Fast and Scalable Score-Based Calibration Tests (Supplementary Material)

A CONDITIONAL GOODNESS-OF-FIT: GENERAL OPERATOR-VALUED KERNEL

Assume that

- kernel $l \in \mathcal{C}^2(\mathcal{Y} \times \mathcal{Y}, \mathbb{R})$,
- densities $P_{|x} \in C^1(\mathcal{Y}, \mathbb{R})$ for $\mathbb{P}(X)$-almost all x, and that
- $\mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \left\| K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right\|_{\mathcal{F}_K} < \infty$.

Due to the Bochner integrability of $(x, y) \mapsto K_{P_{|x}} \xi_{P_{|x}}(y, \cdot)$ expectation and inner product commute [see Andreas Christmann, 2008, Definition A.5.20], and hence we have

$$\begin{aligned}
C_{P_{|\cdot}}(\mathbb{P}) &= \left\| \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \left[K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right] \right\|^2_{\mathcal{F}_K} \\
&= \left\langle \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \left[K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right], \mathbb{E}_{(x',y')\sim\mathbb{P}(X,Y)} \left[K_{P_{|x'}} \xi_{P_{|x'}}(y', \cdot) \right] \right\rangle_{\mathcal{F}_K} \\
&= \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \mathbb{E}_{(x',y')\sim\mathbb{P}(X,Y)} \left\langle K_{P_{|x}} \xi_{P_{|x}}(y, \cdot), K_{P_{|x'}} \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_K} \\
&= \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \mathbb{E}_{(x',y')\sim\mathbb{P}(X,Y)} \left\langle K^*_{P_{|x'}} K_{P_{|x}} \xi_{P_{|x}}(y, \cdot), \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}},
\end{aligned}$$

where $K^*_{P_{|x'}}$ is the adjoint of $K_{P_{|x'}}$. The reproducing property implies $K^*_{P_{|x'}} K_{P_{|x}} = K(P_{|x}, P_{|x'})$, and therefore we get

$$\begin{aligned}
C_{P_{|\cdot}}(\mathbb{P}) &= \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \mathbb{E}_{(x',y')\sim\mathbb{P}(X,Y)} \left\langle K(P_{|x}, P_{|x'}) \xi_{P_{|x}}(y, \cdot), \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}} \\
&= \mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)} \mathbb{E}_{(x',y')\sim\mathbb{P}(X,Y)} H((P_{|x}, y), (P_{|x'}, y'))
\end{aligned}$$

where

$$\begin{aligned}
H((p, y), (p', y')) &\coloneqq \left\langle K(p, p') \xi_p(y, \cdot), \xi_{p'}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}} \\
&= \left\langle K(p, p') \xi_p(y, \cdot), l(y', \cdot) \nabla_{y'} \log f_{p'}(y') + \nabla_{y'} l(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}}.
\end{aligned}$$

For $i \in \{1, \dots, d_y\}$, let $\mathrm{proj}_i \colon \mathcal{F}_l^{d_y} \to \mathcal{F}_l$ be the projection map to the ith subspace of the product space $\mathcal{F}_l^{d_y}$, and similarly let $\iota_i \colon \mathcal{F}_l \to \mathcal{F}_l^{d_y}$ be the embedding of $\mathcal{F}_l$ in the ith subspace of $\mathcal{F}_l^{d_y}$ via $x \mapsto (0, \dots, 0, x, 0, \dots, 0)$. Then we can write

$$\begin{aligned} H((p,y),(p',y')) &= \sum_{i=1}^{d_y} \left\langle \mathrm{proj}_i K(p,p')\xi_p(y,\cdot), l(y',\cdot)\frac{\partial}{\partial y_i'} \log f_{p'}(y') + \frac{\partial}{\partial y_i'} l(y',\cdot) \right\rangle_{\mathcal{F}_l} \\ &= \sum_{i=1}^{d_y} \left[(\mathrm{proj}_i K(p,p')\xi_p(y,\cdot))(y') \frac{\partial}{\partial y_i'} \log f_{p'}(y') + \frac{\partial}{\partial y_i'} (\mathrm{proj}_i K(p,p')\xi_p(y,\cdot))(y') \right]. \end{aligned}$$

Since $K(p,p') \in \mathcal{L}(\mathcal{F}_l^{d_y})$ is a linear operator, we have

$$K(p,p')\xi_p(y,\cdot) = K(p,p')(l(y,\cdot)\nabla_y \log f_p(y)) + K(p,p')\nabla_y l(y,\cdot).$$

For $1 \le i,j \le d_y$, define $K_{i,j}(p,p') \colon \mathcal{F}_l \to \mathcal{F}_l$ as the continuous linear operator

$$K_{i,j}(p,p') := \mathrm{proj}_i K(p,p')\iota_j.$$

Thus we have

$$\mathrm{proj}_i K(p,p')\xi_p(y,\cdot) = \sum_{j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] K_{i,j}(p,p')l(y,\cdot) + \sum_{j=1}^{d_y} \frac{\partial}{\partial y_j} K_{i,j}(p,p')l(y,\cdot),$$

and therefore

$$(\mathrm{proj}_i K(p,p')\xi_p(y,\cdot))(y') = \sum_{j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log p(y) \right] (K_{i,j}(p,p')l(y,\cdot))(y') + \sum_{j=1}^{d_y} \frac{\partial}{\partial y_j} (K_{i,j}(p,p')l(y,\cdot))(y').$$

Due to the differentiability of kernel l we can interchange inner product and differentiation [Andreas Christmann, 2008, Lemma 4.34], and thus we obtain

$$\begin{aligned} H((p,y),(p',y')) = &\sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] \left[\frac{\partial}{\partial y_i'} \log f_{p'}(y') \right] (K_{i,j}(p,p')l(y,\cdot))(y') \\ &+ \sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y_i'} \log f_{p'}(y') \right] \frac{\partial}{\partial y_j} (K_{i,j}(p,p')l(y,\cdot))(y') \\ &+ \sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] \frac{\partial}{\partial y_i'} (K_{i,j}(p,p')l(y,\cdot))(y') \\ &+ \sum_{i,j=1}^{d_y} \frac{\partial}{\partial y_i'} \frac{\partial}{\partial y_j} (K_{i,j}(p,p')l(y,\cdot))(y'), \end{aligned}$$

Define $A \colon (P_{|\mathcal{X}} \times \mathcal{Y})^2 \to \mathbb{R}^{d_y \times d_y}$ by

$$[A((p,y),(p',y'))]_{i,j} := (K_{i,j}(p,p')l(y,\cdot))(y') \qquad (1 \le i,j \le d_y).$$

Thus we obtain

$$H((p,y),(p',y')) = (s_{p'}(y') + \nabla_{y'})^\top A((p,y),(p',y'))(s_p(y) + \nabla_y), \tag{A.1}$$

where for $x, x' \in \mathbb{R}^d$, $M(x,x') \in \mathbb{R}^{d \times d}$ we use the notation

$$\nabla_x^\top M(x,x') = \begin{bmatrix} \nabla_x^\top [M(x,x')]_{:,1} & \cdots & \nabla_x^\top [M(x,x')]_{:,d} \end{bmatrix} = \begin{bmatrix} \mathrm{div}_x [M(x,x')]_{:,1} & \cdots & \mathrm{div}_x [M(x,x')]_{:,d} \end{bmatrix},$$

and similarly

$$M(x,x')\nabla_{x'} = \left(\nabla_{x'}^\top M(x,x')^\top \right)^\top = \begin{bmatrix} \mathrm{div}_{x'} [M(x,x')]_{1,:} & \cdots & \mathrm{div}_{x'} [M(x,x')]_{d,:} \end{bmatrix}^\top$$

and

$$\nabla_x^\top M(x,x')\nabla_{x'} = \nabla_x^\top (M(x,x')\nabla_{x'}^\top) = \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x'_j}[M(x,x')]_{i,j}.$$

Thus, given samples $\{(P_{|x^i}, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$, an unbiased estimator of statistic $C_{P_{|\cdot}}(\mathbb{P})$ is

$$\widehat{C_{P_{|\cdot}}} = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} H((P_{|x^i}, y^i), (P_{|x^j}, y^j)),$$

where H is given by Equation (A.1).

If kernel K is of the form in **??**, we recover the simpler formula in **??**. In this case $A((p,y),(p',y')) = k(p,p')l(y,y')I_{d_y} \in \mathbb{R}^{d_y \times d_y}$, i.e., A is a scaled identity matrix.

