}{d^{2\alpha}}}{1+\frac{2s}{d^{2\alpha}}} \right)^{-d/2} = \left(\frac{1+\frac{4s}{d^{2\alpha}}+\frac{3s^2}{d^{4\alpha}}}{1+\frac{4s}{d^{2\alpha}}+\frac{4s^2}{d^{4\alpha}}} \right)^{-d/2},$$

converges to 1 if $\alpha > \frac{1}{4}$, $s^2/2$ if $\alpha = 1/4$ and ∞ if $\alpha < 1/4$. Moreover, the convergence in each case is exponentially fast. We can express the asymptotic variance as $C = (E(d)-1)B(d) + B(d) - 1/A(d)$, where $A(d) = ((d+2)^2 s^2)/(d^{2\alpha}+2s)^2$ and

$$B(d) = \frac{\left(d^{2\alpha}+2s \right)^2\left((d^{2\alpha}+2s)^2+\frac{2s^2}{d} \right)}{(d^{2\alpha}+s)^2(d^{2\alpha}+3s)^2} = \frac{\left(2sd^{-2\alpha}+1 \right)^2\left(2s^2 d^{-4\alpha-1}+(2sd^{-2\alpha}+1)^2 \right)}{(sd^{-2\alpha}+1)^2(3sd^{-2\alpha}+1)^2},$$

52

By the mean value theorem, $B(d) - 1$ is $O(d^{-1-4\alpha})$. Since $1/A(d)$ is $O(d^{4\alpha-2})$, it follows that $(B(d) - 1)/A(d)$ is $O(d^{-1})$. It follows that C converges to zero for $\alpha > 1/4$ and to infinity for $\alpha < 1/4$. When $\alpha = 1/4$ since $B(d)/A(d) = O(d^{-1})$ it follows that $C \to 0$, completing the proof. $\qquad\square$

D   Additional Details for Numerical Experiments

In this section, we provide additional details and simulation results for experiments in the paper.

D.1   Gaussian Distributions

In this subsection we extend Figure 1 for the Gaussian location model with different classes of kernels. In Figure 8 we plot the loss landscape in each case for different dimensions and different parameter choices. We note that the inverse multiquadric kernel suffers less from vanishing gradients. In Figure 9 and 10 we compute the error in estimating the parameter of a Gaussian location model, as a function of the location of the Dirac contamination and the percentage of corrupted samples. As the estimator is qualitatively robust, this influence will be bounded independently of this location, but the maximum error will depend strongly on the choice of kernel and kernel parameters.

Finally, Figure 11 provides plots demonstrating the strong lack of robustness of the Sinkhorn algorithm as studied in Genevay et al. [2018]. These results demonstrate that this lack of robustness occurs for a large range of regularisation parameter ϵ. The experiments were performed using an l_2 cost, which is standard in this literature. Other cost functions could potentially be used to improve the robustness of this estimator, but this is currently an open question.

D.2   G-and-k Distribution

In order to implement MMD estimators, we will need to have access to derivatives of the generator, which are given as follows: $\partial G_\theta(u)/\partial\theta_1 = 1$ and

$$
\frac{\partial G_\theta(u)}{\partial\theta_2} = \left(1 + \frac{4}{5}\frac{\left(1 - \exp(-\theta_3 z(u))\right)}{\left(1 + \exp(-\theta_3 z(u))\right)}\right)\left(1 + z(u)^2\right)^{\theta_4} z(u)
$$

$$
\frac{\partial G_\theta(u)}{\partial\theta_3} = \frac{8}{5}\theta_2\frac{\exp(\theta_3 z(u))}{\left(1 + \exp(\theta_3 z(u))\right)^2}\left(1 + z(u)^2\right)^{\theta_4} z(u)^2
$$

$$
\frac{\partial G_\theta(u)}{\partial\theta_4} = \theta_2\left(1 + 0.8\frac{\left(1 - \exp(-\theta_3 z(u))\right)}{\left(1 + \exp(-\theta_3 z(u))\right)}\right)\left(1 + z(u)^2\right)^{\theta_4}\log(1 + z(u)^2)z(u)
$$

Note that these could also be obtained using automatic differentiation.