B KCCSD AS A SPECIAL CASE OF SKCE

We prove the following general lemma that establishes the KCSD as a special case of the MMD. Then **??** follows immediately by considering random variables $Z = P_{|X}$ and Y, and models $Q_{|z} = z = P_{|x}$.

Lemma B.1 (KCSD as a special case of the MMD)**.** *Let $Q_{|z}$ be models of the conditional distributions $\mathbb{P}(Y \in \cdot \mid Z = z)$. Moreover, we assume that*

- *$Q_{|z}$ has a density $f_{Q_{|z}} \in C^1(\mathcal{Y}, \mathbb{R})$ for $\mathbb{P}(Z)$-almost all z,*
- *kernel $l \in C^2(\mathcal{Y}\times\mathcal{Y}, \mathbb{R})$,*
- *$\mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)} \left\| K_z \xi_{Q_{|z}}(y,\cdot)\right\|_{\mathcal{F}_K} < \infty$, and*
- *$\oint_{\partial\mathcal{Y}} l(y,y') f_{Q_{|z}}(y) n(y)\,\mathrm{d}S(y') = 0$ and $\oint_{\partial\mathcal{Y}} \nabla_y l(y,y') f_{Q_{|z}}(y') n(y')\,\mathrm{d}S(y') = 0$ for $\mathbb{P}(Z)$-almost all z,*

where $n(y)$ is the unit vector normal to the boundary $\partial\mathcal{Y}$ of $\mathcal{Y}$ at $y \in \mathcal{Y}$.[1]

Then

$$D_{Q_{|\cdot}}(\mathbb{P}) = \mathrm{MMD}^2_{k_{Q_{|\cdot}}}(\mathbb{P}(Z,Y), \mathbb{P}_{Q_{|\cdot}}(Z,Y))$$

where we define distribution $\mathbb{P}_{Q_{|\cdot}}$ by

$$\mathbb{P}_{Q_{|\cdot}}(Z \in A, Y \in B) := \int_A Q_{|z}(Y \in B)\,\mathbb{P}(Z \in \mathrm{d}z)$$

and kernel $k_{Q_{|\cdot}} : (\mathcal{Z}\times\mathcal{Y})\times(\mathcal{Z}\times\mathcal{Y}) \to \mathbb{R}$ as

$$k_{Q_{|\cdot}}((z,y),(z',y')) := (s_{Q_{|z'}}(y') + \nabla_{y'})^{\mathsf{T}} A((z,y),(z',y'))(s_{Q_{|z}}(y) + \nabla_y),$$

using the same notation as in Appendix A and similarly defining $A((z,y),(z',y')) \in \mathbb{R}^{d_y\times d_y}$ by

$$[A((z,y),(z',y'))]_{i,j} := (K_{i,j}(z,z')l(y,\cdot))(y') \qquad (1 \le i,j \le d_y).$$

If K is of the form $k(\cdot,\cdot)I_{\mathcal{F}_l^{d_y}}$, function A simplifies to

$$A((z,y),(z',y')) = k(z,z')l(y,y')I_{d_y}$$

and kernel $k_{Q_{|\cdot}}$ is given by

$$k_{Q_{|\cdot}}((z,y),(z',y')) = k(z,z')\left[l(y,y')s_{Q_{|z}}(y)^{\mathsf{T}} s_{Q_{|z'}}(y') + s_{Q_{|z}}(y)^{\mathsf{T}}\nabla_{y'} l(y,y') + s_{Q_{|z'}}(y')^{\mathsf{T}}\nabla_y l(y,y') + \sum_{i=1}^{d_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(y,y')\right].$$

[1] These assumptions are not restrictive in practice since they are satisfied if the conditions of [Jitkrittum et al., 2020, Theorem 1] hold which are required to ensure that $D_{Q_{|\cdot}}(\mathbb{P}) = 0$ if and only if $Q_{|Z}(\cdot) = \mathbb{P}(Y \in \cdot|Z)$ $\mathbb{P}(Z)$-almost surely.

Proof. From a similar calculation as in Appendix A [cf. Jitkrittum et al., 2020, Section A.2] we obtain that

$$k_{Q_{|\cdot}}((z,y),(z',y')) = \left\langle K_z \xi_{Q_{|z}}(y,\cdot), K_{z'} \xi_{Q_{|z'}}(y',\cdot) \right\rangle_{\mathcal{F}_K}.$$

Thus $k_{Q_{|\cdot}}$ is an inner product of the features of (z,y) and (z',y') given by the feature map $(z,y) \mapsto K_z \xi_{Q_{|z}}(y,\cdot) \in \mathcal{F}_K$, and therefore $k_{Q_{|\cdot}}$ is a positive-definite kernel. Moreover, from our assumption we obtain

$$\mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)} \left| k_{Q_{|\cdot}}((z,y),(z,y)) \right|^{1/2} = \mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)} \left\| K_z \xi_{Q_{|z}}(y,\cdot) \right\|_{\mathcal{F}_K} < \infty.$$

Thus the mean embedding $\mu_{\mathbb{P}(Z,Y)} \in \mathcal{F}_K$ of $\mathbb{P}(Z,Y)$ exists [Gretton et al., 2012, Lemma 3].

Due to the Bochner integrability of $(z,y) \mapsto K_z \xi_{Q_{|z}}(y,\cdot)$ expectation and inner product commute [see Andreas Christmann, 2008, Definition A.5.20], and hence we have

$$\begin{aligned}
\mathbb{E}_{(z,y)\sim\mathbb{P}_{Q_{|\cdot}}(Z,Y)} \mathbb{E}_{(z',y')\sim\mathbb{P}_{Q_{|\cdot}}(Z,Y)} k_{Q_{|\cdot}}((z,y),(z',y')) &= \left\| \mathbb{E}_{(z,y)\sim\mathbb{P}_{Q_{|\cdot}}(Z,Y)} K_z \xi_{Q_{|z}}(y,\cdot) \right\|_{\mathcal{F}_K}^2 \\
&= \left\| \mathbb{E}_{z\sim\mathbb{P}(Z)} \mathbb{E}_{y\sim Q_{|z}} K_z \xi_{Q_{|z}}(y,\cdot) \right\|_{\mathcal{F}_K}^2 \\
&= \left\| \mathbb{E}_{z\sim\mathbb{P}(Z)} K_z \, \mathbb{E}_{y\sim Q_{|z}} \xi_{Q_{|z}}(y,\cdot) \right\|_{\mathcal{F}_K}^2.
\end{aligned}$$

Due to the last assumption [Chwialkowski et al., 2016, Lemma 5.1] we know that

$$\mathbb{E}_{y\sim Q_{|z}} \xi_{Q_{|z}}(y,\cdot) = 0,$$

which implies

$$\mathbb{E}_{(z,y)\sim\mathbb{P}_{Q_{|\cdot}}(Z,Y)} \mathbb{E}_{(z',y')\sim\mathbb{P}_{Q_{|\cdot}}(Z,Y)} k_{Q_{|\cdot}}((z,y),(z',y')) = 0.$$

Thus the mean embedding $\mu_{\mathbb{P}_{Q_{|\cdot}}(Z,Y)} \in \mathcal{F}_K$ of $\mathbb{P}_{Q_{|\cdot}(Z,Y)}$ exists and satisfies $\|\mu_{\mathbb{P}_{Q_{|\cdot}}(Z,Y)}\|_{\mathcal{F}_K}^2 = 0$, and hence $\mu_{\mathbb{P}_{Q|\cdot}(Z,Y)} = 0$. We obtain [Gretton et al., 2012, Lemma 4] that

$$\begin{aligned}
\mathrm{MMD}^2_{k_{Q_{|\cdot}}}(\mathbb{P}(Z,Y), \mathbb{P}_{Q_{|\cdot}}(Z,Y)) &= \|\mu_{\mathbb{P}(Z,Y)} - \mu_{\mathbb{P}_{Q_{|\cdot}}(Z,Y)}\|_{\mathcal{F}_K}^2 \\
&= \|\mu_{\mathbb{P}(Z,Y)}\|_{\mathcal{F}_K}^2 \\
&= \mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)} \mathbb{E}_{(z',y')\sim\mathbb{P}(Z,Y)} k_{Q_{|\cdot}}((z,y),(z',y')) \\
&= \mathbb{E}_{(z,y)\sim\mathbb{P}(Z,Y)} \mathbb{E}_{(z',y')\sim\mathbb{P}(Z,Y)} \left\langle K_z \xi_{Q_{|z}}(y,\cdot), K_{z'} \xi_{Q_{|z'}}(y',\cdot) \right\rangle_{\mathcal{F}_K} \\
&= D_{Q_{|\cdot}}(\mathbb{P}),
\end{aligned}$$

where the last equality follows from [Jitkrittum et al., 2020, Section A.2]. □

C CALIBRATION IMPLIES EXPECTED COVERAGE

We show that the sense of calibration employed by our tests implies posterior coverage in the sense of Hermans et al. [2021]. Again let us note $P_{|x}(\cdot)$ for a model of the conditional distribution $\mathbb{P}(Y \in \cdot \mid X = x)$. Moreover, we assume that $P_{|x}$ has a density $f_{P_{|x}}$ for $\mathbb{P}(X)$-almost every x.

For level $1-\alpha \in [0,1]$, let $\Theta_{P_{|x}}(1-\alpha)$ be the highest density region of a probabilistic model $P_{|x}$ with density $f_{P_|}$. It is defined [see, e.g., Hyndman, 1996] by

$$\Theta_{P_{|x}}(1-\alpha) \coloneqq \left\{ y \colon f_{P_{|x}}(y) \geq c_{P_{|x}}(1-\alpha) \right\}$$

where

$$c_{P_{|x}}(1-\alpha) := \sup \left\{ c \colon \int_{\left\{ \tilde{y} \colon f_{P_{|x}}(\tilde{y}) \geq c \right\}} P_{|x}(\mathrm{d}y) \geq 1-\alpha \right\}.$$

Hence, by definition [see, e.g., Hermans et al., 2021]

$$\mathbb{E}_{y\sim P_{|x}}\,\mathbb{1}\{y\in\Theta_{P_{|x}}(1-\alpha)\}=\int_{\Theta_{P_{|x}}(1-\alpha)}P_{|x}(\mathrm{d}y)\geq 1-\alpha.$$

Assume that model $P_{|\cdot}$ is calibrated. By definition, it satisfies

$$\mathbb{P}(Y\in\cdot\mid P_{|X})=P_{|X}\qquad\mathbb{P}(X)\text{-almost surely.}$$

Hence, for all $\alpha\in[0,1]$, we obtain

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathbb{P}(X,Y)}\,\mathbb{1}\{y\in\Theta_{P_{|x}}(1-\alpha)\}&=\mathbb{E}_{(P_{|x},y)\sim\mathbb{P}(P_{|X},Y)}\,\mathbb{1}\{y\in\Theta_{P_{|x}}(1-\alpha)\}\\&=\mathbb{E}_{P_{|x}\sim\mathbb{P}(P_{|X})}\,\mathbb{E}_{y\sim P_{|x}}\,\mathbb{1}\{y\in\Theta_{P_{|x}}(1-\alpha)\}\\&\geq\mathbb{E}_{P_{|x}\sim\mathbb{P}(P_{|X})}\left[1-\alpha\right]\\&=1-\alpha.\end{aligned}$$

Thus model $P_{|\cdot}$ has expected coverage for all $\alpha\in[0,1]$.

D DIFFUSION-LIMIT AND UNIVERSALITY

D.1 FISHER DIVERGENCE AS A DIFFUSION LIMIT

We recall that for a map f and a measure μ, the push-forward measure of μ by f, noted $f_{\#}\mu$, is the measure on the image space of f which verifies, for any measurable function g

$$\int g(x)\,f_{\#}\mu(\mathrm{d}x)=\int g(f(x))\,\mu(\mathrm{d}x).$$

To prove the differential inequality linking the MMD and the KGFD, we rely on the following reformulation of the Fokker-Planck equation:

$$\begin{aligned}\frac{\partial\mu(x,t)}{\partial t}&=\mathrm{div}_x(-\mu(x,t)s_p(x))+\Delta_x\mu(x,t)\\&=\mathrm{div}_x(-\mu(x,t)s_p(x))+\mathrm{div}_x\,\nabla_x\mu(x,t)\\&=\mathrm{div}_x(-\mu(x,t)s_p(x))+\mathrm{div}_x(\mu(x,t)\nabla_x\log\mu(x,t))\\&=\mathrm{div}_x(-\mu(x,t)(s_p(x)-\nabla_x\log\mu(x,t)).\end{aligned}$$

We remark that since the density $\mu(x,t)$ is twice differentiable in x and differentiable in t [Johnson, 2004], this equation holds in the strong sense, and not only in the sense of distributions. Because of that, one has

$$\partial_t\mu(x,t)=\lim_{\Delta\to 0}\frac{\mu(x,t+\Delta)-\mu(x,t)}{\Delta}.$$

Let us consider an RKHS $\mathcal{H}$ with kernel k, and let $h\in\mathcal{H}$. Let us define $m_t(x):=m(x,t):=\mu_{\nu,p}(x,t)-\mu_{\nu,q}(x,t)$ and we note $\mathrm{MMD}(m_t)$ the function given by

$$\mathrm{MMD}(m_t)=\left[\iint k(x,y)m_t(x)m_t(y)\,\mathrm{d}x\,\mathrm{d}y\right]^{1/2}=\mathrm{MMD}(\mu_{\nu,p}(\cdot,t),\mu_{\nu,q}(\cdot,t)).$$

To show that $\lim_{t\to 0}\frac{d}{\mathrm{d}t}\,\mathrm{MMD}(m_t)=\mathrm{KGFD}(p,q)$, we first analyze the differential properties of the easier to handle MMD^2 and complete the proof using a chain rule argument. The first variation (also called Gateaux Derivative) of $m\mapsto\mathrm{MMD}^2(m)$ is a linear functional on the space of functions

$$\left\{f-g\;\middle|\;f,g\colon\mathcal{X}\times[0,\infty)\to\mathbb{R}\quad\text{with}\quad\forall t\geq 0\colon\int_{\mathcal{X}}f(x,t)\,\mathrm{d}x=\int_{\mathcal{X}}g(x,t)\,\mathrm{d}x=1\right\},$$

given by

$$\frac{\delta \,\mathrm{MMD}^2}{\delta m} : f \mapsto \int 2k(x,y)m_t(x)f(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Using the chain rule for Gateaux derivatives, we have that

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{MMD}^2(m)}{\mathrm{d}t} &= \frac{\mathrm{d}\,\mathrm{MMD}^2}{\mathrm{d}m}(m)\frac{\mathrm{d}m}{\mathrm{d}t} \\ &= \int 2k(x,y)m_t(x)\frac{\mathrm{d}m}{\mathrm{d}t}(y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

From the Fokker-Planck Equation, we have that

$$\begin{aligned}\frac{\mathrm{d}m}{\mathrm{d}t} &= \partial_t \mu_{\nu,p} - \partial_t \mu_{\nu,q} \\ &= \mathrm{div}_x(\mu_{\nu,p}\nabla_x \log\frac{p}{\mu_{\nu,p}}) - \mathrm{div}_x(\mu_{\nu,q}\nabla_x \log\frac{q}{\mu_{\nu,q}}) \\ &= \mathrm{div}_x(\nu\nabla_x \log\frac{p}{\nu}) - \mathrm{div}_x(\nu\nabla_x \log\frac{q}{\nu}) + o(1) \\ &= \mathrm{div}_x(\nu\nabla_x \log\frac{p}{q}) + o(1)\end{aligned}$$

Plugging the last equation in the chain rule, we have:

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{MMD}^2(m)}{\mathrm{d}t} &= \int 2m_t(x)\mathrm{div}_y\nu(y)\nabla_y \log\frac{p}{q}(y)k(x,y)\mathrm{d}x\mathrm{d}y + o(1) \\ &= \int 2m_t(x)\left\langle \nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle \mathrm{d}x\,\mathrm{d}y + o(1).\end{aligned}$$

Similarly, since $m_0 = \mu_{\nu,p}(\cdot,0) - \mu_{\nu,q}(\cdot,0) = \nu - \nu = 0$, we have $m_t(x) = t\partial_t m(x,0) + o_x(t)$. The calculation follows as:

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{MMD}^2(m)}{\mathrm{d}t} &= \int 2t\times\partial_t m(x,t)\left\langle \nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle \mathrm{d}x\,\mathrm{d}y + o(t) \\ &= \int 2t\times \mathrm{div}_x\,\nu(x)\nabla_x \log\frac{p}{q}(x)\left\langle \nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle \mathrm{d}x\,\mathrm{d}y + o(t) \\ &= \int 2t\times \left\langle \nu(x)\nabla_x \log\frac{p}{q}(x), \nabla_x\left\langle \nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle\right\rangle \mathrm{d}x\,\mathrm{d}y + o(t) \\ &= \int 2t\times \left\langle \nu(x)\nabla_x \log\frac{p}{q}(x), \nabla_x\nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle \mathrm{d}x\,\mathrm{d}y + o(t).\end{aligned}$$

To get rid of the degenerate scaling as $t\to 0$, we now focus on (the derivative of) $\sqrt{\mathrm{MMD}^2(m_t)}$ as $t\to 0$. Notice that since $\mathrm{MMD}(m_0)=0$, the derivative of $\sqrt{\mathrm{MMD}^2(m_t)}$ does not exist a priori for $t=0$: we consider instead $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\mathrm{MMD}^2(m_t)}\Big|_{t=t}$, and extend it by continuity by setting $t\to 0$. We have:

$$\frac{\mathrm{d}\sqrt{\mathrm{MMD}^2(m_t)}}{\mathrm{d}t} = \frac{1}{2\sqrt{\mathrm{MMD}^2(m_t)}}\frac{\mathrm{d}\,\mathrm{MMD}^2(m_t)}{\mathrm{d}t}.$$

As

$$\mathrm{MMD}^2(m_t) = \int k(x,y)m_t(x)m_t(y)\,\mathrm{d}x\,\mathrm{d}y$$

we obtain through similar calculations that

$$\mathrm{MMD}^2(m_t) = \int\int t^2\left\langle \nu(x)\nabla_x \log\frac{p}{q}(x), \nabla_x\nabla_y k(x,y), \nu(y)\nabla_y \log\frac{p}{q}(y)\right\rangle \mathrm{d}x\,\mathrm{d}y + o(t)$$

from which the results follows. Note that the matrix-valued kernel $(K(x,y))_{ij} = (\nabla_x \nabla_y k(x,y))_{ij}$ is positive definite, a result akin to one of Zhou [2008] but for the matrix-valued case. Indeed, for all $x, y \in \mathcal{X}$, $z, t \in \mathbb{R}^d$,

$$zK(x,y)t = \left\langle \sum_{i=1}^{d} z_i \partial_i k(x,\cdot), \sum_{i=1}^{d} t_i \partial_i k(y,\cdot) \right\rangle_{\mathcal{H}}$$

where $\partial_i k(x,\cdot) \in \mathcal{H}$ [Zhou, 2008]. In the following, we write $\phi(x,y) = \sum_{i=1}^{d} y_i \partial_i k(x_i,\cdot)$. Now, for all sets of $\{x^i\}_{i=1}^n \in \mathcal{X}$, $\{y^j\}_{i=1}^n \in \mathbb{R}^d$, we have

$$\begin{aligned}\sum_{i,j=1}^{n} \left\langle K(x^i,x^j)y^j, y^i \right\rangle_{\mathbb{R}^d} &= \sum_{i,j=1}^{n} \left\langle \phi(x^i,y^i), \phi(x^j,y^j) \right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \phi(x^i,y^i), \sum_{i=1}^{n} \phi(x^i,y^i) \right\rangle_{\mathcal{H}} \geq 0\end{aligned}$$

from which it follows that K is indeed positive definite [Micchelli and Pontil, 2005, Theorem 2.1].

D.2 UNIVERSALITY OF THE EXPONENTIATED-GFD AND EXPONENTIATED-KGFD KERNEL

To prove the universality of K_ν and $K_{\nu,K}$ under the assumptions discussed in the related propositions, we rely on the following theorem [Christmann and Steinwart, 2010, Theorem 2.2].

Theorem D.1. *On a compact metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ and for a continuous and injective map $\phi : \mathcal{Z} \mapsto H$, where H is a separable Hilbert space, the kernel $K(z,z') = e^{-\gamma \|\phi(z) - \phi(z')\|_H^2}$ is universal.*

We first focus on the universality of K_ν. We set as our goal to apply that theorem to our setting, in which $\mathcal{Z} := \mathcal{P}_\mathcal{X}$ is a (sub)set of probability densities, which needs to be associated with a suitably chosen metric in order to make $\mathcal{P}_\mathcal{X}$ to be compact, and ϕ continuous. As bounded subsets of differentiable densities, whose elements can be framed as elements of the Sobolev space of first order $\mathcal{W}^{2,1}(\nu)$ [Taylor, 1996]), are not compact a priori, we restrict ourselves to twice-differentiable densities with bounded Sobolev norm of second order, i.e., to $\mathcal{W}^{2,2}(\nu)$ with norm $\|p\|_{\mathcal{W}^{2,2}}^2 := \|p\|_{\mathcal{L}_2(\nu)}^2 + \sum_{i=1}^{d} \|\partial_i p\|_{\mathcal{L}_2(\nu)}^2 + \sum_{i,j=1}^{d} \|\partial_i \partial_j p\|_{\mathcal{L}_2(\nu)}^2$. From the Rellich-Kondrachov theorem [Taylor, 1996], we know that when ν has compact support, the canonical canonical injection $I\colon \mathcal{W}^{2,2}(\nu) \to \mathcal{W}^{2,1}(\nu)$ is a compact operator. As a consequence, for any bounded subset A of $\mathcal{P}_\mathcal{X}$ we thus have that $I(A)$ is compact for $\|f\|_{\mathcal{W}^{2,1}}^2 := \|f\|_{\mathcal{L}_2(\nu)}^2 + \sum_{i=1}^{d} \|\partial_i f\|_{\mathcal{L}_2(\nu)}^2$, which implies that any bounded subset A of $P_\mathcal{X}$ is compact for $d(z,z') = \|z - z'\|_{\mathcal{W}^{2,1}}$. To apply the above theorem, it remains to prove the continuity and injectivity of $\phi : p \mapsto \nabla \log p$ under this metric (in that case the separable Hilbert space H is set to $\mathcal{L}_2(\nu)$). And indeed, for such a choice of d, ϕ and H, ϕ is continuous. To prove this fact, remark that differentiable densities with full support on $\mathcal{X}$ are bounded away from 0, making the use of a $\phi\colon p \mapsto \nabla \log p = \nabla p / p$ continuous. Moreover, ϕ is injective as $d_{W^{2,1}}(p,q) := \|p - q\|_{W^{2,1}} \neq 0$ implies $\|\nabla \log p - \nabla \log q\|_{\mathcal{L}_2(\nu)} \neq 0$. Thus, all conditions of [Christmann and Steinwart, 2010, Theorem 2.2] are satisfied, and the result follows as a consequence.

We now move on to prove the universality of $K_{\nu,K}$. The proof follows the same reasoning as the proof of the universality of K_ν, the only difference being the fact that the feature map $\tilde{\phi}$ of $K_{K,\nu}$ is given by $T_\nu \circ \phi$, where $\phi\colon p \mapsto \nabla \log p$ and $T_{K,\nu}\colon \mathcal{L}(\mathcal{X}, \mathbb{R}^d) \to \mathcal{H}_K$ is given by

$$T_{K,\nu}\colon f \mapsto \int_{\mathcal{X}} K_x f(x)\, \nu(\mathrm{d}x).$$

However, if ν is a probability measure and K is bounded, then $T_{K,\nu}$ is a bounded operator, and thus continuous, making $\tilde{\phi}$ continuous. Moreover, if K is characteristic, $T_{K,\nu}$ is injective. Thus $\tilde{\phi}$ is injective and continuous, from which the result follows by Christmann and Steinwart [2010].

E BACKGROUND ON STEIN AND FISHER DIVERGENCES

The Fisher Divergence Consider two continuously differentiable densities p and q on $\mathbb{R}^d$. Then the Fisher divergence [Sriperumbudur et al., 2017, Johnson, 2004] between p and q is defined as:

$$\mathrm{FD}(p||q) = \int_{\mathbb{R}^d} \|\nabla \log p(x) - \nabla \log q(x)\|_2^2\, p(x)\, \mathrm{d}x.$$

We refer to Sriperumbudur et al. [2017] for an overview of the properties of the Fisher divergence, including its relative strength w.r.t. other divergences, and other formulations. The Fisher divergence was used for learning statistical models of some training data in Hyvärinen [2005], Sriperumbudur et al. [2017], and more recently in Song and Ermon [2019].