D.3   Stochastic Volatility Model

We can see the stochastic volatility model as a generative model with parameters $\theta = (\theta_1, \theta_2, \theta_3)$, which maps a sample $u = (u_0, u_1, \ldots, u_{2T})$ which is $\mathcal{N}(0, I_{2T\times 2T})$ distributed to a realisation $y = (y_1, \ldots, y_T)$ of the stochastic volatility model. Here, $\epsilon_t = u_t$ for $t \geq 1$, $h_1 =$
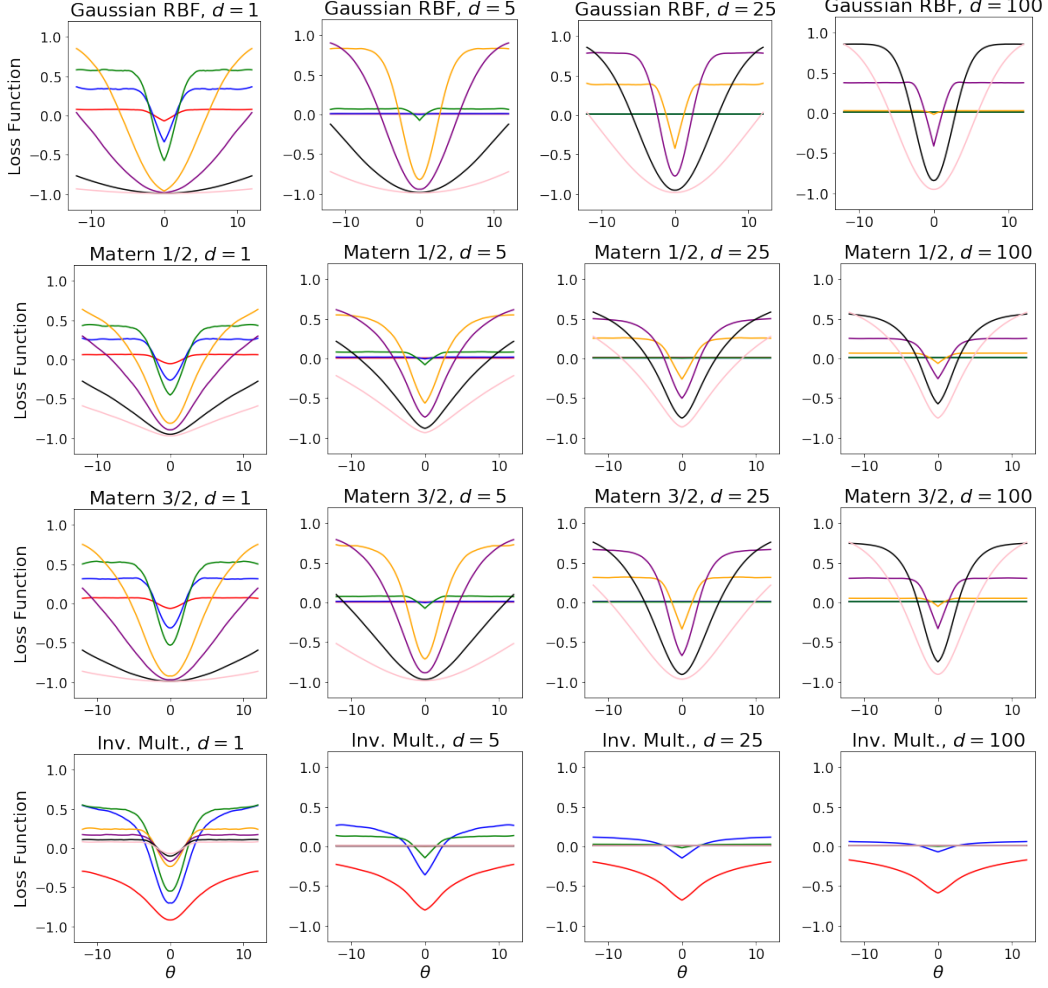
Figure 8: MMD Loss landscape for the Gaussian location model in dimensions $d = 1, 5, 25, 100$. The landscape is plotted for varying choices of kernels including a Gaussian RBF kernel, a Matérn kernel with smoothness $\frac{1}{2}$ or $\frac{3}{2}$ and an inverse-multiquadric kernel. For each kernel, we plot the loss function for varying values of the lengthscale parameter including $l = 0.1$ (red), $l = 0.5$ (blue), $l = 1$ (green), $l = 5$ (orange), $l = 10$ (purple), $l = 25$ (black) and $l = 50$ (pink).