Stein Discrepancies Of proximity to the Fisher divergence is the family of Stein discrepancies [Anastasiou et al., 2022]. Stein discrepancies build upon the concept of Stein operators, which are operators $\mathcal{A}_{\mathbb{P}}$ such that

$$\mathbb{E}_{\mathbb{Q}}\left[\mathcal{A}_{\mathbb{P}} f\right] = 0 \iff \mathbb{Q} = \mathbb{P}$$

for any f within a set $\mathcal{G}(\mathcal{A}_{\mathbb{P}}) \subset \mathrm{dom}(\mathcal{A}_{\mathbb{P}})$ called the *Stein class* of $\mathcal{A}_{\mathbb{P}}$. Following this definition, the $\mathcal{A}_{\mathbb{P}}$-stein discrepancy is defined as

$$\mathrm{SD}_{\mathcal{A}_{\mathbb{P}}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in G(\mathcal{A})} \|\mathbb{E}_{\mathbb{Q}}\, \mathcal{A} f\|$$

which satisfies by construction the axioms of a *dissimilarity* (or *divergence*) measure between $\mathbb{P}$ and $\mathbb{Q}$.

Link Between the Fisher divergence and Diffusion Stein Discrepancies Perhaps the most famous Stein discrepancy is the one that sets $\mathcal{A}_{\mathbb{P}}$ to be the infinitesimal generator of the isotropic diffusion process toward $\mathbb{P}$ [Gorham et al., 2019]:

$$\begin{cases} \mathrm{d}X_t & = \nabla \log p(X_t)\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t \\ (\mathcal{A}_{d,\mathbb{P}} f)(\cdot) & = \langle \nabla \log p(\cdot), \nabla f \rangle + \langle \nabla, \nabla f \rangle \end{cases}$$

Recalling that $\mathbb{E}_{\mathbb{P}}\left[\mathcal{A}_{d,\mathbb{P}} f\right] = 0$ for all $f \in \mathcal{G}(\mathcal{A}_{d,\mathbb{P}})$, we obtain the following formulation for the diffusion Stein discrepancy

$$\begin{aligned} \mathrm{SD}_{\mathcal{A}_{d,\mathbb{P}}}(\mathbb{P}, \mathbb{Q}) &\coloneqq \sup_f \|\mathbb{E}_{\mathbb{Q}}\, \mathcal{A}_{d,\mathbb{P}} f\| = \sup_f \left\|\mathbb{E}_{\mathbb{Q}}(\nabla \log p - \nabla \log q)^\top \nabla f\right\| \\ &= \sup_{g = \nabla f} \left\|\mathbb{E}_{\mathbb{Q}}(\nabla \log p - \nabla \log q)^\top g\right\|, \end{aligned}$$

highlighting the connection between the Fisher divergence and the diffusion Stein discrepancy.

Link Between the Fisher divergence and the Kernelized Stein Discrepancy Given a RKHS $\mathcal{H}$ such that $B_{\mathcal{H}^{\otimes d}}(0_{\mathcal{H}^{\otimes d}}, 1)$ is a Stein class for $\mathcal{A}_{d,\mathbb{P}}$, the kernelized Stein discrepancy [Gorham and Mackey, 2017] is given by

$$\begin{aligned} \mathrm{KSD}(\mathbb{P}, \mathbb{Q}) &\coloneqq \sup_{h = \nabla f \in \mathcal{H}^{\otimes d} : \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}} \langle \nabla \log p(x) - \nabla \log q(x), h(x) \rangle\| \\ &= \sup_{h = \nabla f \in \mathcal{H}^{\otimes d} : \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \langle h, \mathbb{E}_{\mathbb{Q}}(\nabla \log p(x) - \nabla \log q(x)) k(x, \cdot) \rangle_{\mathcal{H}^{\otimes d}}^{1/2} \\ &= \|\mathbb{E}_{\mathbb{Q}}\left[(\nabla \log p(x) - \nabla \log q(x)) k(x, \cdot)\right]\|_{\mathcal{H}^{\otimes d}} \\ &= \left\|I^\star_{k,\mathbb{Q}}(\nabla \log p - \nabla \log q)\right\|_{\mathcal{H}^{\otimes d}} \end{aligned}$$

where $I^\star_{k,\mathbb{Q}}$ is the adjoint of the canonical injection from $\mathcal{H}^{\otimes d}$ to $(L^2(\mathbb{Q}))^{\otimes d}$, also known as the *kernel integral operator*. This derivation shows that the KSD can be seen as a kernelized version of the Fisher divergence.

Link between MMD and KSD It is possible [Gorham and Mackey, 2017] to reframe the KSD as an MMD with a specific kernel. Indeed, given some base kernel $k(x, y)$, define the following "Stein" kernel

$$\tilde{k}(x, y) = \langle \nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot), \nabla \log p(y) k(y, \cdot) + \nabla \log k(y, \cdot) \rangle_{\mathcal{H}^{\otimes d}}$$

which is positive definite as an inner product of a feature map of x. Then $\mathcal{H}_{\tilde{k}} = \mathcal{A}_{d,\mathbb{P}}(\mathcal{H})$ and $\|f\|_{\mathcal{H}_{\tilde{k}}} = \|\mathcal{A} f\|_{\mathcal{H}_k^{\otimes d}}$. Moreover, we have that $\mathbb{E}_{\mathbb{P}}\, \tilde{h} = 0$ for all $\tilde{h} \in \mathcal{H}_{\tilde{k}}$. By the definition of the KSD, we have that

$$\begin{aligned} \mathrm{KSD}(\mathbb{P}, \mathbb{Q}) &= \sup_{h \in \mathcal{H}^{\otimes d} :\ \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}}\, \mathcal{A}_{d,\mathbb{P}} h\| \\ &= \sup_{h \in \mathcal{H}^{\otimes d} :\ \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}}\, \mathcal{A}_{d,\mathbb{P}} h - \mathbb{E}_{\mathbb{P}}\, \mathcal{A}_{d,\mathbb{P}} h\|_{\mathcal{H}} \\ &= \sup_{h \in \mathcal{H}_{\tilde{k}} :\ \|h\|_{\mathcal{H}_{\tilde{k}}} \leq 1} \|\mathbb{E}_{\mathbb{Q}}\, h - \mathbb{E}_{\mathbb{P}}\, h\|_{\mathcal{H}_{\tilde{k}}} \\ &= \mathrm{MMD}_{\tilde{k}}(\mathbb{P}, \mathbb{Q}). \end{aligned}$$

Differential Inequalities between the KL and the Fisher Divergence It is well known [Carrillo et al., 2003] that the KL divergence can be related to the Fisher divergence by considering the evolution of $\mathrm{KL}(\mathbb{P}_t||\mathbb{Q})$ when $\mathbb{P}_t$ evolves according to the Fokker-Planck equation

$$\partial_t p_t(x) = \mathrm{div}(p_t(x)(\nabla \log q_t(x) - \nabla \log p_t(x))), \quad \mathbb{P}_0 = \mathbb{P}. \tag{E.1}$$

(Two relevant side notes: for any $t \geq 0$, $\mathbb{P}_t$ is the law at time t of the Markov process $(X_t)_{t\geq 0}$ such that $X_0 \sim \mathbb{P}$ and undergoing an isotropic diffusion towards $\mathbb{Q}$. Moreover, Equation (E.1) is also the Wasserstein gradient flow equation of $\mathrm{KL}(\cdot||\mathbb{Q})$ starting from $\mathbb{P}$). Recalling that Equation (E.1) is satisfied in the sense of distributions, and relying on Gateaux-Derivative formulas for Free Energy-type functionals [see Ambrosio et al., 2005, for more precise statements], we have:

$$\begin{aligned}
\frac{\mathrm{dKL}(\mathbb{P}_t||\mathbb{Q})}{\mathrm{d}t} &= \frac{\partial \,\mathrm{KL}}{\mathrm{d}\mathbb{P}}\bigg|_{\mathbb{P}_t} \frac{d\mathbb{P}_t}{\mathrm{d}t} \\
&= \int \langle \nabla(\log p_t(x) - \log q_t(x)), (\nabla \log q_t - \nabla \log p_t) \rangle \, \mathrm{d}\mathbb{P}_t(x) \\
&= -\,\mathrm{FD}(\mathbb{P}_t, \mathbb{Q}).
\end{aligned}$$

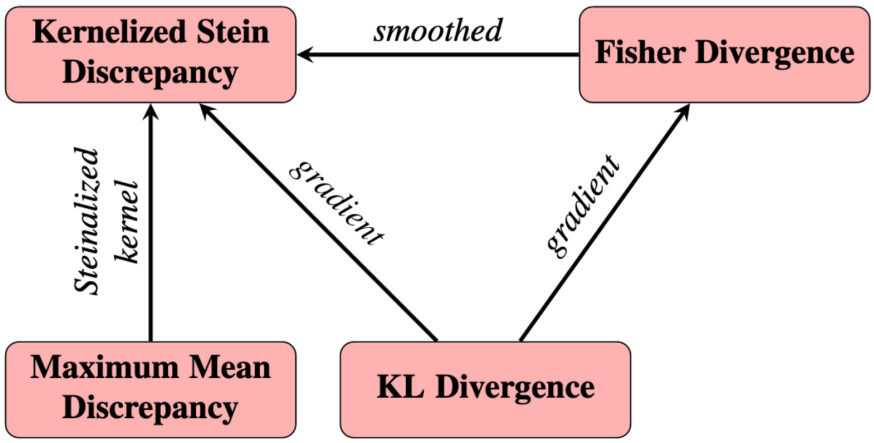

Figure E.1: Relationships between the Fisher divergence, the KL divergence, the MMD, and the KSD [Liu, 2016].

F EXPERIMENTAL RESULTS

This section contains visualizations of all experiments discussed in **??**, including figures contained in the main text. In all experiments we set the significance level to $\alpha = 0.05$. Every experiment is repeated for 100 randomly sampled datasets and with 500 bootstrap iterations for estimating the quantile of the test statistic.

We use Gaussian distributions and compare the KCCSD and the SKCE with different combinations of kernels. For the KCCSD, for Gaussian distributions all considered test statistics can be evaluated exactly. Alternatively, for the exponentiated (kernelized) Fisher kernel and the exponentiated MMD kernel one can resort to approximations using samples from the base measure. For the SKCE, however, the test statistic can be evaluated exactly on in special cases such as Gaussian kernels on the target space. All approximate evaluations are performed with 10 samples.

F.1 MEAN GAUSSIAN MODEL

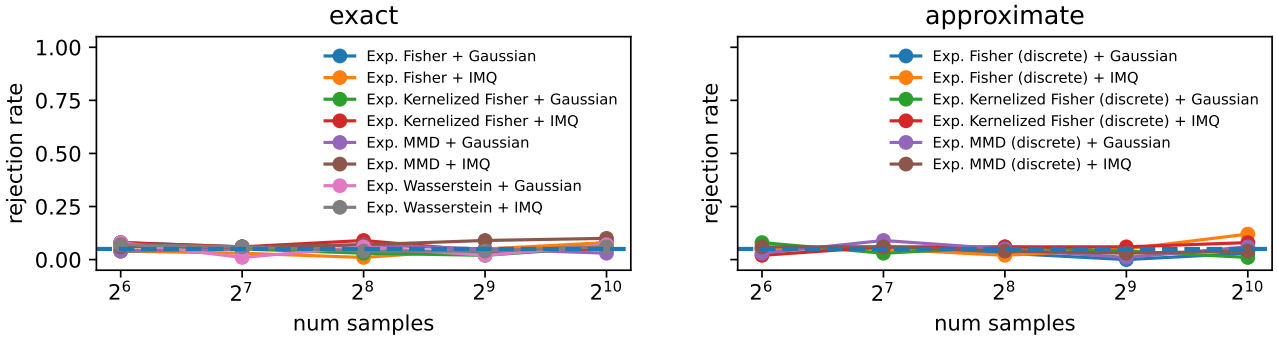

Figure F.1: False rejection rate of the KCCSD for MGM ($\delta = 0$).

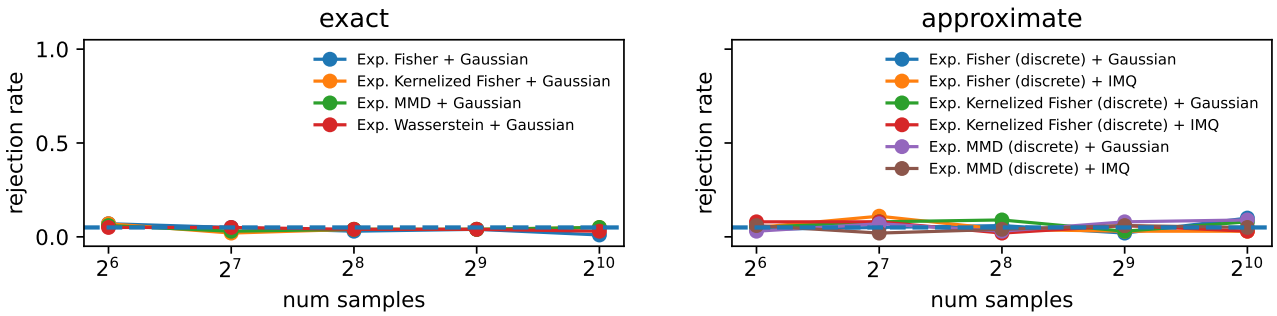

Figure F.2: False rejection rate of the SKCE for MGM ($\delta = 0$).

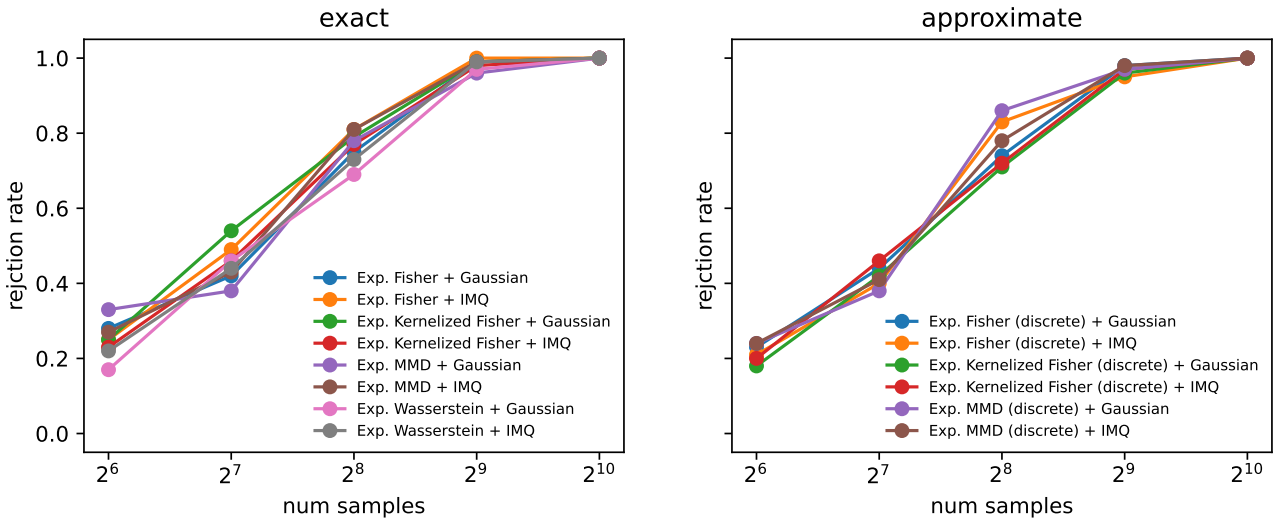

Figure F.3: Rejection rate of the KCCSD for MGM ($\delta = 0.1$, $c = \mathbf{1}_d$).