$u_{T+1}\sqrt{\sigma^2/(1-\phi^2)}$ and $\eta_t = \sigma u_{T+t}$ for all $t \geq 2$. Note that it is possible to go back to the original parameterisation using $\phi = (\exp(\theta_1) - 1)/(\exp(\theta_1) + 1)$, $\kappa = \exp(\theta_2)$ and $\sigma = \exp(\theta_3/2)$.

We can obtain the derivative process as follows: $\partial_{\theta_1} y_t = y_t(\partial_{\theta_1} h_t)/2$, $\partial_{\theta_1} h_1 = [(\exp(\theta_1/2) - \exp(-\theta_1/2))/(\exp(\theta_1/2) + \exp(-\theta_1/2))](h_1/2)$, $\partial_{\theta_1} h_t = (\partial_{\theta_1}\phi)h_{t-1} + \phi(\partial_{\theta_1} h_{t-1})$ for $t > 1$, $\partial_{\theta_1}\phi = 2\exp(\theta_1)/(\exp(\theta_1) + 1)^2$, $\partial_{\theta_2} y_t = y_t$, $\partial_{\theta_2} h_t = 0$, $\partial_{\theta_3} y_t = y_t(\partial_{\theta_3} h_t)/2$, $\partial_{\theta_3} h_1 = h_1/2$, $\partial_{\theta_3} h_t = \phi(\partial_{\theta_3} h_{t-1}) + (\partial_{\theta_3}\eta_t)$ for $t > 1$ and $\partial_{\theta_3}\eta_t = \eta_t/2$.
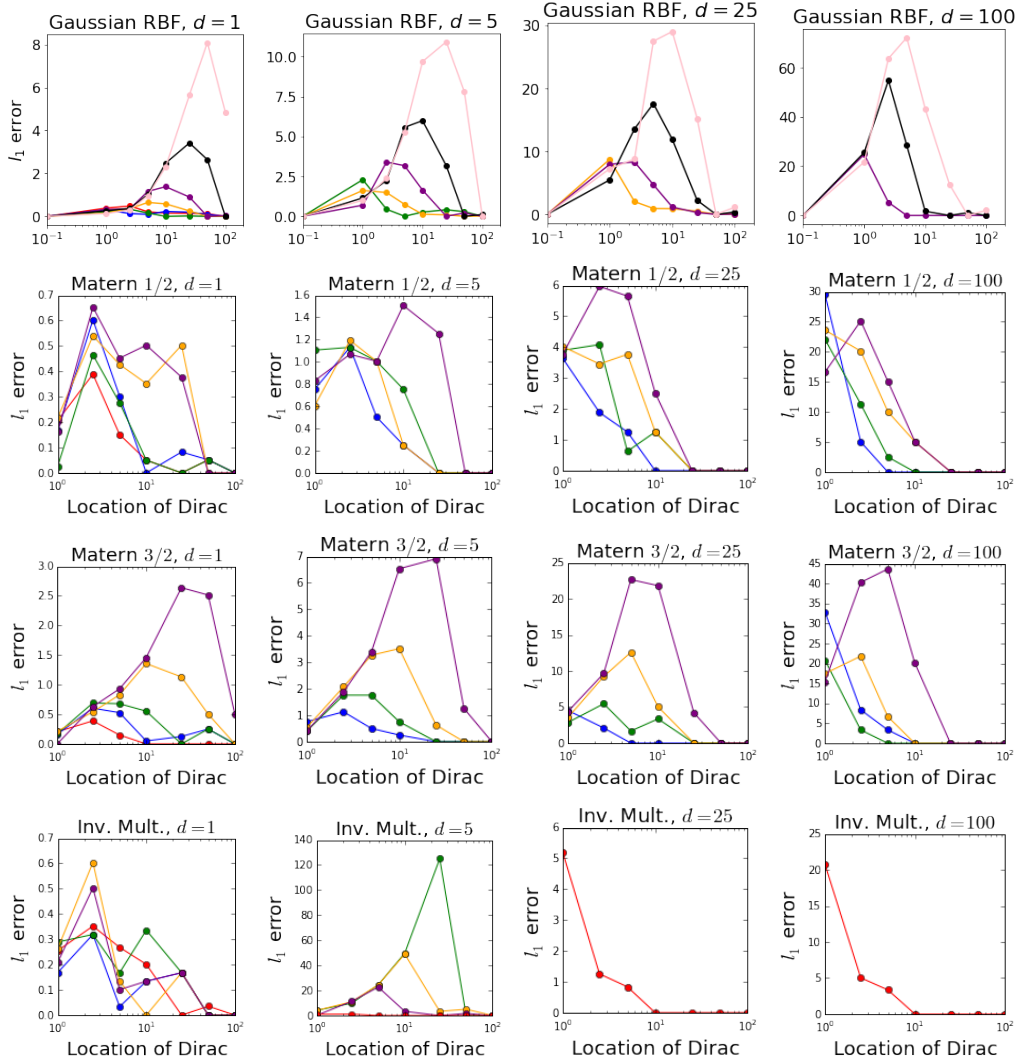
Figure 9: Gaussian distribution with unknown mean: Robustness as a function of the location of the Dirac for varying kernel and kernel lengthscales in dimensions $d = 1, 5, 25, 100$.