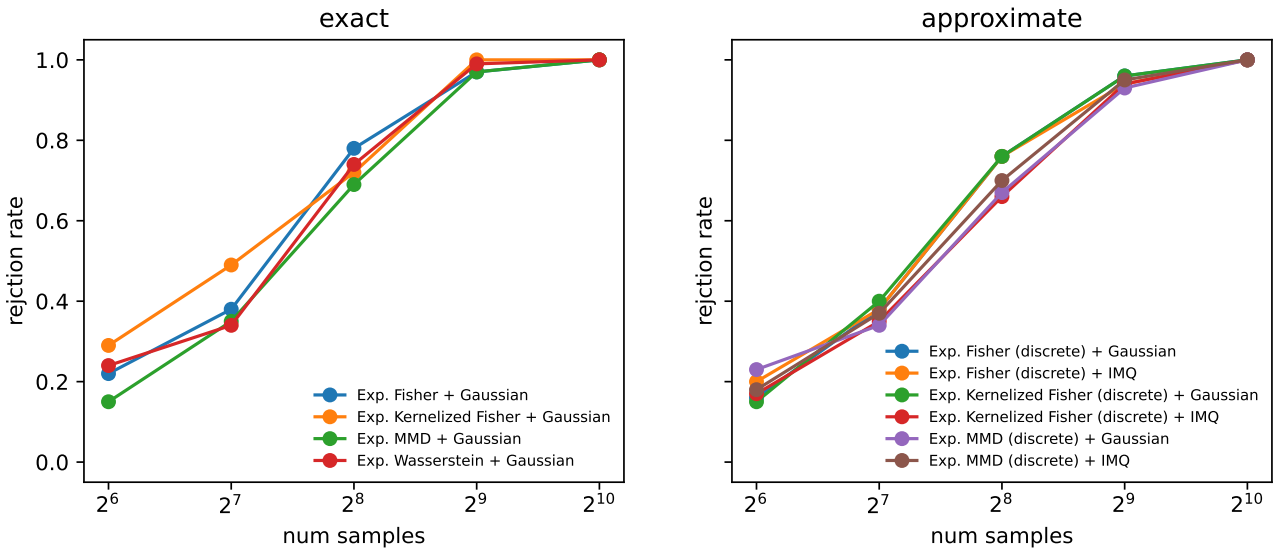

Figure F.4: Rejection rate of the SKCE for MGM ($\delta = 0.1$, $c = \mathbf{1}_d$).

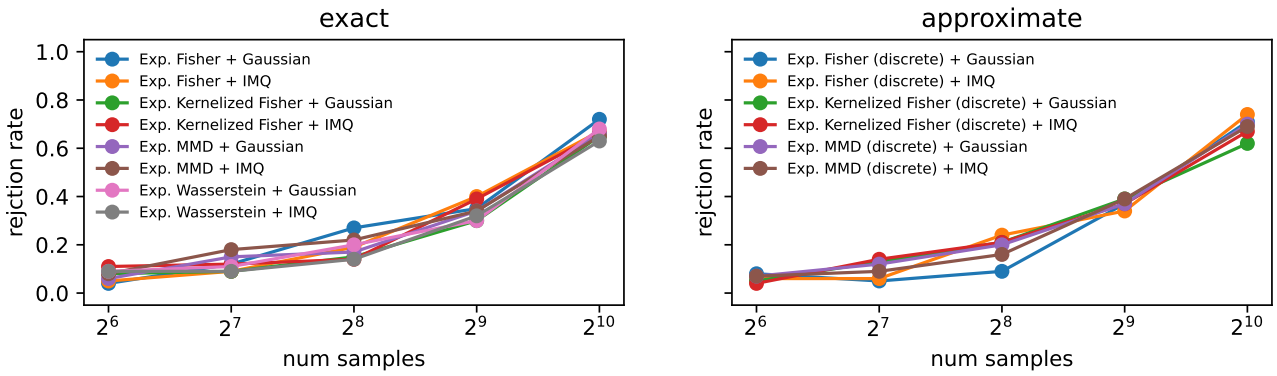

Figure F.5: Rejection rate of the KCCSD for MGM ($\delta = 0.1$, $c = e_1$).

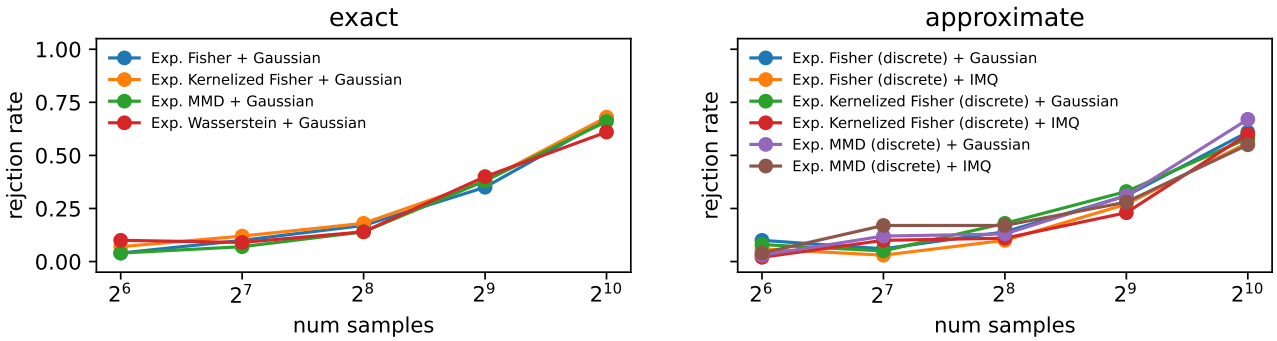

Figure F.6: Rejection rate of the SKCE for MGM ($\delta = 0.1$, $c = e_1$).

F.2 LINEAR GAUSSIAN MODEL

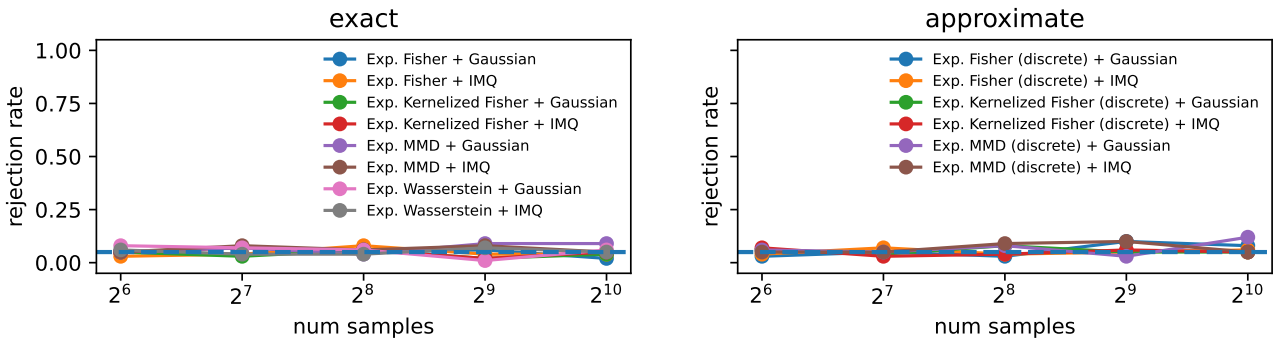

Figure F.7: False rejection rate of the KCCSD for LGM ($\delta = 0$).

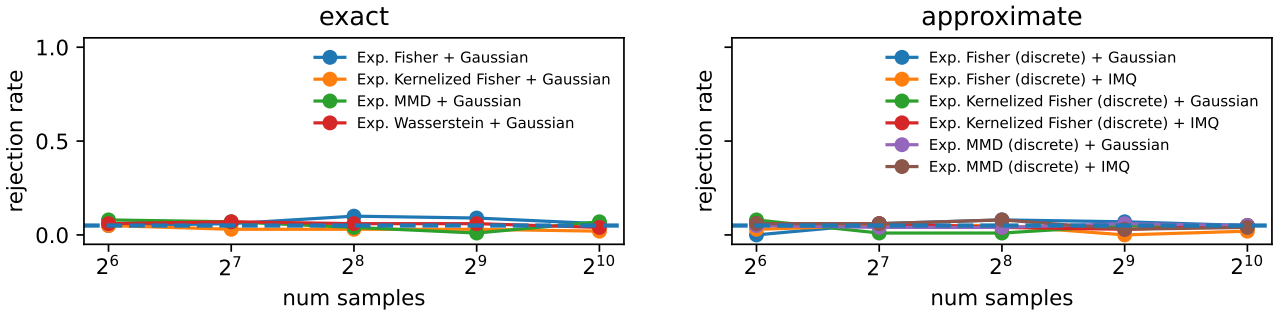

Figure F.8: False rejection rate of the SKCE for LGM ($\delta = 0$).