D.4 Stochastic Lotka-Volterra Model

Besides simulating $X_{1,t}$ and $X_{2,t}$ we also required the coupled matrix diffusion process J_t taking values in $\mathbb{R}^{2 \times 2}$ which satisfies the following SDE:

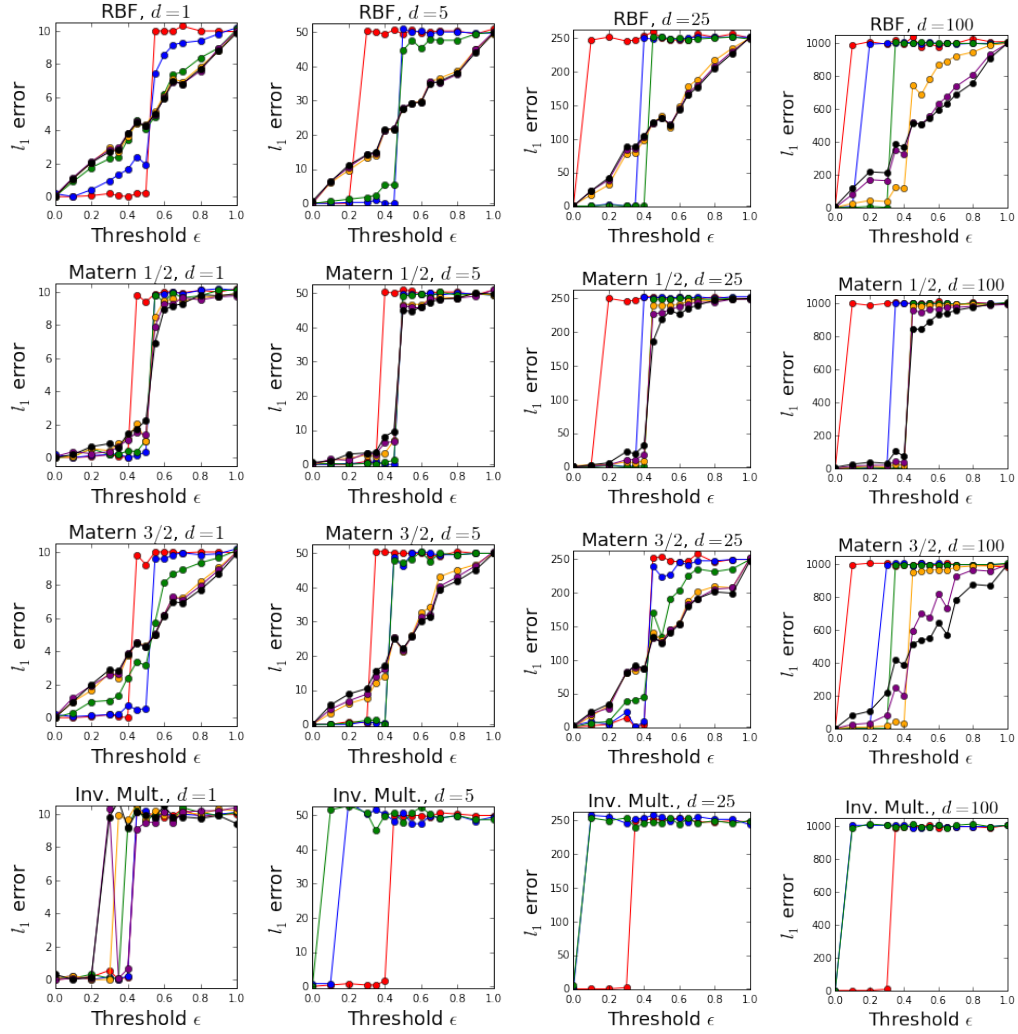$$dJ_t = J_t A(X_{1,t}, X_{2,t}) \, dt + \sum_{i=1}^{3} J_t B_i(X_{1,t}, X_{2,t}) dW_{i,t},$$

Figure 10: Gaussian distribution with unknown mean: Error in estimator as a function of the threshold ϵ for varying kernel and kernel lengthscales in dimensions $d = 1, 5, 25, 100$.

where

$$A(x,y) = \begin{pmatrix} c_1 - c_2 y & c_2 x \\ c_2 y & c_2 x - c_3 \end{pmatrix}, \qquad B_1(x,y) = \frac{\sqrt{c_1}}{2} \begin{pmatrix} \frac{1}{\sqrt{x}} & 0 \\ 0 & 0 \end{pmatrix},$$

$$B_2(x,y) = \frac{\sqrt{c_2}}{2} \begin{pmatrix} -\sqrt{\frac{y}{x}} & -\sqrt{\frac{x}{y}} \\ \sqrt{\frac{y}{x}} & \sqrt{\frac{x}{y}} \end{pmatrix}, \qquad B_3(x,y) = \frac{\sqrt{c_3}}{2} \begin{pmatrix} 0 & 0 \\ 0 & -\frac{1}{\sqrt{x}} \end{pmatrix},$$

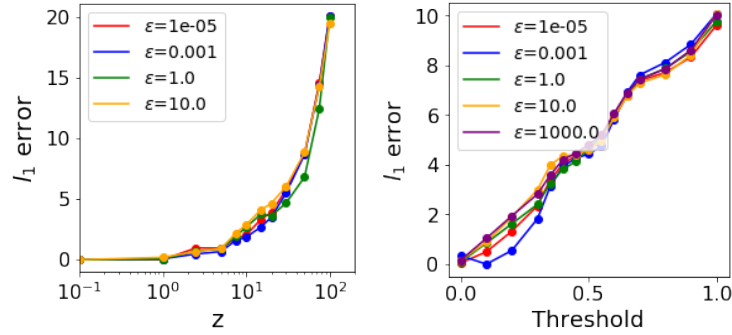and subject to the initial condition $J_0 = I_{2\times 2}$.

Figure 11: *Gaussian location models - Performance of Sinkhorn Estimators in $d = 1$ for varying values of the regularisation parameter ϵ. Left:* l_1 error as a function of the location of the Dirac. *Right:* l_1 error as a function of the percentage of corrupted data points.
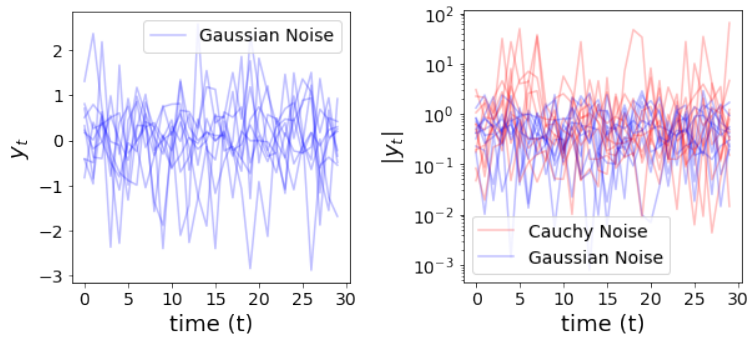

Figure 12: *Realisations from the stochastic volatility model.* Left: 10 realisations from the assumed model $\mathbb{P}_{\theta*}$ (i.e. stochastic volatility model with Gaussian noise). Right: Absolute value of these same realisations and 10 realisations from the data generating process $\mathbb{Q}$ (i.e. stochastic volatility model with Cauchy noise).