F.3 HETEROSCEDASTIC GAUSSIAN MODEL

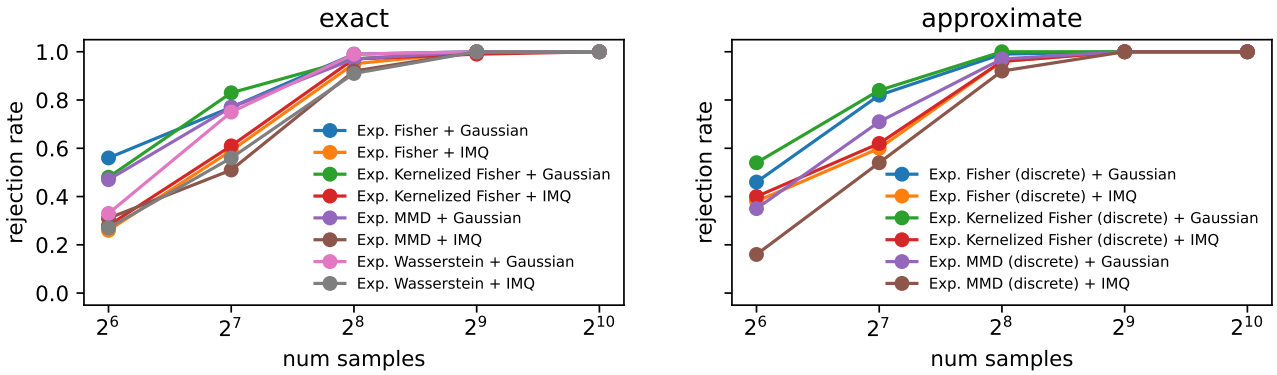

Figure F.9: Rejection rate of the KCCSD for HGM ($\delta = 1$).

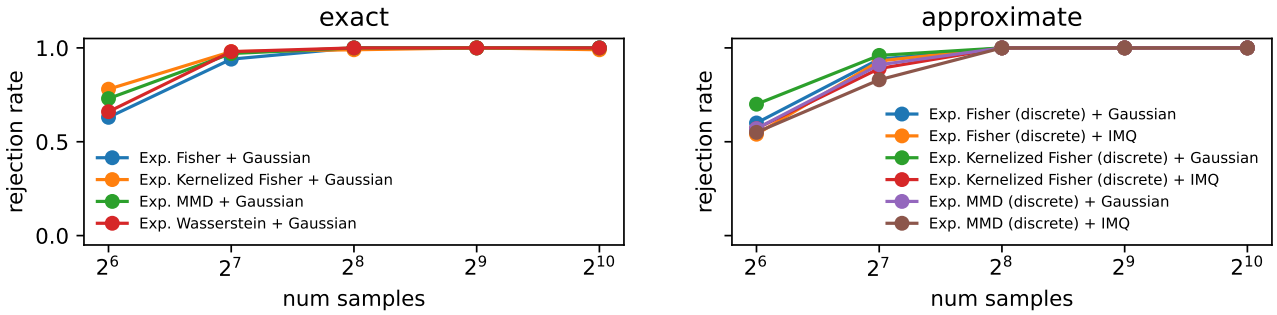

Figure F.10: Rejection rate of the SKCE for HGM ($\delta = 1$).

F.4 QUADRATIC GAUSSIAN MODEL

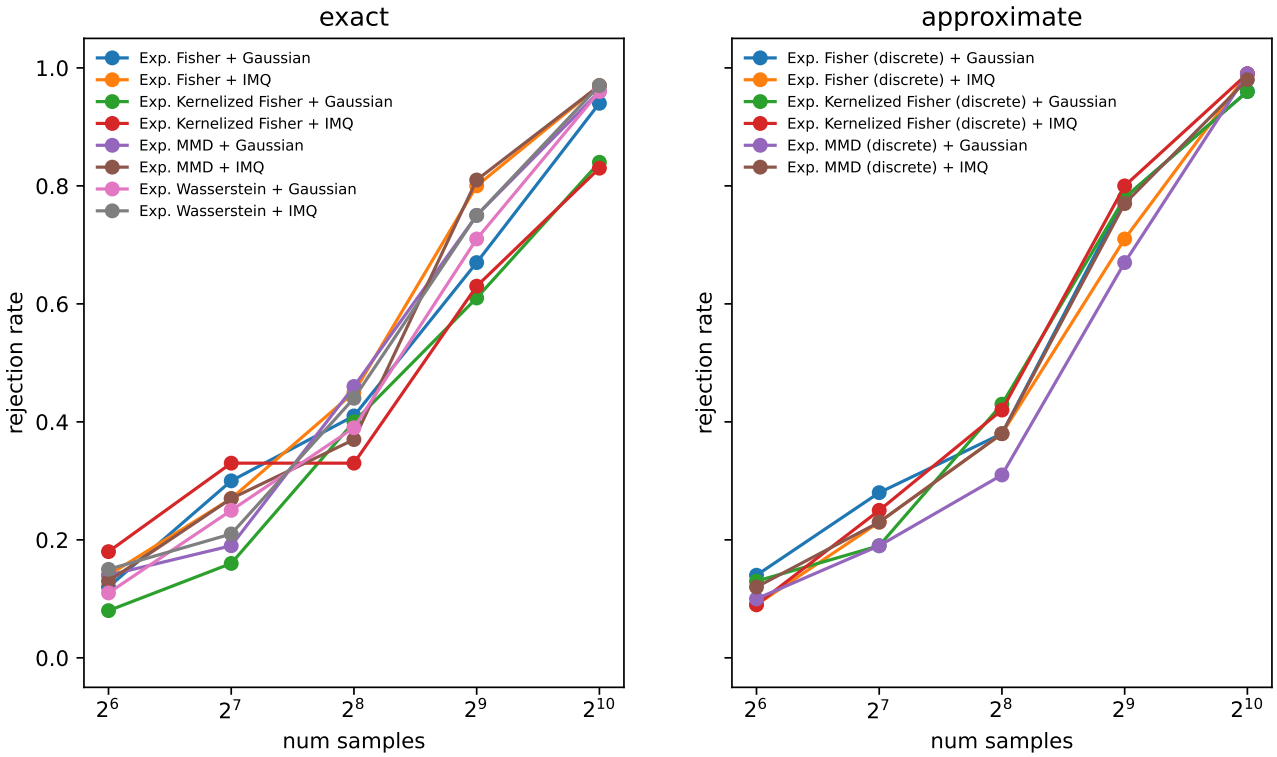

Figure F.11: Rejection rate of the KCCSD for QGM ($\delta = 1$).

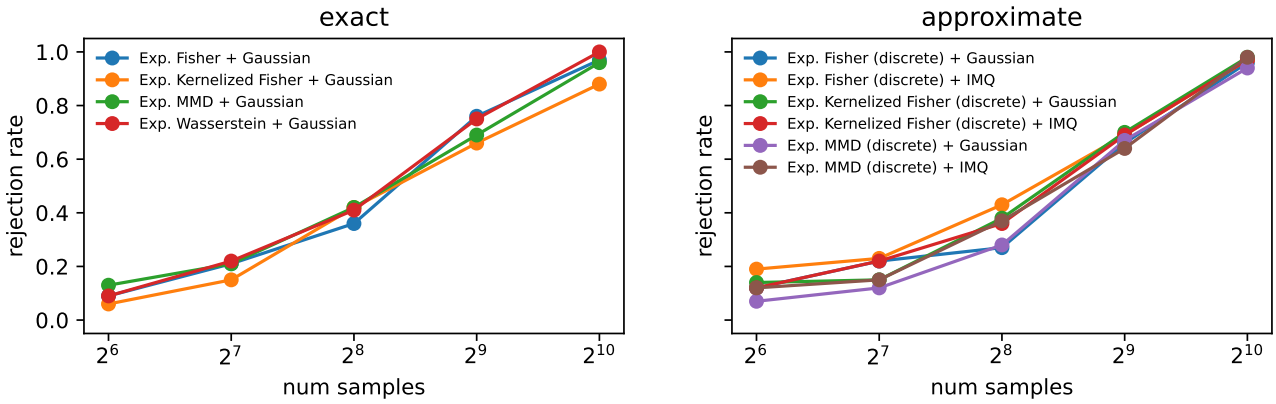

Figure F.12: Rejection rate of the SKCE for QGM ($\delta = 1$).

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: a review of some recent developments. *Statistical Science*, 2022.

Ingo Steinwart Andreas Christmann. *Support Vector Machines*. Springer New York, 2008.

José A Carrillo, Robert J McCann, and Cédric Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Revista Matematica Iberoamericana*, 2003.

Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. *Advances in neural information processing systems*, 23, 2010.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, 2017.

Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference, 2021.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 1996.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6(24):695–709, 2005.

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Oliver Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

Qiang Liu. A short introduction to kernelized Stein discrepancy, 2016.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 2005.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.*, 2017.

Michael Eugene Taylor. *Partial differential equations. 1, Basic theory*. Springer, 1996.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 2008.